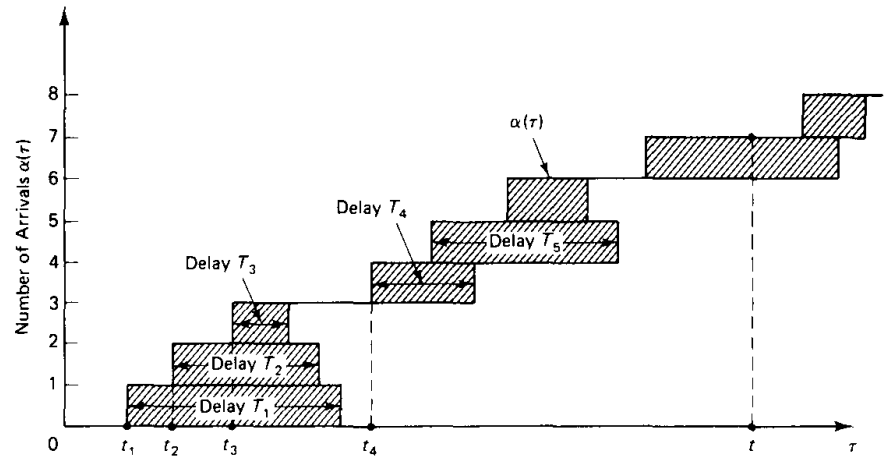
# 3




# Delay Models in Data Networks

## 3.1 INTRODUCTION

One of the most important performance measures of a data network is the average delay required to deliver a packet from origin to destination. Furthermore, delay considerations strongly influence the choice and performance of network algorithms, such as routing and flow control. For these reasons, it is important to understand the nature and mechanism of delay, and the manner in which it depends on the characteristics of the network.

Queueing theory is the primary methodological framework for analyzing network delay. Its use often requires simplifying assumptions since, unfortunately, more realistic assumptions make meaningful analysis extremely difficult. For this reason, it is sometimes impossible to obtain accurate quantitative delay predictions on the basis of queueing models. Nevertheless, these models often provide a basis for adequate delay approximations, as well as valuable qualitative results and worthwhile insights.

In what follows, we will mostly focus on packet delay within the communication subnet (*i.e.*, the network layer). This delay is the sum of delays on each subnet link traversed by the packet. Each link delay in turn consists of four components.

1. The *processing delay* between the time the packet is correctly received at the head node of the link and the time the packet is assigned to an outgoing link queue for transmission. (In some systems, we must add to this delay some additional processing time at the DLC and physical layers.)
2. The *queueing delay* between the time the packet is assigned to a queue for transmission and the time it starts being transmitted. During this time, the packet waits while other packets in the transmission queue are transmitted.
3. The *transmission delay* between the times that the first and last bits of the packet are transmitted.
4. The *propagation delay* between the time the last bit is transmitted at the head node of the link and the time the last bit is received at the tail node. This is proportional to the physical distance between transmitter and receiver; it can be relatively substantial, particularly for a satellite link or a very high speed link.

This accounting neglects the possibility that a packet may require retransmission on a link due to transmission errors or various other causes. For most links in practice, other than multiaccess links to be considered in Chapter 4, retransmissions are rare and will be neglected. The propagation delay depends on the physical characteristics of the link and is independent of the traffic carried by the link. The processing delay is also independent of the amount of traffic handled by the corresponding node if computation power is not a limiting resource. This will be assumed in our discussion. Otherwise, a separate processing queue must be introduced prior to the transmission queues. Most of our subsequent analysis focuses on the queueing and transmission delays. We first consider a single transmission line and analyze some classical queueing models. We then take up the network case and discuss the type of approximations involved in deriving analytical delay models.

While our primary emphasis is on packet-switched network models, some of the models developed are useful in a circuit-switched network context. Indeed, queueing theory was developed extensively in response to the need for performance models in telephony.

### 3.1.1 Multiplexing of Traffic on a Communication Link

The communication link considered is viewed as a bit pipe over which a given number of bits per second can be transmitted. This number is called the *transmission capacity* of the link. It depends on both the physical channel and the interface (*e.g.*, modems), and is simply the rate at which the interface accepts bits. The link capacity may serve several traffic streams (*e.g.*, virtual circuits or groups of virtual circuits) multiplexed on the link. The manner of allocation of capacity among these traffic streams has a profound effect on packet delay.

In the most common scheme, *statistical multiplexing*, the packets of all traffic streams are merged into a single queue and transmitted on a first-come first-serve basis. A variation of this scheme, which has roughly the same average delay per packet, maintains

a separate queue for each traffic stream and serves the queues in sequence one packet at a time. However, if the queue of a traffic stream is empty, the next traffic stream is served and no communication resource is wasted. Since the entire transmission capacity $C$ (bits/sec) is allocated to a single packet at a time, it takes $L/C$ seconds to transmit a packet that is $L$ bits long.

In *time-division* (TDM) *and frequency-division multiplexing* (FDM) with $m$ traffic streams, the link capacity is essentially subdivided into $m$ portions—one per traffic stream. In FDM, the channel bandwidth $W$ is subdivided into $m$ channels each with bandwidth $W/m$ (actually slightly less because of the need for guard bands between channels). The transmission capacity of each channel is roughly $C/m$, where $C$ is the capacity that would be obtained if the entire bandwidth were allocated to a single channel. The transmission time of a packet that is $L$ bits long is $Lm/C$, or $m$ times larger than in the corresponding statistical multiplexing scheme. In TDM, allocation is done by dividing the time axis into slots of fixed length (*e.g.*, one bit or one byte long, or perhaps one packet long for fixed length packets). Again, conceptually, we may view the communication link as consisting of $m$ separate links with capacity $C/m$. In the case where the slots are short relative to packet length, we may again regard the transmission time of a packet $L$ bits long as $Lm/C$. In the case where the slots are of packet length, the transmission time of an $L$ bit packet is $L/C$, but there is a wait of $(m-1)$ packet transmission times between packets of the same stream.

One of the themes that will emerge from our queueing analysis is that statistical multiplexing has smaller average delay per packet than either TDM or FDM. This is particularly true when the traffic streams multiplexed have a relatively low duty cycle. The main reason for the poor delay performance of TDM and FDM is that communication resources are wasted when allocated to a traffic stream with a momentarily empty queue, while other traffic streams have packets waiting in their queue. For a traffic analogy, consider an $m$-lane highway and two cases. In one case, cars are not allowed to cross over to other lanes (this corresponds to TDM or FDM), while in the other case, cars can change lanes (this corresponds roughly to statistical multiplexing). Restricting crossover increases travel time for the same reason that the delay characteristics of TDM or FDM are poor: namely, some system resources (highway lanes or communication channels) may not be utilized, while others are momentarily stressed.

Under certain circumstances, TDM or FDM may have an advantage. Suppose that each traffic stream has a "regular" character (*i.e.*, all packets arrive sufficiently apart so that no packet has to wait while the preceding packet is transmitted.) If these traffic streams are merged into a single queue, it can be shown that the average delay per packet will decrease, but the variance of waiting time in queue will generally become positive (for an illustration, see Prob. 3.7). Therefore, if maintaining a small variability of delay is more important than decreasing delay, it may be preferable to use TDM or FDM. Another advantage of TDM and FDM is that there is no need to include identification of the traffic stream on each packet, thereby saving some overhead and simplifying packet processing at the nodes. Note also that when overhead is negligible, one can afford to make packets very small, thereby reducing delay through pipelining (cf. Fig. 2.37).

## 3.2 QUEUEING MODELS—LITTLE'S THEOREM

We consider queueing systems where customers arrive at random times to obtain service. In the context of a data network, customers represent packets assigned to a communication link for transmission. Service time corresponds to the packet transmission time and is equal to $L/C$, where $L$ is the packet length in bits and $C$ is the link transmission capacity in bits/sec. In this chapter it is convenient to ignore the layer 2 distinction between packets and frames; thus packet lengths are taken to include frame headers and trailers. In a somewhat different context (which we will not emphasize very much), customers represent ongoing conversations (or virtual circuits) between points in a network and service time corresponds to the duration of a conversation. In a related context, customers represent active calls in a telephone or circuit switched network and again service time corresponds to the duration of the call.

We shall be typically interested in estimating quantities such as:

1. The average number of customers in the system (*i.e.*, the "typical" number of customers either waiting in queue or undergoing service)
2. The average delay per customer (*i.e.*, the "typical" time a customer spends waiting in queue plus the service time).

These quantities will be estimated in terms of known information such as:

1. The customer arrival rate (*i.e.*, the "typical" number of customers entering the system per unit time)
2. The customer service rate (*i.e.*, the "typical" number of customers the system serves per unit time when it is constantly busy)

In many cases the customer arrival and service rates are not sufficient to determine the delay characteristics of the system. For example, if customers tend to arrive in groups, the average customer delay will tend to be larger than when their arrival times are regularly spaced apart. Thus to predict average delay, we will typically need more detailed (statistical) information about the customer interarrival and service times. In this section, however, we will largely ignore the availability of such information and see how far we can go without it.

### 3.2.1 Little's Theorem

We proceed to clarify the meaning of the terms "average" and "typical" that we used somewhat liberally above in connection with the number of customers in the system, the customer delay, and so on. In doing so we will derive an important result known as *Little's Theorem*.

Suppose that we observe a sample history of the system from time $t = 0$ to the indefinite future and we record the values of various quantities of interest as time

progresses. In particular, let

$$N(t) = \text{Number of customers in the system at time } t$$

$$\alpha(t) = \text{Number of customers who arrived in the interval } [0, t]$$

$$T_i = \text{Time spent in the system by the } i^{\text{th}} \text{ arriving customer}$$

Our intuitive notion of the "typical" number of customers in the system observed up to time $t$ is

$$N_t = \frac{1}{t} \int_0^t N(\tau)\, d\tau$$

which we call the *time average of* $N(\tau)$ *up to time* $t$. Naturally, $N_t$ changes with the time $t$, but in many systems of interest, $N_t$ tends to a steady-state $N$ as $t$ increases, that is,

$$N = \lim_{t \to \infty} N_t$$

In this case, we call $N$ the *steady-state time average* (or simply time average) of $N(\tau)$. It is also natural to view

$$\lambda_t = \frac{\alpha(t)}{t}$$

as the *time average arrival rate* over the interval $[0, t]$. The *steady-state arrival rate* is defined as

$$\lambda = \lim_{t \to \infty} \lambda_t$$

(assuming that the limit exists). The *time average of the customer delay up to time* $t$ is similarly defined as

$$T_t = \frac{\sum_{i=0}^{\alpha(t)} T_i}{\alpha(t)} \tag{3.1}$$

that is, the average time spent in the system per customer up to time $t$. The *steady-state time average customer delay* is defined as

$$T = \lim_{t \to \infty} T_t$$

(assuming that the limit exists).

It turns out that the quantities $N$, $\lambda$, and $T$ above are related by a simple formula that makes it possible to determine one given the other. This result, known as Little's Theorem, has the form

$$N = \lambda T$$

Little's Theorem expresses the natural idea that crowded systems (large $N$) are associated with long customer delays (large $T$) and reversely. For example, on a rainy day, traffic on a rush hour moves slower than average (large $T$), while the streets are more crowded (large $N$). Similarly, a fast-food restaurant (small $T$) needs a smaller waiting room (small $N$) than a regular restaurant for the same customer arrival rate.

The theorem is really an accounting identity and its derivation is very simple. We will give a graphical proof under some simplifying assumptions. Suppose that the system is initially empty $[N(0) = 0]$ and that customers depart from the system in the order they arrive. Then the number of arrivals $\alpha(t)$ and departures $\beta(t)$ up to time $t$ form a staircase graph as shown in Fig. 3.1. The difference $\alpha(t) - \beta(t)$ is the number in the system $N(t)$ at time $t$. The shaded area between the graphs of $\alpha(\tau)$ and $\beta(\tau)$ can be expressed as

$$\int_0^t N(\tau)\, d\tau$$

and if $t$ is any time for which the system is empty $[N(t) = 0]$, the shaded area is also equal to

$$\sum_{i=1}^{\alpha(t)} T_i$$

Dividing both expressions above with $t$, we obtain

$$\frac{1}{t}\int_0^t N(\tau)\, d\tau = \frac{1}{t}\sum_{i=1}^{\alpha(t)} T_i = \frac{\alpha(t)}{t}\frac{\sum_{i=1}^{\alpha(t)} T_i}{\alpha(t)}$$

or equivalently,

$$N_t = \lambda_t T_t \tag{3.2}$$

Little's Theorem is obtained assuming that

$$N_t \to N,\ \lambda_t \to \lambda,\ T_t \to T$$

and that the system becomes empty infinitely often at arbitrarily large times. With a minor modification in the preceding argument, the latter assumption becomes unnecessary. To see this, note that the shaded area in Fig. 3.1 lies between $\sum_{i=1}^{\alpha(t)} T_i$ and $\sum_{i=1}^{\beta(t)} T_i$, so we obtain

$$\frac{\beta(t)}{t}\frac{\sum_{i=1}^{\beta(t)} T_i}{\beta(t)} \le N_t \le \lambda_t T_t$$

Assuming that $N_t \to N$, $\lambda_t \to \lambda$, $T_t \to T$, and that the departure rate $\beta(t)/t$ up to time $t$ tends to the steady-state arrival rate $\lambda$, we obtain Little's Theorem.

The simplifying assumptions used in the preceding graphical proof can be relaxed considerably, and one can construct an analytical proof that requires only that the limits $\lambda = \lim_{t\to\infty} \alpha(t)/t$, $\delta = \lim_{t\to\infty} \beta(t)/t$, and $T = \lim_{t\to\infty} T_t$ exist, and that $\lambda = \delta$. In particular, it is not necessary that customers are served in the order they arrive, and that the system is initially empty (see Problem 3.41). Figure 3.2 explains why the order of customer service is not essential for the validity of Little's Theorem.

### 3.2.2 Probabilistic Form of Little's Theorem

Little's Theorem admits also a probabilistic interpretation provided that we can replace time averages with statistical or ensemble averages, as we now discuss. Our preceding

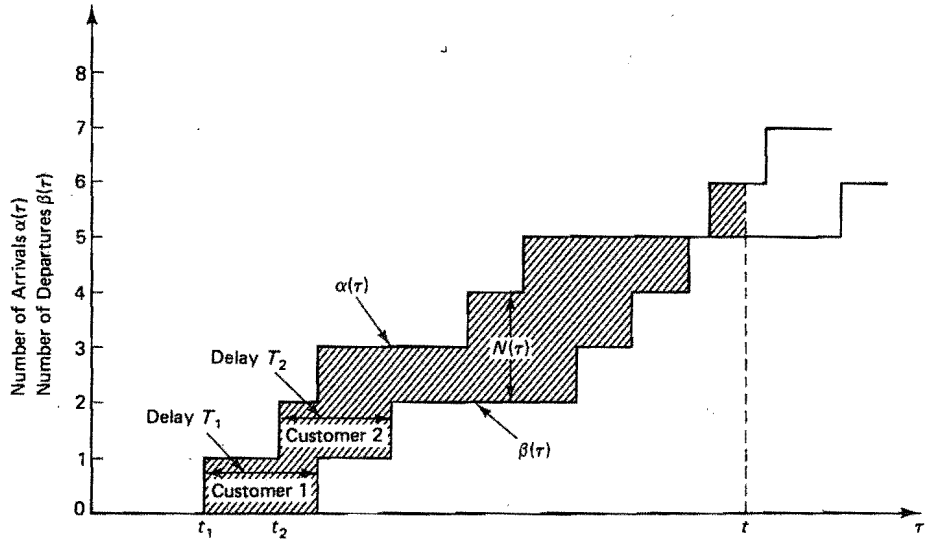



**Figure 3.1** Proof of Little's Theorem. If the system is empty at time $t$ $[N(t) = 0]$, the shaded area can be expressed both as $\int_0^t N(\tau)\,d\tau$ and as $\sum_{i=1}^{\alpha(t)} T_i$. Dividing both expressions by $t$, equating them, and taking the limit as $t \to \infty$ gives Little's Theorem. If $N(t) > 0$, we have

$$\sum_{i=1}^{\beta(t)} T_i \le \int_0^t N(\tau)\,d\tau \le \sum_{i=1}^{\alpha(t)} T_i$$

and assuming that the departure rate $\beta(t)/t$ up to time $t$ tends to the steady-state arrival rate $\lambda$, the same argument applies.

analysis deals with a single sample function; now we will look at the probabilities of many sample functions and other events.

We first need to clarify the meaning of an ensemble average. Let us denote

$$p_n(t) = \text{Probability of } n \text{ customers in the system at time } t$$

$$\text{(waiting in queue or under service)}$$

In a typical situation we are given the initial probabilities $p_n(0)$ at time 0, together with enough statistical information to determine, at least in principle, the probabilities $p_n(t)$ for all times $t$. For example, the probability distribution of the time between two successive arrivals (the interarrival time), and the probability distribution of the customers' service time at various parts of the queueing system may be given. Then the average number in the system at time $t$ is given by

$$\overline{N}(t) = \sum_{n=0}^{\infty} n p_n(t)$$

Note that both $\overline{N}(t)$ and $p_n(t)$ depend on $t$ as well as the initial probability distribution

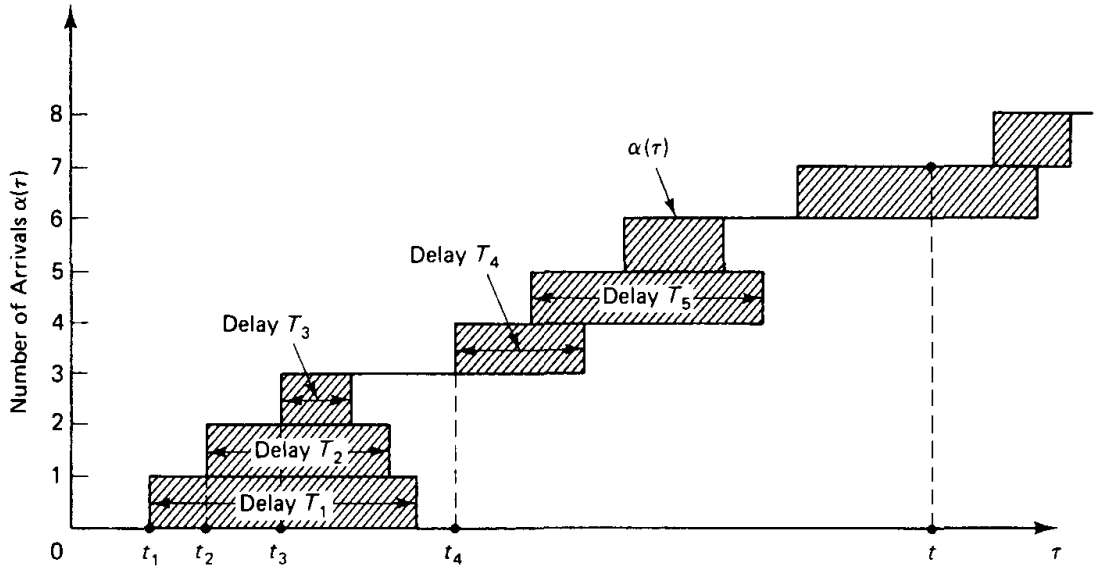



**Figure 3.2** Informal justification of Little's Theorem without assuming first-in first-out customer service. The shaded area can be expressed both as $\int_0^t N(\tau)\,d\tau$ and as $\sum_{i \in D(t)} T_i + \sum_{i \in \overline{D}(t)} (t - t_i)$, where $D(t)$ is the set of customers that have departed up to time $t$, $\overline{D}(t)$ is the set of customers that are still in the system at time $t$, and $t_i$ is the time of arrival of the $i^{\text{th}}$ customer. Dividing both expressions by $t$, equating them, and taking the limit as $t \to \infty$ gives Little's Theorem.

$\{p_0(0),\ p_1(0), \ldots\}$. However, the queueing systems that we will consider typically reach a steady-state in the sense that for some $p_n$ (independent of the initial distribution), we have

$$\lim_{t \to \infty} p_n(t) = p_n\,, \quad n = 0, 1, \ldots.$$

The average number in the system at steady-state is given by

$$\overline{N} = \sum_{n=0}^{\infty} n p_n$$

and we typically have

$$\overline{N} = \lim_{t \to \infty} \overline{N}(t)$$

Regarding average delay per customer, we are typically given enough statistical information to determine in principle the probability distribution of delay of each individual customer (*i.e.*, the first, second, etc.). From this, we can determine the average delay of each customer. The average delay of the $k^{\text{th}}$ customer, denoted $\overline{T}_k$, typically converges as $k \to \infty$ to a steady-state value

$$\overline{T} = \lim_{k \to \infty} \overline{T}_k$$

To make the connection with time averages, we note that almost every system of interest to us is *ergodic* in the sense that the time average, $N = \lim_{t \to \infty} N_t$, of a sample

function is, with probability 1, equal to the steady-state average $\overline{N} = \lim_{t\to\infty} \overline{N}(t)$, that is,

$$N = \lim_{t\to\infty} N_t = \lim_{t\to\infty} \overline{N}(t) = \overline{N}$$

Similarly, for the systems of interest to us, the time average of customer delay $T$ is also equal (with probability 1) to the steady-state average delay $\overline{T}$, that is,

$$T = \lim_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} T_i = \lim_{k\to\infty} \overline{T}_k = \overline{T}$$

Under these circumstances, Little's formula, $N = \lambda T$, holds with $N$ and $T$ being stochastic averages and with $\lambda$ given by

$$\lambda = \lim_{t\to\infty} \frac{\text{Expected number of arrivals in the interval } [0, t]}{t}$$

The equality of long term time and ensemble averages of various stochastic processes will often be accepted in this chapter on intuitive grounds. This equality can often be shown by appealing to general results from the theory of Markov chains (see Appendix A, at the end of this chapter, which states these results without proof). In other cases, this equality, though highly plausible, requires a specialized mathematical proof. Such a proof is typically straightforward for an expert in stochastic processes but requires background that is beyond what is assumed in this book. In what follows we will generally use the time average notation $T$ and $N$ in place of the ensemble average notation $\overline{T}$ and $\overline{N}$, respectively, implicitly assuming the equality of the corresponding time and ensemble averages.

### 3.2.3 Applications of Little's Theorem

The significance of Little's Theorem is due in large measure to its generality. It holds for almost every queueing system that reaches a steady-state. The system need not consist of just a single queue. Indeed, with appropriate interpretation of the terms $N$, $\lambda$, and $T$, the theorem holds for many complex arrival–departure systems. The following examples illustrate its broad applicability.

**Example 3.1**

If $\lambda$ is the arrival rate in a transmission line, $N_Q$ is the average number of packets waiting in queue (but not under transmission), and $W$ is the average time spent by a packet waiting in queue (not including the transmission time), Little's Theorem gives

$$N_Q = \lambda W$$

Furthermore, if $\overline{X}$ is the average transmission time, then Little's Theorem gives the average number of packets under transmission as

$$\rho = \lambda \overline{X}$$

Since at most one packet can be under transmission, $\rho$ is also the line's *utilization factor*, (*i.e.*, the proportion of time that the line is busy transmitting a packet).

**Example 3.2**

Consider a network of transmission lines where packets arrive at $n$ different nodes with corresponding rates $\lambda_1, \ldots, \lambda_n$. If $N$ is the average total number of packets inside the network, then (regardless of the packet length distribution and method for routing packets) the average delay per packet is

$$T = \frac{N}{\sum_{i=1}^{n} \lambda_i}$$

Furthermore, Little's Theorem also yields $N_i = \lambda_i T_i$, where $N_i$ and $T_i$ are the average number in the system and average delay of packets arriving at node $i$, respectively.

**Example 3.3**

A packet arrives at a transmission line every $K$ seconds with the first packet arriving at time 0. All packets have equal length and require $\alpha K$ seconds for transmission where $\alpha < 1$. The processing and propagation delay per packet is $P$ seconds. The arrival rate here is $\lambda = 1/K$. Because packets arrive at a regular rate (equal interarrival times), there is no delay for queueing, so the time $T$ a packet spends in the system (including the propagation delay) is

$$T = \alpha K + P$$

According to Little's Theorem, we have

$$N = \lambda T = \alpha + \frac{P}{K}$$

Here the number in the system $N(t)$ is a deterministic function of time. Its form is shown in Fig. 3.3 for the case where $K < \alpha K + P < 2K$, and it can be seen that $N(t)$ does not converge to any value (the system never reaches statistical equilibrium). However, Little's Theorem holds with $N$ viewed as a time average.

**Example 3.4**

Consider a window flow control system (as described in Section 2.8.1) with a window of size $W$ for each session. Since the number of packets in the system per session is always no more than $W$, Little's Theorem asserts that the arrival rate $\lambda$ of packets into the system for each session, and the average packet delay are related by $W \geq \lambda T$. Thus, if congestion builds up in the network and $T$ increases, $\lambda$ must eventually decrease. Next, suppose that the network is congested and capable of delivering only $\lambda$ packets per unit time for each session. Assuming that acknowledgment delays are negligible relative to the forward packet delays, we have $W \simeq \lambda T$. Then, increasing the window size $W$ for all sessions merely serves to increase the delay $T$ without appreciably changing $\lambda$.

**Example 3.5**

Consider a queueing system with $K$ servers, and with room for at most $N \geq K$ customers (either in queue or in service). The system is always full; we assume that it starts with $N$ customers and that a departing customer is immediately replaced by a new customer. (Queueing systems of this type are called *closed* and are discussed in detail in Section 3.8.) Suppose that the average customer service time is $\overline{X}$. We want to find the average

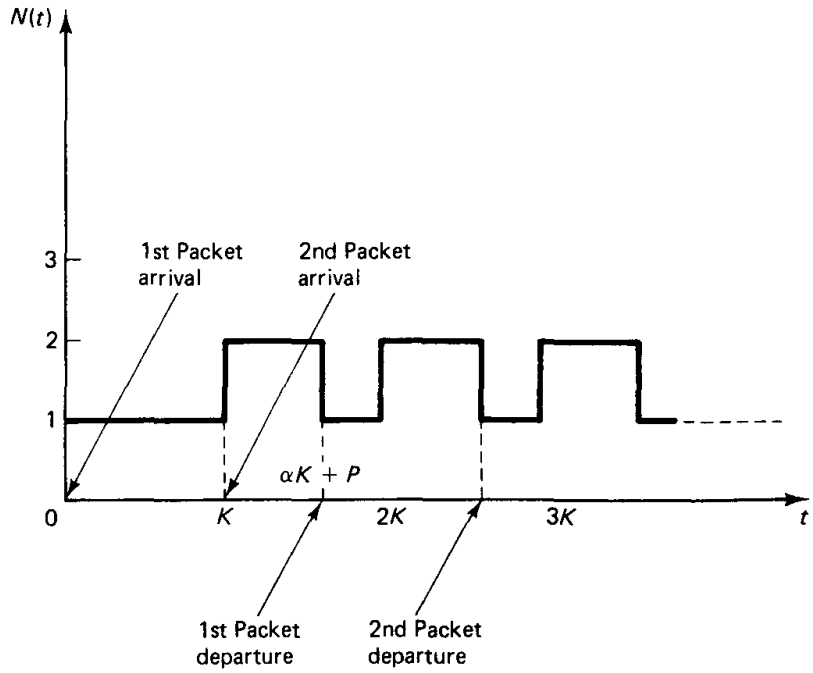



**Figure 3.3** The number in the system in Example 3.3, $N(t)$, is deterministic and does not converge as $t \to \infty$. Little's Theorem holds with $N$, $\lambda$, and $T$ viewed as time averages.

customer time in the system $T$. We apply Little's Theorem twice, first for the entire system, obtaining $N = \lambda T$, and then for the service portion of the system, obtaining $K = \lambda \overline{X}$ (since all servers are constantly busy). By eliminating $\lambda$ in these two relations we have

$$T = \frac{N\overline{X}}{K}$$

Consider also the same system but under different customer arrival assumptions. In particular, assume that customers arrive at a rate $\lambda$ but are blocked (and lost) from the system if they find the system full. Then the number of servers that are busy may be less than $K$. Let $\overline{K}$ be the average number of busy servers and let $\beta$ be the proportion of customers that are blocked from entering the system. Applying Little's Theorem to the service portion of the system, we obtain

$$\overline{K} = (1 - \beta)\lambda\overline{X}$$

from which

$$\beta = 1 - \frac{\overline{K}}{\lambda\overline{X}}$$

Since $\overline{K} \leq K$, we obtain a lower bound on the blocking probability, namely,

$$\beta \geq 1 - \frac{K}{\lambda\overline{X}}$$

**Example 3.6**

A transmission line serves $m$ packet streams, also called users, in round-robin cycles. In each cycle, some packets of user 1 are transmitted, followed by some packets of user 2, and

so on, until finally, some packets of user $m$ are transmitted. An overhead period of average length $A_i$ precedes the transmission of the packets of user $i$ in each cycle. Systems of this type are called *polling systems* and are discussed in detail in Section 3.5.2.

The arrival rate and the average transmission time of the packets of user $i$ are $\lambda_i$ and $\overline{X}_i$, respectively. From Little's theorem we know that the fraction of time the transmission line is busy transmitting packets of user $i$ is $\lambda_i \overline{X}_i$. Consider now the time intervals used for overhead of user $i$. We can view these intervals as "packets" with average transmission time $A_i$. The arrival rate of these "packets" is $1/L$, where $L$ is the average cycle length, and as before, we may use Little's theorem to assert that the fraction of time used for transmission of these "packets" is $A/L$, where $A = A_1 + A_2 + \cdots + A_m$. Therefore, we have

$$1 = \frac{A}{L} + \sum_{i=1}^{m} \lambda_i \overline{X}_i$$

which yields the average cycle length

$$L = \frac{A}{1 - \sum_{i=1}^{m} \lambda_i \overline{X}_i}$$

**Example 3.7 Estimating Throughput in a Time-Sharing System**

Little's Theorem can sometimes be used to provide bounds on the attainable system throughput $\lambda$. In particular, known bounds on $N$ and $T$ can be translated into throughput bounds via $\lambda = N/T$. As an example, consider a time-sharing computer system with $N$ terminals. A user logs into the system through a terminal, and after an initial reflection period of average length $R$, submits a job that requires an average processing time $P$ at the computer. Jobs queue up inside the computer and are served by a single CPU according to some unspecified priority or time-sharing rule.

We would like to get estimates of the throughput sustainable by the system (in jobs per unit time), and corresponding estimates of the average delay of a user. Since we are interested in maximum attainable throughput, we assume that there is always a user ready to take the place of a departing user, so the number of users in the system is always $N$. For this reason, it is appropriate to adopt a model whereby a departing user immediately reenters the system as shown in Fig. 3.4.

Applying Little's Theorem to the portion of the system between the entry to the terminals and the exit of the system (points $A$ and $C$ in Fig. 3.4), we have

$$\lambda = \frac{N}{T} \tag{3.3}$$

where $T$ is the average time a user spends in the system. We have

$$T = R + D \tag{3.4}$$

where $D$ is the average delay between the time a job is submitted to the computer and the time its execution is completed. Since $D$ can vary between $P$ (case where the user's job does not have to wait for other jobs to be completed) and $NP$ (case where the user's job has to wait for the jobs of all the other users; compare with Example 3.5), we have

$$R + P \leq T \leq R + NP \tag{3.5}$$

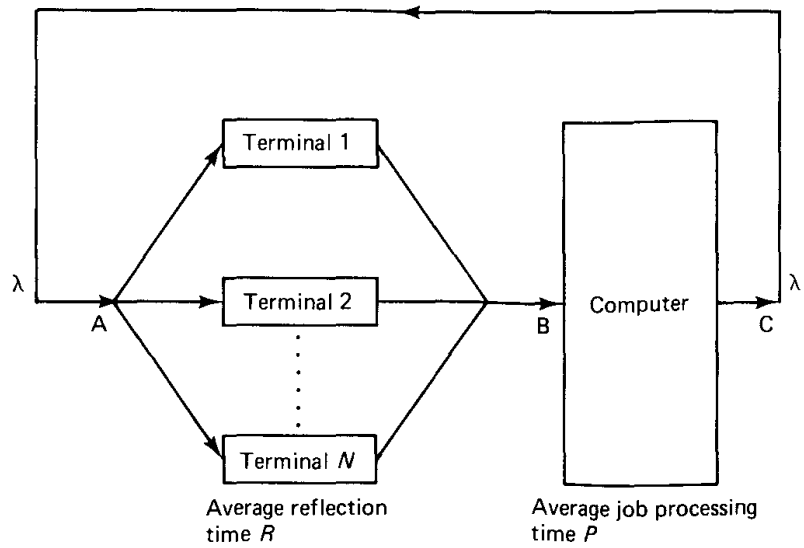



**Figure 3.4** $N$ terminals connected with a time-sharing computer system. To estimate maximum attainable throughput, we assume that a departing user immediately reenters the system or, equivalently, is immediately replaced by a new user.

Combining this relation with $\lambda = N/T$ [cf. Eq. (3.3)], we obtain

$$\frac{N}{R+NP} \leq \lambda \leq \frac{N}{R+P} \tag{3.6}$$

The throughput $\lambda$ is also bounded above by the processing capacity of the computer. In particular, since the execution time of a job is $P$ units on the average, it follows that the computer cannot process in the long run more than $1/P$ jobs per unit time, that is,

$$\lambda \leq \frac{1}{P} \tag{3.7}$$

(This conclusion can also be reached by applying Little's Theorem between the entry and exit points of the computer's CPU.)

By combining the preceding two relations, we obtain the bounds

$$\frac{N}{R+NP} \leq \lambda \leq \min\left\{\frac{1}{P}, \frac{N}{R+P}\right\} \tag{3.8}$$

for the throughput $\lambda$. By using $T = N/\lambda$, we also obtain bounds for the average user delay when the system is fully loaded:

$$\max\{NP,\ R+P\} \leq T \leq R+NP \tag{3.9}$$

These relations are illustrated in Fig. 3.5.

It can be seen that as the number of terminals $N$ increases, the throughput approaches the maximum $1/P$, while the average user delay rises essentially in direct proportion with $N$. The number of terminals becomes a throughput bottleneck when $N < 1 + R/P$, in which case the computer resource stays idle for a substantial portion of the time while all users are engaged in reflection. In contrast, the limited processing power of the computer becomes the bottleneck when $N > 1 + R/P$. It is interesting to note that while the exact maximum attainable throughput depends on system parameters, such as the statistics of the reflection and processing times, and the manner in which jobs are served by the CPU, the

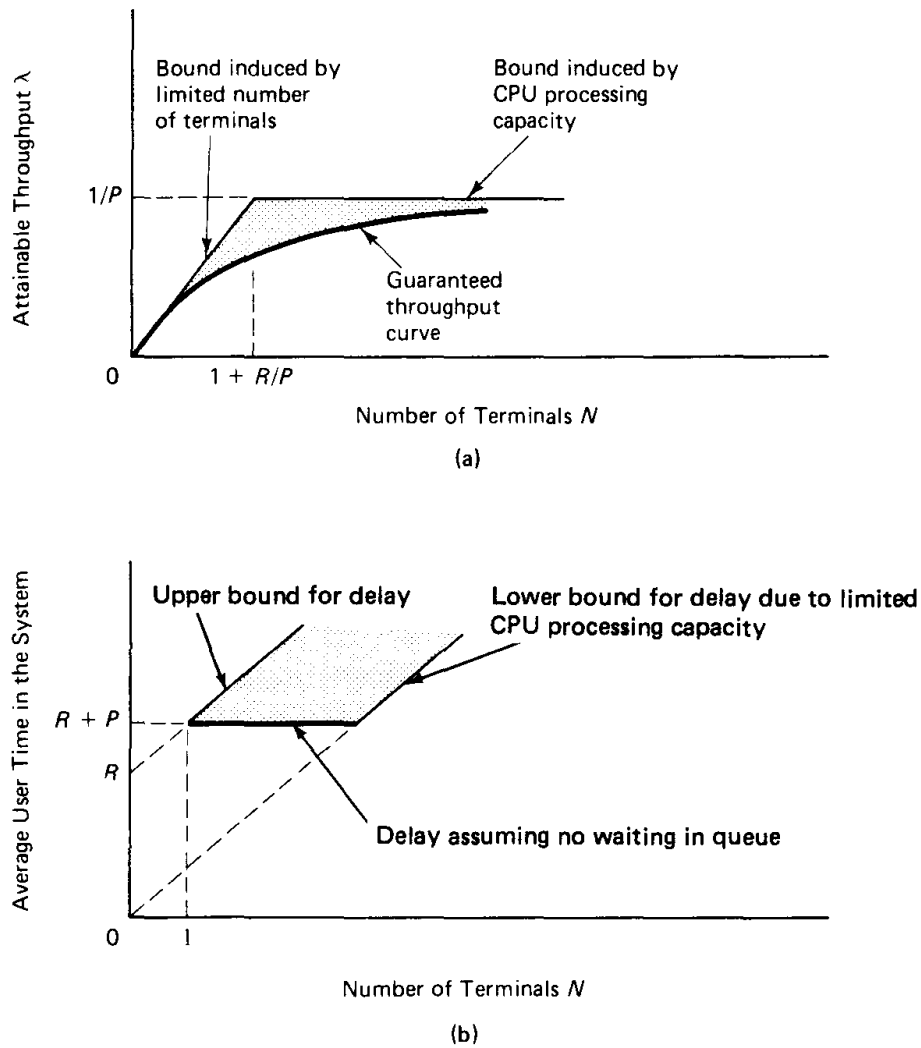



**Figure 3.5** Bounds on throughput and average user delay in a time-sharing system. (a) Bounds on attainable throughput [Eq. (3.8)]. (b) Bounds on average user time in a fully loaded system [Eq. (3.9)]. The time increases essentially in proportion with the number of terminals $N$.

bounds obtained are independent of these parameters. We owe this convenient situation to the generality of Little's Theorem.

## 3.3 THE $M/M/1$ QUEUEING SYSTEM

The $M/M/1$ queueing system consists of a single queueing station with a single server (in a communication context, a single transmission line). Customers arrive according to a Poisson process with rate $\lambda$, and the probability distribution of the service time is exponential with mean $1/\mu$ sec. We will explain the meaning of these terms shortly. The name $M/M/1$ reflects standard queueing theory nomenclature whereby:

1. The first letter indicates the nature of the arrival process [*e.g.*, $M$ stands for memoryless, which here means a Poisson process (*i.e.*, exponentially distributed inter-

arrival times), $G$ stands for a general distribution of interarrival times, $D$ stands for deterministic interarrival times].

2. The second letter indicates the nature of the probability distribution of the service times (*e.g.*, $M$, $G$, and $D$ stand for exponential, general, and deterministic distributions, respectively). In all cases, successive interarrival times and service times are assumed to be statistically independent of each other.
3. The last number indicates the number of servers.

We have already established, via Little's Theorem, the relations

$$N = \lambda T, \qquad N_Q = \lambda W$$

between the basic quantities,

$$\begin{aligned} N &= \text{Average number of customers in the system} \\ T &= \text{Average customer time in the system} \\ N_Q &= \text{Average number of customers waiting in queue} \\ W &= \text{Average customer waiting time in queue} \end{aligned}$$

However, $N$, $T$, $N_Q$, and $W$ cannot be specified further unless we know something more about the statistics of the system. Given these statistics, we will be able to derive the steady-state probabilities

$$p_n = \text{Probability of } n \text{ customers in the system, } n = 0, 1, \ldots.$$

From these probabilities, we can get

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

and using Little's Theorem,

$$T = \frac{N}{\lambda}$$

Similar formulas exist for $N_Q$ and $W$. Appendix B provides a summary of the results for the $M/M/1$ system and the other major systems analyzed later.

The analysis of the $M/M/1$ system as well as several other related systems, such as the $M/M/m$ or $M/M/\infty$ systems, is based on the theory of Markov chains summarized in Appendix A. An alternative approach is to use simple graphical arguments based on the concept of mean residual time introduced in Section 3.5. This approach does not require that the service times are exponentially distributed (*i.e.*, it applies to the $M/G/1$ system). The price paid for this generality is that the characterization of the steady-state probabilities is more complicated than for the $M/M/1$ system. The reader wishing to circumvent the Markov chain analysis may start directly with the $M/G/1$ system in Section 3.5 after a reading of the preliminary facts on the Poisson process given in Sections 3.3.1 and 3.3.2.

### 3.3.1 Main Results

We first introduce our assumptions on the arrival and service statistics of the $M/M/1$ system.

**Arrival statistics—the Poisson process.** In the $M/M/1$ system, customers arrive according to a Poisson process which we now define:

A stochastic process $\{A(t) \mid t \geq 0\}$ taking nonnegative integer values is said to be a *Poisson process* with rate $\lambda$ if

1. $A(t)$ is a counting process that represents the total number of arrivals that have occurred from 0 to time $t$ [*i.e.*, $A(0) = 0$], and for $s < t$, $A(t) - A(s)$ equals the numbers of arrivals in the interval $(s, t]$.
2. The numbers of arrivals that occur in disjoint time intervals are independent.
3. The number of arrivals in any interval of length $\tau$ is Poisson distributed with parameter $\lambda\tau$. That is, for all $t$, $\tau > 0$,

$$P\{A(t+\tau) - A(t) = n\} = e^{-\lambda\tau}\frac{(\lambda\tau)^n}{n!}, \qquad n = 0, 1, \ldots \tag{3.10}$$

The average number of arrivals within an interval of length $\tau$ is $\lambda\tau$ (based on the mean of the Poisson distribution). This leads to the interpretation of $\lambda$ as an arrival rate (average number of arrivals per unit time).

We list some of the properties of the Poisson process that will be of interest:

1. Interarrival times are independent and exponentially distributed with parameter $\lambda$; that is, if $t_n$ denotes the time of the $n^{\text{th}}$ arrival, the intervals $\tau_n = t_{n+1} - t_n$ have the probability distribution

$$P\{\tau_n \leq s\} = 1 - e^{-\lambda s}, \qquad s \geq 0 \tag{3.11}$$

and are mutually independent. [The corresponding probability density function is $p(\tau_n) = \lambda e^{-\lambda\tau_n}$. The mean and variance of $\tau_n$ are $1/\lambda$ and $1/\lambda^2$, respectively.] For a proof of this property, see [Ros83], p. 35.

2. For every $t \geq 0$ and $\delta \geq 0$,

$$P\{A(t+\delta) - A(t) = 0\} = 1 - \lambda\delta + o(\delta) \tag{3.12}$$

$$P\{A(t+\delta) - A(t) = 1\} = \lambda\delta + o(\delta) \tag{3.13}$$

$$P\{A(t+\delta) - A(t) \geq 2\} = o(\delta) \tag{3.14}$$

where we generically denote by $o(\delta)$ a function of $\delta$ such that

$$\lim_{\delta \to 0} \frac{o(\delta)}{\delta} = 0$$

These equations can be verified by expanding the Poisson distribution on the number of arrivals in an interval of length $\delta$ [Eq. (3.10)] in a Taylor series [or equivalently, by writing $e^{-\lambda\delta} = 1 - \lambda\delta + (\lambda\delta)^2/2 - \cdots$].

3. If two or more independent Poisson processes $A_1, \ldots, A_k$ are merged into a single process $A = A_1 + A_2 + \cdots + A_k$, the latter process is Poisson with a rate equal to the sum of the rates of its components (see Problem 3.10).
4. If a Poisson process is split into two other processes by independently assigning each arrival to the first (second) of these processes with probability $p$ ($1 - p$, respectively), the two arrival processes thus obtained are Poisson (see Problem 3.11). (For this it is essential that the assignment of each arrival be independent of the assignment of other arrivals. If, for example, the assignment is done by alternation, with even-numbered arrivals assigned to one process and odd-numbered arrivals assigned to the other, the two generated processes are not Poisson. This will prove to be significant in the context of data networks; see Example 3.17 in Section 3.6.)

A Poisson process is generally considered to be a good model for the aggregate traffic of a large number of similar and independent users. In particular, suppose that we merge $n$ independent and identically distributed packet arrival processes. Each process has arrival rate $\lambda/n$, so that the aggregate process has arrival rate $\lambda$. The interarrival times $\tau$ between packets of the same process have a given distribution $F(s) = P\{\tau \leq s\}$ and are independent [$F(s)$ need not be an exponential distribution]. Then under relatively mild conditions on $F$ [*e.g.*, $F(0) = 0$, $dF(0)/ds > 0$], the aggregate arrival process can be approximated well by a Poisson process with rate $\lambda$ as $n \to \infty$ (see [KaT75], p. 221).

**Service statistics.** Our assumption regarding the service process is that *the customer service times have an exponential distribution with parameter* $\mu$, that is, if $s_n$ is the service time of the $n^{\text{th}}$ customer,

$$P\{s_n \leq s\} = 1 - e^{-\mu s}, \qquad s \geq 0$$

[The probability density function of $s_n$ is $p(s_n) = \mu e^{-\mu s_n}$, and its mean and variance are $1/\mu$ and $1/\mu^2$, respectively.] Furthermore, *the service times* $s_n$ *are mutually independent and also independent of all interarrival times.* The parameter $\mu$ is called the *service rate* and represents the rate (in customers served per unit time) at which the server operates when busy. In the context of a packet transmission system, the independence of interarrival and service times implies, among other things, that the length of an arriving packet does not affect the arrival time of the next packet. It will be seen in Section 3.6 that this condition is often violated in practice, particularly when the arriving packets have just departed from another queue.

An important fact regarding the exponential distribution is its *memoryless* character, which can be expressed as

$$P\{\tau_n > r + t \mid \tau_n > t\} = P\{\tau_n > r\}, \qquad \text{for } r, t \geq 0$$

$$P\{s_n > r + t \mid s_n > t\} = P\{s_n > r\}, \qquad \text{for } r, t \geq 0$$

for the interarrival and service times $\tau_n$ and $s_n$, respectively. This means that the additional time needed to complete a customer's service in progress is independent of when the service started. Similarly, the time up to the next arrival is independent of when the previous arrival occurred. Verification of the memoryless property follows from the calculation

$$P\{\tau_n > r + t \mid \tau_n > t\} = \frac{P\{\tau_n > r + t\}}{P\{\tau_n > t\}} = \frac{e^{-\lambda(r+t)}}{e^{-\lambda t}} = e^{-\lambda r} = P\{\tau_n > r\}$$

**Markov chain formulation.** An important consequence of the memoryless property is that it allows the use of the theory of Markov chains. Indeed, this property, together with our earlier independence assumptions on interarrival and service times, imply that once we know the number $N(t)$ of customers in the system at time $t$, the times at which customers will arrive or complete service in the future are independent of the arrival times of the customers presently in the system and of how much service the customer currently in service (if any) has already received. This means that the future numbers of customers depend on past numbers only through the present number; that is, $\{N(t) \mid t \geq 0\}$ *is a continuous-time Markov chain.*

We could analyze the process $N(t)$ in terms of continuous-time Markov chain methodology; most of the queueing literature follows this line of analysis (see also Problem 3.12). It is sufficient, however, for our purposes in this section to use the simpler theory of discrete-time Markov chains (briefly summarized in Appendix A).

Let us focus attention at the times

$$0,\ \delta,\ 2\delta, \ldots, k\delta, \ldots$$

where $\delta$ is a small positive number. We denote

$$N_k = \text{ Number of customers in the system at time } k\delta$$

Since $N_k = N(k\delta)$ and, as discussed, $N(t)$ is a continuous-time Markov chain, we see that $\{N_k \mid k = 0, 1, \ldots\}$ is a discrete-time Markov chain with steady-state occupancy probabilities equal to those of the continuous chain. Let $P_{ij}$ denote the corresponding transition probabilities

$$P_{ij} = P\{N_{k+1} = j \mid N_k = i\}$$

Note that $P_{ij}$ depends on $\delta$, but to keep notation simple, we do not show this dependence. By using Eqs. (3.12) through (3.14), one can show that

$$P_{oo} = 1 - \lambda\delta + o(\delta) \tag{3.15}$$

$$P_{ii} = 1 - \lambda\delta - \mu\delta + o(\delta), \qquad i \geq 1 \tag{3.16}$$

$$P_{i,i+1} = \lambda\delta + o(\delta), \qquad i \geq 0 \tag{3.17}$$

$$P_{i,i-1} = \mu\delta + o(\delta), \qquad i \geq 1 \tag{3.18}$$

$$P_{ij} = o(\delta), \qquad i \text{ and } j \neq i,\ i+1,\ i-1$$

To see how these equations are verified, note that when at a state $i \geq 1$, the probability of 0 arrivals and 0 departures in a $\delta$-interval $I_k = (k\delta, (k+1)\delta]$ is $(e^{-\lambda\delta})(e^{-\mu\delta})$; this is because the number of arrivals and the number of departures are Poisson distributed and independent of each other. Expanding this in a power series in $\delta$,

$$P\{0 \text{ customers arrive and } 0 \text{ depart in } I_k\} = 1 - \lambda\delta - \mu\delta + o(\delta) \tag{3.19}$$

The probability of 0 arrivals and 1 departure in the interval $I_k$ is $e^{-\lambda\delta}(1 - e^{-\mu\delta})$ if $i = 1$ (since $1 - e^{-\mu\delta}$ is the probability that the customer in service will complete its service within $I_k$), and $e^{-\lambda\delta}(\mu\delta e^{-\mu\delta})$ if $i > 1$ (since $\mu\delta e^{-\mu\delta}$ is the probability that within the interval $I_k$, the customer in service will complete its service while the subsequent customer will not). In both cases we have

$$P\{0 \text{ customers arrive and } 1 \text{ departs in } I_k\} = \mu\delta + o(\delta)$$

Similarly, the probability of 1 arrival and 0 departures in $I_k$ is $(\lambda\delta e^{-\lambda\delta})e^{-\mu\delta}$, so

$$P\{1 \text{ customer arrives and } 0 \text{ depart in } I_k\} = \lambda\delta + o(\delta)$$

These probabilities add up to 1 plus $o(\delta)$. Thus, the probability of more than one arrival or departure is negligible for $\delta$ small. It follows that for $i \geq 1$, $P_{ii}$, which is the probability of an equal number of arrivals and departures in $I_k$, is within $o(\delta)$ of the value in Eq. (3.19); this verifies Eq. (3.16). Equations (3.15), (3.17), and (3.18) are verified in the same way.

The state transition diagram for the Markov chain $\{N_k\}$ is shown in Fig. 3.6, where we have omitted the terms $o(\delta)$.

**Derivation of the stationary distribution.** Consider now the steady-state probabilities

$$p_n = \lim_{k\to\infty} P\{N_k = n\} = \lim_{t\to\infty} P\{N(t) = n\}$$

Note that during any time interval, the total number of transitions from state $n$ to $n+1$ must differ from the total number of transitions from $n+1$ to $n$ by at most 1. Thus asymptotically, the frequency of transitions from $n$ to $n+1$ is equal to the frequency of transitions from $n+1$ to $n$. Equivalently, the probability that the system is in state $n$ and makes a transition to $n+1$ in the next transition interval is the same as the probability that the system is in state $n+1$ and makes a transition to $n$, that is,

$$p_n\lambda\delta + o(\delta) = p_{n+1}\mu\delta + o(\delta)$$

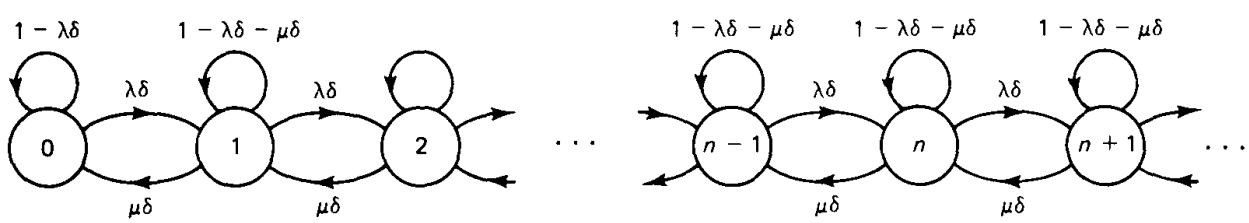



**Figure 3.6** Discrete-time Markov chain for the $M/M/1$ system. The state $n$ corresponds to $n$ customers in the system. Transition probabilities shown are correct up to an $o(\delta)$ term.

By taking the limit in this equation as $\delta \to 0$, we obtain

$$p_n \lambda = p_{n+1} \mu \tag{3.20}$$

(The preceding equations are called *global balance equations*, corresponding to the set of states $\{0, 1, \ldots, n\}$ and $\{n+1, n+2, \ldots\}$. See Appendix A for a more general statement of these equations and for an interpretation that parallels the argument given above.) These equations can also be written as

$$p_{n+1} = \rho p_n, \qquad n = 0, 1, \ldots$$

where

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

It follows that

$$p_{n+1} = \rho^{n+1} p_0, \quad n = 0, 1, \ldots \tag{3.21}$$

If $\rho < 1$ (service rate exceeds arrival rate), the probabilities $p_n$ are all positive and add up to unity, so

$$1 = \sum_{n=0}^{\infty} p_n = \sum_{n=0}^{\infty} \rho^n p_0 = \frac{p_0}{1-\rho} \tag{3.22}$$

Combining the last two equations, we finally obtain

$$p_n = \rho^n (1-\rho), \qquad n = 0, 1, \ldots \tag{3.23}$$

We can now calculate the average number of customers in the system in steady-state:

$$\begin{aligned} N &= \lim_{t \to \infty} E\{N(t)\} = \sum_{n=0}^{\infty} n p_n = \sum_{n=0}^{\infty} n \rho^n (1-\rho) \\ &= \rho(1-\rho) \sum_{n=0}^{\infty} n \rho^{n-1} = \rho(1-\rho) \frac{\partial}{\partial \rho} \left( \sum_{n=0}^{\infty} \rho^n \right) \\ &= \rho(1-\rho) \frac{\partial}{\partial \rho} \left( \frac{1}{1-\rho} \right) = \rho(1-\rho) \frac{1}{(1-\rho)^2} \end{aligned}$$

and finally, using $\rho = \lambda/\mu$, we have

$$N = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda} \tag{3.24}$$

The graph of this equation is shown in Fig. 3.7. As $\rho$ increases, so does $N$, and as $\rho \to 1$, we have $N \to \infty$. The graph is valid for $\rho < 1$. If $\rho > 1$, the server cannot keep up with the arrival rate and the queue length increases without bound. In the context of a packet transmission system, $\rho > 1$ means that $\lambda L > C$, where $\lambda$ is the arrival rate in packets/sec, $L$ is the average packet length in bits, and $C$ is the transmission capacity in bits/sec.

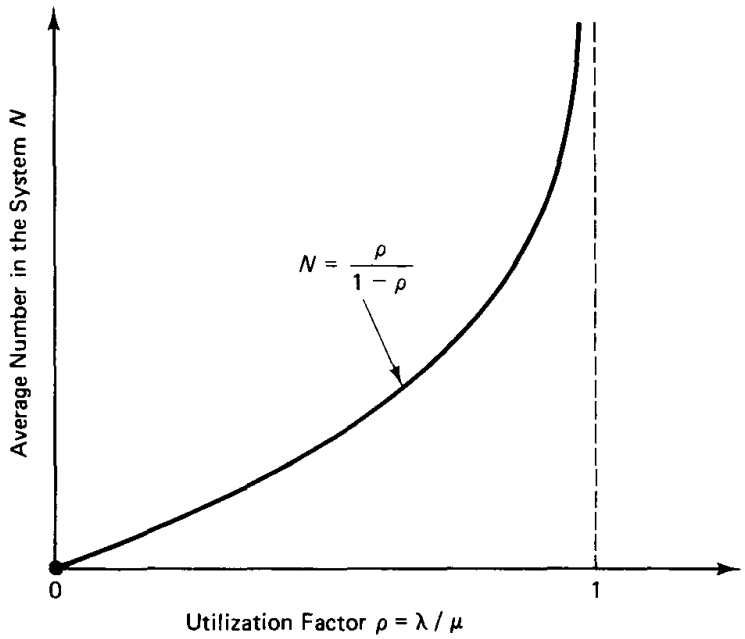



**Figure 3.7** Average number in the system versus the utilization factor in the $M/M/1$ system. As $\rho \to 1$, $N \to \infty$.

The average delay per customer (waiting time in queue plus service time) is given by Little's Theorem,

$$T = \frac{N}{\lambda} = \frac{\rho}{\lambda(1-\rho)} \tag{3.25}$$

Using $\rho = \lambda/\mu$, this becomes

$$T = \frac{1}{\mu - \lambda} \tag{3.26}$$

We note that it is actually possible to show that the customer delay is exponentially distributed in steady-state [see Problem 3.11(b)].

The average waiting time in queue, $W$, is the average delay $T$ less the average service time $1/\mu$, so

$$W = \frac{1}{\mu - \lambda} - \frac{1}{\mu} = \frac{\rho}{\mu - \lambda}$$

By Little's Theorem, the average number of customers in queue is

$$N_Q = \lambda W = \frac{\rho^2}{1-\rho}$$

A very useful interpretation is to view the quantity $\rho$ as the *utilization factor* of the queueing system, (*i.e.*, the long-term proportion of time the server is busy). We showed this earlier in a broader context by using Little's Theorem (Example 3.1). Based on this interpretation, it follows that $\rho = 1 - p_0$, where $p_0$ is the probability of having no customers in the system, and we obtain an alternative verification of the formula derived for $p_0$ [Eq. (3.22)].

We illustrate these results by means of some examples from data networks.

**Example 3.8 Increasing the Arrival and Transmission Rates by the Same Factor**

Consider a packet transmission system whose arrival rate (in packets/sec) is increased from $\lambda$ to $K\lambda$, where $K > 1$ is some scalar factor. The packet length distribution remains the same but the transmission capacity is increased by a factor of $K$, so the average packet transmission time is now $1/(K\mu)$ instead of $1/\mu$. It follows that the utilization factor $\rho$, and therefore the average number of packets in the system, remain the same:

$$N = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}$$

However, the average delay per packet is now $T = N/(K\lambda)$ and is therefore decreased by a factor of $K$. In other words, *a transmission line $K$ times as fast will accommodate $K$ times as many packets/sec at $K$ times smaller average delay per packet.* This result is quite general, even applying to networks of queues. What is happening, as illustrated in Fig. 3.8, is that by increasing arrival rate and service rate by a factor $K$, the statistical characteristics of the queueing process are unaffected except for a change in time scale—the process is speeded up by a factor $K$. Thus, when a packet arrives, it will see ahead of it statistically the same number of packets as with a slower transmission line. However, the packets ahead of it will be moving $K$ times faster.

**Example 3.9 Statistical Multiplexing Compared with Time- and Frequency-Division Multiplexing**

Assume that $m$ statistically identical and independent Poisson packet streams each with an arrival rate of $\lambda/m$ packets/sec are transmitted over a communication line. The packet lengths for all streams are independent and exponentially distributed. The average transmission time is $1/\mu$. If the streams are merged into a single Poisson stream, with rate $\lambda$, as in statistical multiplexing, the average delay per packet is

$$T = \frac{1}{\mu - \lambda}$$

If, instead, the transmission capacity is divided into $m$ equal portions, one per packet stream as in time- and frequency-division multiplexing, each portion behaves like an $M/M/1$ queue with arrival rate $\lambda/m$ and average service rate $\mu/m$. Therefore, the average delay per packet is

$$T = \frac{m}{\mu - \lambda}$$

that is, $m$ times larger than for statistical multiplexing.

The preceding argument indicates that multiplexing a large number of traffic streams on separate channels in a transmission line performs very poorly in terms of delay. The performance is even poorer if the capacity of the channels is not allocated in direct proportion to the arrival rates of the corresponding streams—something that cannot be done (at least in the scheme considered here) if these arrival rates change over time. This is precisely why data networks, which most of the time serve many low duty cycle traffic streams, are typically organized on the basis of some form of statistical multiplexing. An argument in favor of time- and frequency-division multiplexing arises when each traffic stream is "regular" (as opposed to Poisson) in the sense that no packet arrives while another is transmitted, and thus there is no waiting in queue if that stream is transmitted on a dedicated transmission line. If several streams of this type are statistically multiplexed on a single transmission line, the

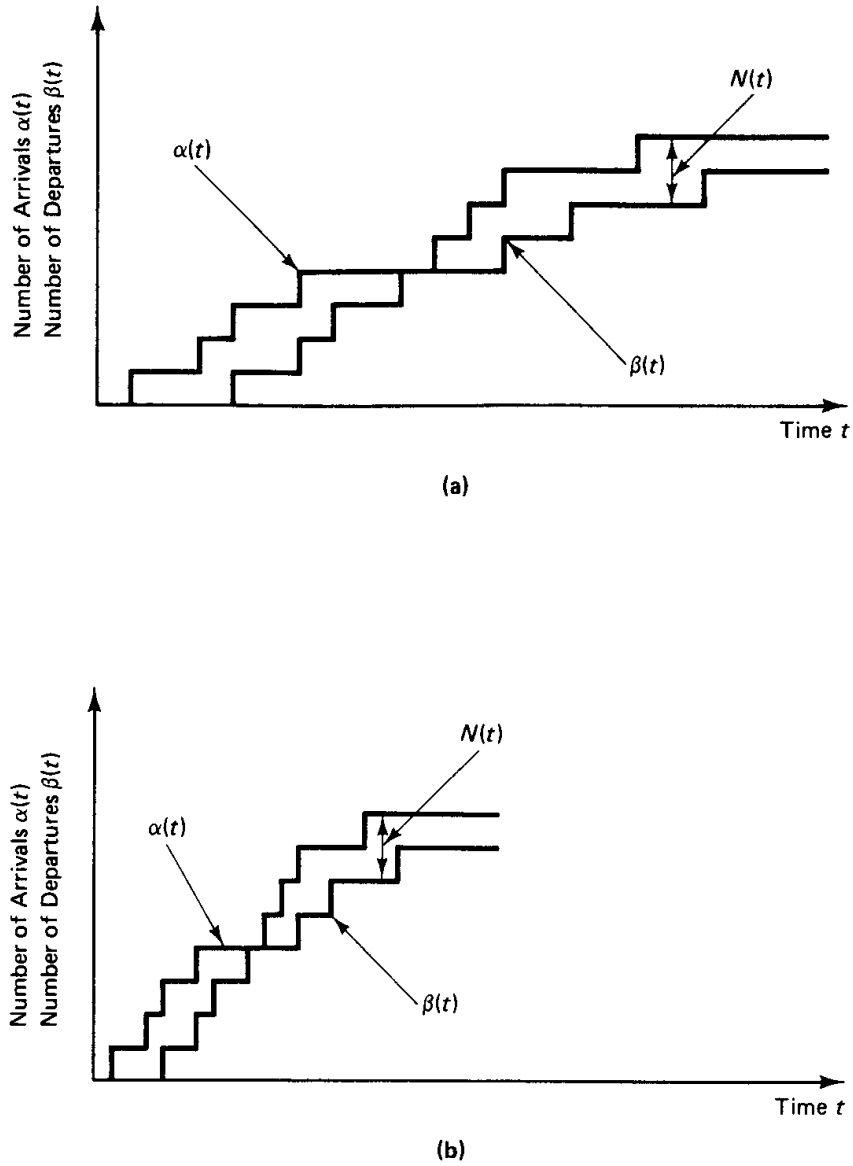



**Figure 3.8** Increasing the arrival rate and the service rate by the same factor (see Example 3.8). (a) Sample paths of number of arrivals $\alpha(t)$ and departures $\beta(t)$ in the original system. (b) Corresponding sample paths of number of arrivals $\alpha(t)$ and departures $\beta(t)$ in the "speeded up" system, where the arrival rate and the service rate have been increased by a factor of 2. The average number in the system is the same as before, but the average delay is reduced by a factor of 2 since customers are moving twice as fast.

average delay per packet will decrease, but the average waiting time in queue will become positive and the variance of delay will also become positive. Thus in telephony, where each traffic stream is a voice conversation that is regular in the sense above and small variability of delay is critical, time- and frequency-division multiplexing are still used widely.

### 3.3.2 Occupancy Distribution upon Arrival

In our subsequent development, there are several situations where we will need a probabilistic characterization of a queueing system as seen by an arriving customer. It is

possible that the times of customer arrivals are in some sense nontypical, so that the steady-state occupancy probabilities upon arrival,

$$a_n = \lim_{t\to\infty} P\{N(t) = n \mid \text{an arrival occurred just after time } t\} \tag{3.27}$$

need not be equal to the corresponding unconditional steady-state probabilities,

$$p_n = \lim_{t\to\infty} P\{N(t) = n\} \tag{3.28}$$

It turns out, however, that for the $M/M/1$ system, we have

$$p_n = a_n, \qquad n = 0, 1, \ldots \tag{3.29}$$

so that an arriving customer finds the system in a "typical" state. Indeed, *this holds under very general conditions for queueing systems with Poisson arrivals regardless of the distribution of the service times.* The only additional requirement we need is that future arrivals are independent of the current number in the system. More precisely, *we assume that for every time $t$ and increment $\delta > 0$, the number of arrivals in the interval $(t, t+\delta)$ is independent of the number in the system at time $t$.* Given the Poisson hypothesis, essentially this amounts to assuming that, at any time, the service times of previously arrived customers and the future interarrival times are independent—something that is reasonable for packet transmission systems. In particular, the assumption holds if the arrival process is Poisson and interarrival times and service times are independent.

For a formal proof of the equality $a_n = p_n$ under the preceding assumption, let $A(t, t+\delta)$ be the event that an arrival occurs in the interval $(t, t+\delta)$. Let

$$p_n(t) = P\{N(t) = n\} \tag{3.30}$$

$$a_n(t) = P\{N(t) = n \mid \text{an arrival occurred just after time } t\} \tag{3.31}$$

We have, using Bayes' rule,

$$\begin{aligned} a_n(t) &= \lim_{\delta\to 0} P\{N(t) = n \mid A(t, t+\delta)\} \\ &= \lim_{\delta\to 0} \frac{P\{N(t) = n,\ A(t, t+\delta)\}}{P\{A(t, t+\delta)\}} \\ &= \lim_{\delta\to 0} \frac{P\{A(t, t+\delta) \mid N(t) = n\}\ P\{N(t) = n\}}{P\{A(t, t+\delta)\}} \end{aligned} \tag{3.32}$$

By assumption, the event $A(t, t+\delta)$ is independent of the number in the system at time $t$. Therefore,

$$P\{A(t, t+\delta) \mid N(t) = n\} = P\{A(t, t+\delta)\}$$

and we obtain from Eq. (3.32)

$$a_n(t) = P\{N(t) = n\} = p_n(t)$$

Taking the limit as $t \to \infty$, we obtain $a_n = p_n$.

As an example of what can happen if the arrival process is not Poisson, suppose that interarrival times are independent and uniformly distributed between 2 and 4 sec,

while customer service times are all equal to 1 sec. Then an arriving customer always finds an empty system. On the other hand, the average number in the system as seen by an outside observer looking at a system at a random time is $1/3$. (The time in the system of each customer is 1 sec, so by Little's Theorem, $N$ is equal to the arrival rate $\lambda$, which is $1/3$ since the expected time between arrivals is 3.)

For a similar example where the arrival process is Poisson but the service times of customers in the system and the future arrival times are correlated, consider a packet transmission system where packets arrive according to a Poisson process. The transmission time of the $n^{\text{th}}$ packet equals one half the interarrival time between packets $n$ and $n+1$. Upon arrival, a packet finds the system empty. However, the average number in the system, as seen by an outside observer, is easily seen to be $1/2$.

### 3.3.3 Occupancy Distribution upon Departure

Let us consider the distribution of the number of customers in the system just after a departure has occurred, that is, the probabilities

$$d_n(t) = P\{N(t) = n \mid \text{a departure occurred just before time } t\}$$

The corresponding steady-state values are denoted

$$d_n = \lim_{t\to\infty} d_n(t), \qquad n = 0, 1, \ldots$$

It turns out that

$$d_n = a_n, \qquad n = 0, 1, \ldots$$

under very general assumptions—the only requirement essentially is that the system reaches a steady-state with all $n$ having positive steady-state probabilities, and that $N(t)$ changes in unit increments. [These assumptions certainly hold for a stable $M/M/1$ system ($\rho < 1$), but they also hold for most stable single-queue systems of interest.] For any sample path of the system and for every $n$, the number in the system will be $n$ infinitely often (with probability 1). This means that for each time the number in the system increases from $n$ to $n+1$ due to an arrival, there will be a corresponding future decrease from $n+1$ to $n$ due to a departure. Therefore, in the long run, the frequency of transitions from $n$ to $n+1$ out of transitions from any $k$ to $k+1$ equals the frequency of transitions from $n+1$ to $n$ out of transitions from any $k+1$ to $k$, which implies that $d_n = a_n$. Therefore, *in steady-state, the system appears statistically identical to an arriving and a departing customer. When arrivals are Poisson,* we saw earlier that $a_n = p_n$; so, in this case, *both an arriving and a departing customer in steady-state see a system that is statistically identical to the one seen by an observer looking at the system at an arbitrary time.*

## 3.4 THE $M/M/m$, $M/M/\infty$, $M/M/m/m$, AND OTHER MARKOV SYSTEMS

We consider now a number of queueing systems that are similar to $M/M/1$ in that the arrival process is Poisson and the service times are independent, exponentially dis-

tributed, and independent of the interarrival times. Because of these assumptions, these systems can be modeled with continuous- or discrete-time Markov chains. From the corresponding state transition diagram, we can derive a set of equations that can be solved for the steady-state occupancy probabilities. Application of Little's Theorem then yields the average delay per customer.

### 3.4.1 $M/M/m$: The $m$-Server Case

The $M/M/m$ queueing system is identical to the $M/M/1$ system except that there are $m$ servers (or channels of a transmission line in a data communication context). A customer at the head of the queue is routed to any server that is available. The corresponding state transition diagram is shown in Fig. 3.9.

By writing down the global balance equations for the steady-state probabilities $p_n$ and taking $\delta \to 0$, we obtain

$$\begin{aligned} \lambda p_{n-1} &= n\,\mu\, p_n, \qquad n \le m \\ \lambda p_{n-1} &= m\,\mu\, p_n, \qquad n > m \end{aligned} \tag{3.33}$$

From these equations we obtain

$$p_n = \begin{cases} p_0 \dfrac{(m\rho)^n}{n!}, & n \le m \\[2ex] p_0 \dfrac{m^m \rho^n}{m!}, & n > m \end{cases} \tag{3.34}$$

where $\rho$ is given by

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

We can calculate $p_0$ using Eq. (3.34) and the condition $\sum_{n=0}^{\infty} p_n = 1$. We obtain

$$p_0 = \left[ 1 + \sum_{n=1}^{m-1} \frac{(m\rho)^n}{n!} + \sum_{n=m}^{\infty} \frac{(m\rho)^n}{m!} \frac{1}{m^{n-m}} \right]^{-1}$$

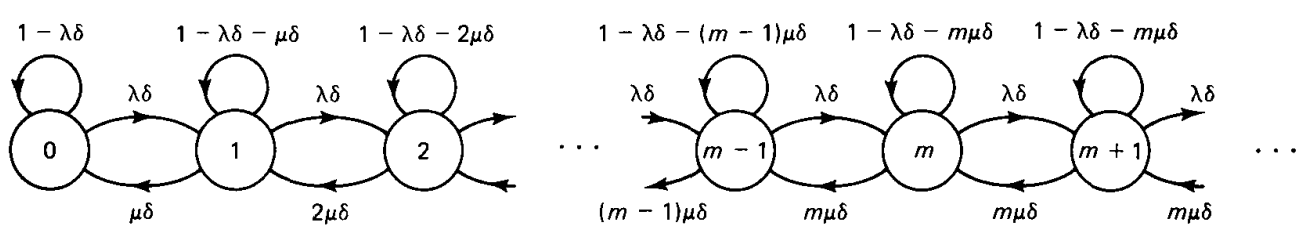



**Figure 3.9** Discrete-time Markov chain for the $M/M/m$ system.

and finally,

$$p_0 = \left[\sum_{n=0}^{m-1} \frac{(m\rho)^n}{n!} + \frac{(m\rho)^m}{m!(1-\rho)}\right]^{-1} \tag{3.35}$$

The probability that an arrival will find all servers busy and will be forced to wait in queue is an important measure of performance of the $M/M/m$ system. Since an arriving customer finds the system in "typical" state (see Section 3.3.2), we have

$$P\{\text{Queueing}\} = \sum_{n=m}^{\infty} p_n = \sum_{n=m}^{\infty} \frac{p_0 m^m \rho^n}{m!} = \frac{p_0 (m\rho)^m}{m!} \sum_{n=m}^{\infty} \rho^{n-m}$$

and, finally,

$$P_Q \triangleq P\{\text{Queueing}\} = \frac{p_0 (m\rho)^m}{m!(1-\rho)} \tag{3.36}$$

where $p_0$ is given by Eq. (3.35). This is known as the *Erlang C formula*, honoring Denmark's A. K. Erlang, the foremost pioneer of queueing theory. This equation is often used in telephony (and more generally in circuit switching systems) to estimate the probability of a call request finding all of the $m$ circuits of a transmission line busy. In an $M/M/m$ model it is assumed that such a call request "remains in queue," that is, continuously attempts to find a free circuit. The alternative model where such a call departs from the system and never returns is discussed in the context of the $M/M/m/m$ system in Section 3.4.3.

The expected number of customers waiting in queue (not in service) is given by

$$N_Q = \sum_{n=0}^{\infty} n p_{m+n}$$

Using the expression for $p_{m+n}$ [Eq. (3.34)], we obtain

$$N_Q = \sum_{n=0}^{\infty} n p_0 \frac{m^m \rho^{m+n}}{m!} = \frac{p_0 (m\rho)^m}{m!} \sum_{n=0}^{\infty} n\rho^n$$

Using the Erlang $C$ formula of Eq. (3.36) to express $p_0$ in terms of $P_Q$, and the equation $(1-\rho)\sum_{n=0}^{\infty} n\rho^n = \rho/(1-\rho)$ encountered in the $M/M/1$ system analysis, we finally obtain

$$N_Q = P_Q \frac{\rho}{1-\rho} \tag{3.37}$$

Note that

$$\frac{N_Q}{P_Q} = \frac{\rho}{1-\rho}$$

represents the expected number found in queue by an arriving customer conditioned on the fact that he is forced to wait in queue, and is independent of the number of servers for a given $\rho = \lambda/m\mu$. This suggests in particular that as long as there are customers waiting in queue, the queue size of the $M/M/m$ system behaves identically as in an $M/M/1$

system with service rate $m\mu$—the aggregate rate of the $m$ servers. Some thought shows that indeed this is true in view of the memoryless property of the exponential distribution.

Using Little's Theorem and the expression (3.37) for $N_Q$, we obtain the average time $W$ a customer has to wait in queue:

$$W = \frac{N_Q}{\lambda} = \frac{\rho P_Q}{\lambda(1-\rho)}$$

The average delay per customer is, therefore,

$$T = \frac{1}{\mu} + W = \frac{1}{\mu} + \frac{\rho P_Q}{\lambda(1-\rho)}$$

and using $\rho = \lambda/m\mu$, we obtain

$$T = \frac{1}{\mu} + W = \frac{1}{\mu} + \frac{P_Q}{m\mu - \lambda} \tag{3.38}$$

Using Little's Theorem again, the average number of customers in the system is

$$N = \lambda T = \frac{\lambda}{\mu} + \frac{\lambda P_Q}{m\mu - \lambda}$$

and using $\rho = \lambda/m\mu$, we finally obtain

$$N = m\rho + \frac{\rho P_Q}{1-\rho}$$

**Example 3.10 Using One vs. Using Multiple Channels in Statistical Multiplexing**

Consider a communication link serving $m$ independent Poisson traffic streams with overall rate $\lambda$. Suppose that the link is divided into $m$ separate channels with one channel assigned to each traffic stream. However, if a traffic stream has no packet awaiting transmission, its corresponding channel is used to transmit a packet of another traffic stream. The transmission times of packets on each of the channels are exponentially distributed with mean $1/\mu$. The system can be modeled by the same Markov chain as the $M/M/m$ queue. Let us compare the average delays per packet of this system, and an $M/M/1$ system with the same arrival rate $\lambda$ and service rate $m\mu$ (statistical multiplexing with one channel having $m$ times larger capacity). In the former case, the average delay per packet is given by the $M/M/m$ average delay expression (3.38)

$$T = \frac{1}{\mu} + \frac{P_Q}{m\mu - \lambda}$$

while in the latter case, the average delay per packet is

$$\hat{T} = \frac{1}{m\mu} + \frac{\hat{P}_Q}{m\mu - \lambda}$$

where $P_Q$ and $\hat{P}_Q$ denote the queueing probability in each case. When $\rho \ll 1$ (lightly loaded system) we have $P_Q \cong 0$, $\hat{P}_Q \cong 0$, and

$$\frac{T}{\hat{T}} \cong m$$

When $\rho$ is only slightly less than 1, we have $P_Q \cong 1$, $\hat{P}_Q \cong 1$, $1/\mu \ll 1/(m\mu - \lambda)$, and

$$\frac{T}{\hat{T}} \cong 1$$

Therefore, for a light load, statistical multiplexing with $m$ channels produces a delay almost $m$ times larger than the delay of statistical multiplexing with the $m$ channels combined in one (about the same as time- and frequency-division multiplexing). For a heavy load, the ratio of the two delays is close to 1.

### 3.4.2 $M/M/\infty$: The Infinite-Server Case

In the limiting case where $m = \infty$ in the $M/M/m$ system, we obtain from the global balance equations (3.33)

$$\lambda p_{n-1} = n\mu p_n , \qquad n = 1, 2, \ldots$$

so

$$p_n = p_0 \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!} , \qquad n = 1, 2, \ldots$$

From the condition $\sum_{n=0}^{\infty} p_n = 1$, we obtain

$$p_0 = \left[1 + \sum_{n=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}\right]^{-1} = e^{-\lambda/\mu}$$

so finally,

$$p_n = \left(\frac{\lambda}{\mu}\right)^n \frac{e^{-\lambda/\mu}}{n!} , \qquad n = 0, 1, \ldots$$

Therefore, in steady-state, *the number in the system is Poisson distributed with parameter* $\lambda/\mu$. The average number in the system is

$$N = \frac{\lambda}{\mu}$$

By Little's Theorem, the average delay is $N/\lambda$ or

$$T = \frac{1}{\mu}$$

This last equation can also be obtained simply by arguing that in an $M/M/\infty$ system, there is no waiting in queue, so $T$ equals the average service time $1/\mu$. It can be shown that the number in the system is Poisson distributed even if the service time distribution is not exponential (*i.e.*, in the $M/G/\infty$ system; see Problem 3.47).

**Example 3.11 The Quasistatic Assumption**

It is often convenient to assume that the external packet traffic entering a subnet node and destined for some other subnet node can be modeled by a stationary stochastic process that has a constant bit arrival rate (average bits/sec). This approximates a situation where the

arrival rate changes slowly with time and constitutes what we refer to as the quasistatic assumption.

When there are only a few active sessions (*i.e.*, user pairs) for the given origin–destination pair, this assumption is seriously violated since the addition or termination of a single session can change the total bit arrival rate by a substantial factor. When, however, there are many active sessions, each with a bit arrival rate that is small relative to the total, it seems plausible that the quasistatic assumption is approximately valid. The reason is that session additions are statistically counterbalanced by session terminations, with variations in the total rate being relatively small. For an analytical substantiation, assume that sessions are generated according to a Poisson process with rate $\lambda$, and terminate after a time which is exponentially distributed with mean $1/\mu$. Then the number of active sessions $n$ evolves like the number of customers in an $M/M/\infty$ system (*i.e.*, is Poisson distributed with parameter $\lambda/\mu$ in steady-state). In particular, the mean and standard deviation of $n$ are

$$N = E\{n\} = \frac{\lambda}{\mu}$$

$$\sigma_n = \left[E\left\{(n-N)^2\right\}\right]^{1/2} = \left(\frac{\lambda}{\mu}\right)^{1/2}$$

Suppose the $i^{\text{th}}$ active session generates traffic according to a stationary stochastic process having a bit arrival rate $\gamma_i$ bits/sec. Assume that the rates $\gamma_i$ are independent random variables with common mean $E\{\gamma_i\} = \Gamma$ and second moment $s_\gamma^2 = E\{\gamma_i^2\}$. Then the total bit arrival rate for $n$ active sessions is the random variable $f = \sum_{i=1}^{n} \gamma_i$, which has mean

$$F = E\{f\} = \frac{\lambda}{\mu}\Gamma$$

The standard deviation of $f$, denoted $\sigma_f$, can be obtained by writing

$$\sigma_f^2 = E\left\{\left(\sum_{i=1}^{n}\gamma_i\right)^2\right\} - F^2$$

and carrying out the corresponding calculations (Problem 3.28). The result is

$$\sigma_f = \left(\frac{\lambda}{\mu}\right)^{1/2} s_\gamma$$

Therefore, we have

$$\frac{\sigma_f}{F} = \left(\frac{s_\gamma}{\Gamma}\right)\left(\frac{\mu}{\lambda}\right)^{1/2}$$

Suppose now that the average bit rate $\Gamma$ of a session is small relative to the total $F$; that is, a "many-small-sessions assumption" holds. Then, since $\Gamma/F = \mu/\lambda$, we have that $\mu/\lambda$ is small. If we reasonably assume that $s_\gamma/\Gamma$ has a moderate value, it follows from the equation above that $\sigma_f/F$ is small. Therefore, the total arrival rate $f$ is approximately constant, thereby justifying the quasistatic assumption.

### 3.4.3 *M/M/m/m*: The *m*-Server Loss System

Consider a system which is identical to the $M/M/m$ system except that if an arrival finds all $m$ servers busy, it does not enter the system and is lost instead; the last $m$ in the

$M/M/m/m$ notation indicates the limit on the number of customers in the system. This model is in wide use in telephony (and also, more generally, in circuit switched networks). In this context, customers in the system correspond to active telephone conversations and the $m$ servers represent a single transmission line consisting of $m$ circuits. The average service time $1/\mu$ is the average duration of a telephone conversation. The principal quantity of interest here is the *blocking probability*, that is, the steady-state probability that all circuits are busy, in which case an arriving call is refused service. Note that in an $M/M/m/m$-based model, the assumption is that blocked calls are lost (not reattempted). This is in contrast with an $M/M/m$-based model, where the assumption is that blocked calls continuously reattempt admission into service. In data networks, the $M/M/m/m$ system can be used as a model where arrivals correspond to requests for virtual circuit connections between two nodes and the maximum number of virtual circuits allowed is $m$.

The corresponding state transition diagram is shown in Fig. 3.10. We have

$$\lambda p_{n-1} = n\mu p_n\,, \qquad n = 1, 2, \ldots, m$$

so

$$p_n = p_0 \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}\,, \qquad n = 1, 2, \ldots, m$$

Solving for $p_0$ in the equation $\sum_{n=0}^{m} p_n = 1$, we obtain

$$p_0 = \left[\sum_{n=0}^{m} \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}\right]^{-1}$$

The probability that an arrival will find all $m$ servers busy and will therefore be lost is

$$p_m = \frac{(\lambda/\mu)^m/m!}{\sum_{n=0}^{m}(\lambda/\mu)^n/n!}$$

This equation is known as the *Erlang B formula* and finds wide use in evaluating the blocking probability of telephone systems. It can be shown to hold even if the service time has mean $1/\mu$ but arbitrary probability distribution (*i.e.*, for an $M/G/m/m$ system; see [Ros83], p. 170).

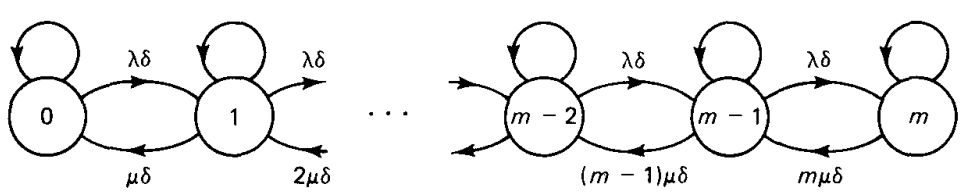



**Figure 3.10** Discrete-time Markov chain for the $M/M/m/m$ system.

### 3.4.4 Multidimensional Markov Chains—Applications in Circuit Switching

We have considered so far queueing systems with a single type of customer where the state can be described by the number of customers in the system. In some important systems there are several classes of customers, each with its own statistical characteristics for arrival and service, which cannot be lumped into a single class for the purpose of analysis. Here are some examples:

**Example 3.12 Two Session Classes in a Circuit Switching System**

Consider a transmission line consisting of $m$ independent circuits of equal capacity. There are two types of sessions arriving with Poisson rates $\lambda_1$ and $\lambda_2$, respectively. A session is blocked and lost for the system if all circuits are busy upon arrival, and is otherwise routed to any free circuit. The durations (or holding times) of the sessions of the two types are exponentially distributed with means $1/\mu_1$ and $1/\mu_2$. We are interested in finding the steady-state blocking probability for this system.

We first note that if $\mu_1 = \mu_2$, the two session types are indistinguishable for queueing purposes and the system can be modeled by an $M/M/m/m$ queue with arrival rate $\lambda_1 + \lambda_2$ and state equal to the total number of busy circuits. The desired blocking probability $p_m$ is then given by the Erlang $B$ formula of the preceding subsection. If, however, $\mu_1 \neq \mu_2$, then the total number of busy circuits does not fully specify the future statistical behavior of the queue; the number of each session type is also important since the duration of a session depends statistically on its type. Thus, the appropriate Markov chain model involves the two-dimensional state $(n_1, n_2)$, where $n_i$ is the number of circuits occupied by a session of type $i$, for $i = 1, 2$. The transition probability diagram for this chain is shown in Fig. 3.11. Generally, for multidimensional chains one may write the global balance equations for the stationary distribution

$$P(n_1, n_2), \qquad n_1 \geq 0,\ n_2 \geq 0,\ n_1 + n_2 \leq m$$

and try to solve them numerically. For this example, however, a closed-form expression is possible. We will demonstrate this shortly, once we develop the appropriate methodology.

**Example 3.13 Two-Class System with Preferential Treatment for One Class**

Consider the system of the preceding example with the difference that there is a limit $k < m$ on the number of circuits that can be used by sessions of the second type, so there are always $m - k$ circuits for use by sessions of the first type. The corresponding two-dimensional Markov chain is shown in Fig. 3.12. Note that here we should distinguish between the blocking probability for the first type of session, which is

$$\sum_{\{(n_1,n_2)|m-k\leq n_1\leq m, n_2=m-n_1\}} P(n_1, n_2)$$

and the blocking probability for the second type of session, which is

$$\sum_{\{(n_1,n_2)|0\leq n_1\leq m, n_2=\min\{k,m-n_1\}\}} P(n_1, n_2)$$

Again, it turns out that there is a closed-form expression for $P(n_1, n_2)$, as will be seen shortly.

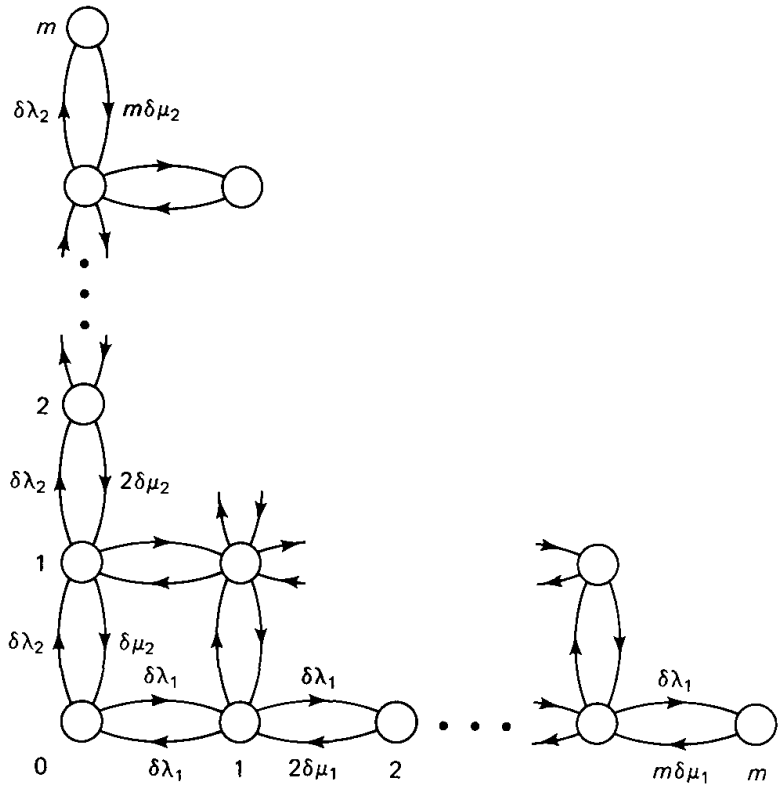



**Figure 3.11** Markov chain for the two-class queue of Example 3.12. To simplify the diagram, we do not show self-transitions and $o(\delta)$ transitions.

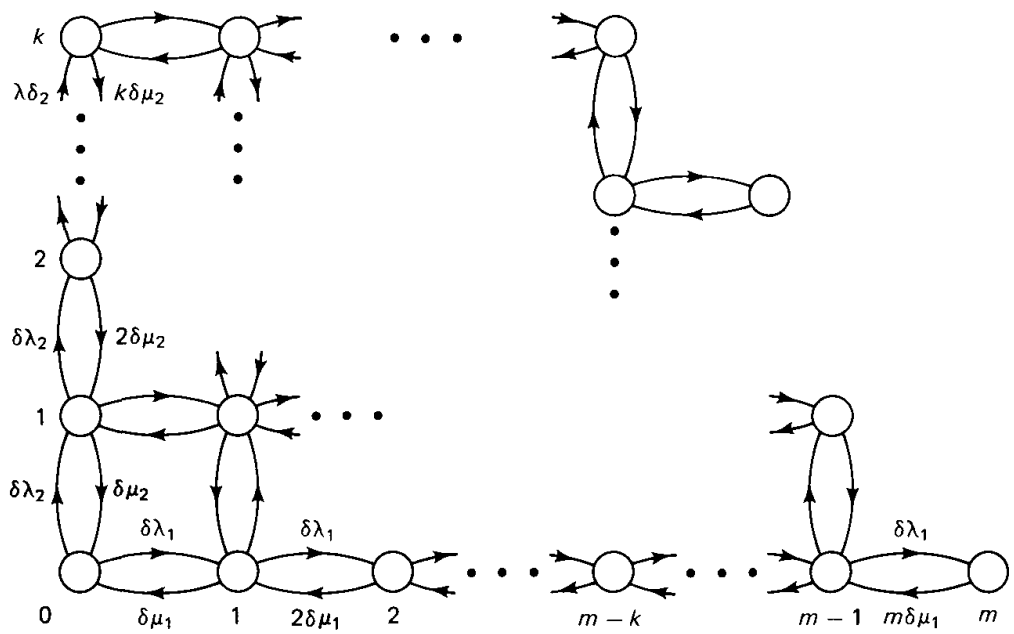



**Figure 3.12** Markov chain for the two class queue with preferential treatment for one class (cf. Example 3.13). Self-transitions and $o(\delta)$ transitions are not shown.

Multidimensional Markov chains usually involve $K$ customer types. Their states are of the form $(n_1, n_2, \ldots, n_K)$, where $n_i$ is the number of customers of type $i$ in the system. Such chains are usually harder to analyze than their one-dimensional counterparts, but in many interesting special cases one can obtain a closed-form solution for the

stationary distribution $P(n_1, n_2, \ldots, n_K)$. Important examples of properties that make this possible are:

1. The *detailed balance equations*

$$\lambda_i P(n_1, \ldots, n_{i-1}, n_i, n_{i+1}, \ldots, n_K) = \mu_i P(n_1, \ldots, n_{i-1}, n_i + 1, n_{i+1}, \ldots, n_K)$$

hold for all pairs of adjacent states

$$(n_1, \ldots, n_{i-1}, n_i, n_{i+1}, \ldots, n_K) \qquad \text{and} \qquad (n_1, \ldots, n_{i-1}, n_i + 1, n_{i+1}, \ldots, n_K)$$

where $\lambda_i$ and $\mu_i$ are the arrival rate and service rate, respectively, of the customers of type $i$. These equations imply that the frequency of transitions between any two adjacent states is the same in both directions (see Appendix A). We will explain in Section 3.7 that chains for which these equations hold are statistically indistinguishable when looked in forward and in reverse time, and for this reason they will be called *reversible*. Note that these equations hold for all the single-customer class systems discussed so far.

2. The stationary distribution can be expressed in *product form*, that is,

$$P(n_1, n_2, \ldots, n_K) = P_1(n_1) P_2(n_2) \cdots P_K(n_K)$$

where for each $i$, $P_i(n_i)$ is an expression depending only on the number $n_i$ of customers of type $i$. Several important types of networks of queues admit product form solutions, as will be seen in Section 3.8.

In this section we restrict ourselves to a class of multidimensional Markov chains, constructed from single-customer class systems using a process called *truncation*, for which we will see that both of the properties above hold.

**Truncation of independent single-class systems.** For a trivial example of a multidimensional Markov chain that admits a product form solution, consider $K$ independent $M/M/1$ queues. The number of customers in the $i^{\text{th}}$ queue has distribution

$$P_i(n_i) = \rho_i^{n_i}(1 - \rho_i)$$

where

$$\rho_i = \frac{\lambda_i}{\mu_i}$$

$\lambda_i$ and $\mu_i$ are the corresponding arrival and service rates, respectively, and we assume that $\rho_i < 1$ for all $i$. Since the $K$ queues are independent, we have the product form

$$P(n_1, n_2, \ldots, n_K) = P_1(n_1) P_2(n_2) \cdots P_K(n_K)$$

Note that the reasoning above would also apply if each of the $M/M/1$ queues were replaced by a birth–death type of queue (two successive states can differ only by a single unit: for example, the $M/M/m$, $M/M/\infty$, and $M/M/m/m$ queues). The only requirement is that the queues are independent.

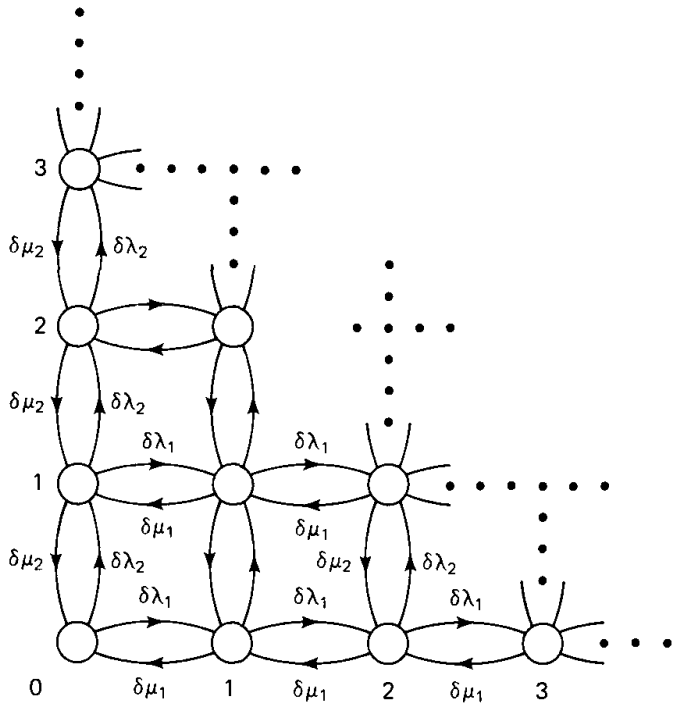



**Figure 3.13** Markov chain for a $K$ independent $M/M/1$ queues system. Self-transitions and $o(\delta)$ transitions are not shown ($K = 2$ in the figure).

Consider now the transition probability diagram of the $K$ independent $M/M/1$ queues system, shown in Fig. 3.13. A *truncation* of this system is a Markov chain having the same transition probability diagram with the only difference that some states have been eliminated, while transitions between all other pairs of states, together with their corresponding transition probabilities, have been left unchanged except for $O(\delta)$ terms (see Fig. 3.14). We require that the truncation is an irreducible Markov chain, that is, all states communicate with each other (see Appendix A).

We claim that the stationary distribution of this truncated system has the product form

$$P(n_1, n_2, \ldots, n_K) = \frac{\rho_1^{n_1} \rho_2^{n_2} \cdots \rho_K^{n_K}}{G} \tag{3.39}$$

where $G$ is a normalization constant guaranteeing that $P(n_1, n_2, \ldots, n_K)$ is a probability distribution, that is,

$$G = \sum_{(n_1, n_2, \ldots, n_K) \in S} \rho_1^{n_1} \rho_2^{n_2} \cdots \rho_K^{n_K} \tag{3.40}$$

where $S$ is the set of states of the truncated system.

To show this, we consider the detailed balance equations

$$\lambda_i P(n_1, \ldots, n_{i-1}, n_i, n_{i+1}, \ldots, n_K) = \mu_i P(n_1, \ldots, n_{i-1}, n_i + 1, n_{i+1}, \ldots, n_K)$$

By substituting the probabilities (3.39) in these equations, we obtain

$$\lambda_i \frac{\rho_1^{n_1} \cdots \rho_{i-1}^{n_{i-1}} \rho_i^{n_i} \rho_{i+1}^{n_{i+1}} \cdots \rho_K^{n_K}}{G} = \mu_i \frac{\rho_1^{n_1} \cdots \rho_{i-1}^{n_{i-1}} \rho_i^{n_i+1} \rho_{i+1}^{n_{i+1}} \cdots \rho_K^{n_K}}{G}$$

which holds as an identity in view of the definition $\rho_i = \lambda_i/\mu_i$. Therefore, the probability

distribution given by the expression (3.39) satisfies the detailed balance equations for the truncated chain, so it must be its unique stationary distribution (see Appendix A).

It should be noted here that there is a generic difficulty with product form solutions. To obtain the stationary distribution, one needs to compute the normalization constant $G$ of Eq. (3.40). For some systems, this involves a large amount of computation. An alternative to computing $G$ directly from Eq. (3.40) is to approximate it using Monte Carlo simulation. Here, a fairly large number of independent samples of $(n_1, \ldots, n_K)$ are generated using the distribution

$$P(n_1, \ldots, n_K) = \prod_{i=1}^{K} (1 - \rho_i)\rho_i^{n_i}$$

and $G$ is approximated by the proportion of samples that belong to the truncated space $S$. We will return to the computation of normalization constants in Section 3.8 in the context of queueing networks; see also Problem 3.51.

The reasoning above can also be used to show that there is a product form solution for any truncation of a system consisting of $K$ independent queues each described by a birth–death Markov chain, such as the $M/M/m$, $M/M/\infty$, and $M/M/m/m$ systems. For example, it is straightforward to verify that the stationary distribution of the $K$ independent $M/M/\infty$ queues system is given by

$$P(n_1, n_2, \ldots, n_K) = \frac{\frac{\rho_1^{n_1}}{n_1!}\frac{\rho_2^{n_2}}{n_2!}\cdots\frac{\rho_K^{n_K}}{n_K!}}{G}$$

where $G$ is a normalization constant,

$$G = \sum_{(n_1, n_2, \ldots, n_K) \in S} \frac{\rho_1^{n_1}}{n_1!}\frac{\rho_2^{n_2}}{n_2!}\cdots\frac{\rho_K^{n_K}}{n_K!}$$

and $S$ is the set of states of the truncated chain.

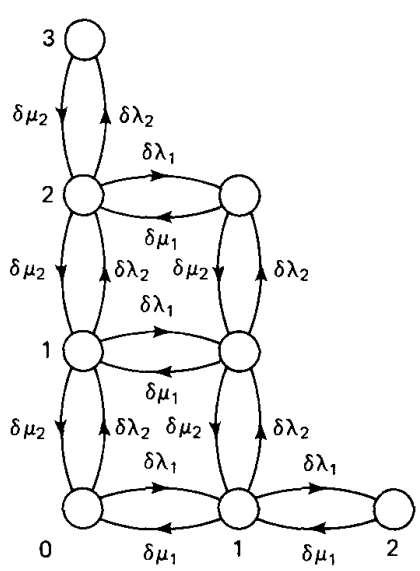



**Figure 3.14** Example of a Markov chain which is a truncation of the $K$ independent $M/M/1$ queues system ($K = 2$ in the figure).

**Blocking probabilities for circuit switching systems.** Using the product form solution (3.39), it is straightforward to write closed-form expressions for the blocking probabilities of the circuit switching systems with two session classes of Examples 3.12 and 3.13. The two-dimensional chains of these examples are truncations of the two independent $M/M/\infty$ queues system (Figs. 3.11 to 3.13). Thus, in the case of Example 3.13, the blocking probability for the first type of session is

$$\sum_{\{(n_1,n_2)|m-k\le n_1\le m, n_2=m-n_1\}} P(n_1,n_2)$$

$$= \frac{\sum_{n_1=k}^{m} \frac{\rho_1^{n_1}}{n_1!}\frac{\rho_2^{m-n_1}}{m-n_1!}}{\sum_{n_1=0}^{k}\sum_{n_2=0}^{k}\frac{\rho_1^{n_1}}{n_1!}\frac{\rho_2^{n_2}}{n_2!} + \sum_{n_1=k+1}^{m}\sum_{n_2=0}^{m-n_1}\frac{\rho_1^{n_1}}{n_1!}\frac{\rho_2^{n_2}}{n_2!}}$$

The blocking probability for the second type of session is

$$\sum_{\{(n_1,n_2)|0\le n_1\le m, n_2=\min\{k,m-n_1\}\}} P(n_1,n_2)$$

$$= \frac{\sum_{n_1=0}^{k}\frac{\rho_1^{n_1}}{n_1!}\frac{\rho_2^{k-n_1}}{m-n_1!} + \sum_{n_1=k+1}^{m}\frac{\rho_1^{n_1}}{n_1!}\frac{\rho_2^{m-n_1}}{m-n_1!}}{\sum_{n_1=0}^{k}\sum_{n_2=0}^{k}\frac{\rho_1^{n_1}}{n_1!}\frac{\rho_2^{n_2}}{n_2!} + \sum_{n_1=k+1}^{m}\sum_{n_2=0}^{m-n_1}\frac{\rho_1^{n_1}}{n_1!}\frac{\rho_2^{n_2}}{n_2!}}$$

The following important example illustrates the wide applicability of product form solutions in circuit switching networks.

**Example 3.14 Circuit Switching Networks with Fixed Routing**

Consider a network of transmission lines shared by sessions of $K$ different types (see Fig. 3.15). Sessions of type $i$ arrive according to a Poisson rate $\lambda_i$ and have an exponentially distributed holding time with mean $1/\mu_i$.

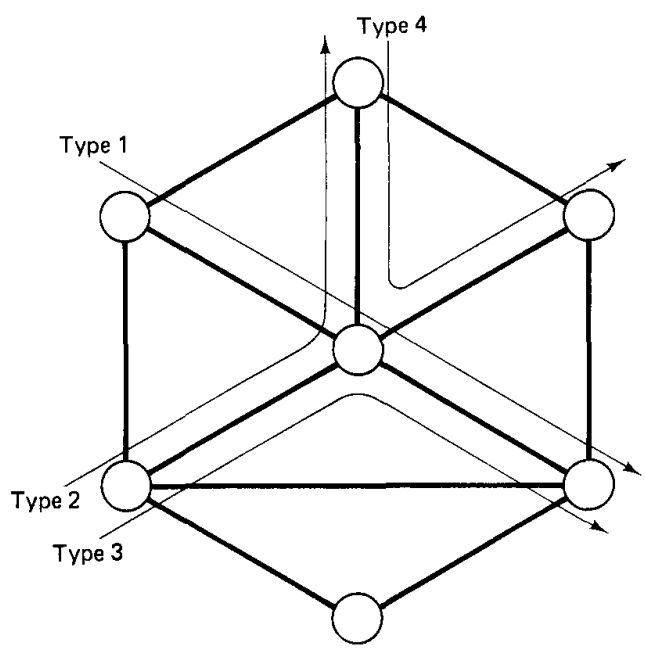



**Figure 3.15** Model of a circuit switching network. There are $K$ different session types. All sessions of the same type go over the same path and reserve the same amount of transmission capacity on each link of their path. A session is blocked if some link on its path is loaded to the point that it cannot accommodate the transmission capacity of the session.

We assume that all sessions of a given type $i$ traverse the same set of network links (fixed routing) and reserve a fixed amount $b_i$ of transmission capacity at each link. Thus if $C_j$ is the transmission capacity of a link $j$ and $I(j)$ is the set of session types using this link, we must have

$$\sum_{i \in I(j)} b_i n_i \leq C_j$$

where $n_i$ is the number of sessions of type $i$ in the network. A session of a given type $m$ is blocked from entering the network (and is assumed lost to the system) if upon arrival it finds that it cannot be accommodated due to insufficient link capacity, that is,

$$b_m + \sum_{i \in I(j)} b_i n_i > C_j$$

for some link $j$ that the session must traverse.

The quality of service of this system may be described by the blocking probabilities for the different session types. To obtain these probabilities, we model the system as a truncation of the $K$ independent $M/M/\infty$ queues system. The truncated chain is the same as for the latter system, except that all states $(n_1, n_2, \ldots, n_K)$ for which the inequality $\sum_{i \in I(j)} b_i n_i \leq C_j$ is violated for some link $j$ have been eliminated. The stationary distribution has a product form, which yields the desired blocking probabilities.

A remarkable fact about the product form solution of this example is that it is valid for a broad class of holding time distributions that includes the exponential as a special case (see [BLL84] and [Kau81]).

## 3.5 THE $M/G/1$ SYSTEM

Consider a single-server queueing system where customers arrive according to a Poisson process with rate $\lambda$, but the customer service times have a general distribution—not necessarily exponential as in the $M/M/1$ system. Suppose that customers are served in the order they arrive and that $X_i$ is the service time of the $i^{\text{th}}$ arrival. We assume that the random variables $(X_1, X_2, \ldots)$ are identically distributed, mutually independent, and independent of the interarrival times.

Let

$$\overline{X} = E\{X\} = \frac{1}{\mu} = \text{ Average service time}$$

$$\overline{X^2} = E\{X^2\} = \text{ Second moment of service time}$$

Our objective is to derive and understand the *Pollaczek–Khinchin (P-K) formula:*

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho)} \tag{3.41}$$

where $W$ is the expected customer waiting time in queue and $\rho = \lambda/\mu = \lambda\overline{X}$. Given the P-K formula (3.41), the total waiting time, in queue and in service, is

$$T = \overline{X} + \frac{\lambda \overline{X^2}}{2(1-\rho)} \tag{3.42}$$

Applying Little's formula to $W$ and $T$, we get the expected number of customers in the queue $N_Q$ and the expected number in the system $N$:

$$N_Q = \frac{\lambda^2 \overline{X^2}}{2(1-\rho)} \tag{3.43}$$

$$N = \rho + \frac{\lambda^2 \overline{X^2}}{2(1-\rho)} \tag{3.44}$$

For example, when service times are exponentially distributed, as in the $M/M/1$ system, we have $\overline{X^2} = 2/\mu^2$, and the P-K formula (3.41) reduces to the equation (see Section 3.3.2)

$$W = \frac{\rho}{\mu(1-\rho)} \qquad (M/M/1)$$

When service times are identical for all customers (the $M/D/1$ system, where $D$ means deterministic), we have $\overline{X^2} = 1/\mu^2$, and

$$W = \frac{\rho}{2\mu(1-\rho)} \qquad (M/D/1) \tag{3.45}$$

Since the $M/D/1$ case yields the minimum possible value of $\overline{X^2}$ for given $\mu$, it follows that the values of $W$, $T$, $N_Q$, and $N$ for an $M/D/1$ queue are lower bounds to the corresponding quantities for an $M/G/1$ queue of the same $\lambda$ and $\mu$. It is interesting to note that $W$ and $N_Q$ for the $M/D/1$ queue are exactly one half their values for the $M/M/1$ queue of the same $\lambda$ and $\mu$. The values of $T$ and $N$ for $M/D/1$, on the other hand, range from the same as $M/M/1$ for small $\rho$ to one half of $M/M/1$ as $\rho$ approaches 1. The reason is that the expected service time is the same in the two cases, and for $\rho$ small, most of the waiting occurs in service, whereas for $\rho$ large, most of the waiting occurs in the queue.

We provide a proof of the Pollaczek–Khinchin formula based on the concept of the *mean residual service time*. This same concept will prove useful in a number of subsequent developments. One example is $M/G/1$ queues with priorities. Another is reservation systems where part of the service time is occupied with sending packets (*i.e.*, serving customers), and part with sending control information or making reservations for sending the packets.

Denote

$W_i$ = Waiting time in queue of the $i$th customer

$R_i$ = Residual service time seen by the $i$th customer. By this we mean that if customer $j$ is already being served when $i$ arrives, $R_i$ is the remaining time until customer $j$'s service time is complete. If no customer is in service (*i.e.*, the system is empty when $i$ arrives), then $R_i$ is zero

$X_i$ = Service time of the $i$th customer

$N_i$ = Number of customers found waiting in queue by the $i$th customer upon arrival

We have

$$W_i = R_i + \sum_{j=i-N_i}^{i-1} X_j$$

By taking expectations and using the independence of the random variables $N_i$ and $X_{i-1}, \ldots, X_{i-N_i}$, we have

$$E\{W_i\} = E\{R_i\} + E\left\{\sum_{j=i-N_i}^{i-1} E\{X_j \mid N_i\}\right\} = E\{R_i\} + \overline{X}E\{N_i\}$$

Taking the limit as $i \to \infty$, we obtain

$$W = R + \frac{1}{\mu} N_Q \tag{3.46}$$

where

$$R = \text{ Mean residual time, defined as } R = \lim_{i\to\infty} E\{R_i\}.$$

In Eq. (3.46) (and throughout this section) all long-term average quantities should be viewed as limits when time or customer index increases to infinity. Thus, $W$, $R$, and $N_Q$ are limits (as $i \to \infty$) of the average waiting time, residual time, and number found in queue, respectively, corresponding to the $i^{\text{th}}$ customer. We assume that these limits exist, and this is true of almost all systems of interest to us provided that $\lambda < \mu$. Note that in the waiting time equation (3.46), the average number in queue $N_Q$ and the mean residual time $R$ as seen by an arriving customer are also equal to the average number in queue and mean residual time seen by an outside observer at a random time. This is due to the Poisson character of the arrival process, which implies that the occupancy distribution upon arrival is typical (see Section 3.3.2).

By Little's Theorem, we have

$$N_Q = \lambda W$$

and by substitution in the waiting time formula (3.46), we obtain

$$W = R + \rho W \tag{3.47}$$

where $\rho = \lambda/\mu$ is the utilization factor; so, finally,

$$W = \frac{R}{1-\rho} \tag{3.48}$$

We can calculate $R$ by a graphical argument. In Fig. 3.16 we plot the residual service time $r(\tau)$ (*i.e.*, the remaining time for completion of the customer in service at time $\tau$) as a function of $\tau$. Note that when a new service of duration $X$ begins, $r(\tau)$ starts at $X$ and decays linearly for $X$ time units. Consider a time $t$ for which $r(t) = 0$. The time average of $r(\tau)$ in the interval $[0, t]$ is

$$\frac{1}{t}\int_0^t r(\tau)\,d\tau = \frac{1}{t}\sum_{i=1}^{M(t)} \frac{1}{2}X_i^2 \tag{3.49}$$

where $M(t)$ is the number of service completions within $[0, t]$, and $X_i$ is the service time of the $i^{\text{th}}$ customer. We can also write this equation as

$$\frac{1}{t}\int_0^t r(\tau)\,d\tau = \frac{1}{2}\frac{M(t)}{t}\frac{\sum_{i=1}^{M(t)} X_i^2}{M(t)} \tag{3.50}$$

and assuming the limits below exist, we obtain

$$\lim_{t\to\infty}\frac{1}{t}\int_0^t r(\tau)\,d\tau = \frac{1}{2}\lim_{t\to\infty}\frac{M(t)}{t}\cdot\lim_{t\to\infty}\frac{\sum_{i=1}^{M(t)} X_i^2}{M(t)} \tag{3.51}$$

The two limits on the right are the time averages of the departure rate (which equals the arrival rate) and the second moment of the service time, respectively, while the limit on the left is the time average of the residual time. Assuming that time averages can be replaced by ensemble averages, we obtain

$$R = \frac{1}{2}\lambda\overline{X^2} \tag{3.52}$$

The P-K formula,

$$W = \frac{\lambda\overline{X^2}}{2(1-\rho)} \tag{3.53}$$

now follows by substituting the expression obtained for $R$ [cf. Eq. (3.52)] into the formula $W = R/(1-\rho)$ [cf. Eq. (3.48)].

Note that our derivation was based on two assumptions:

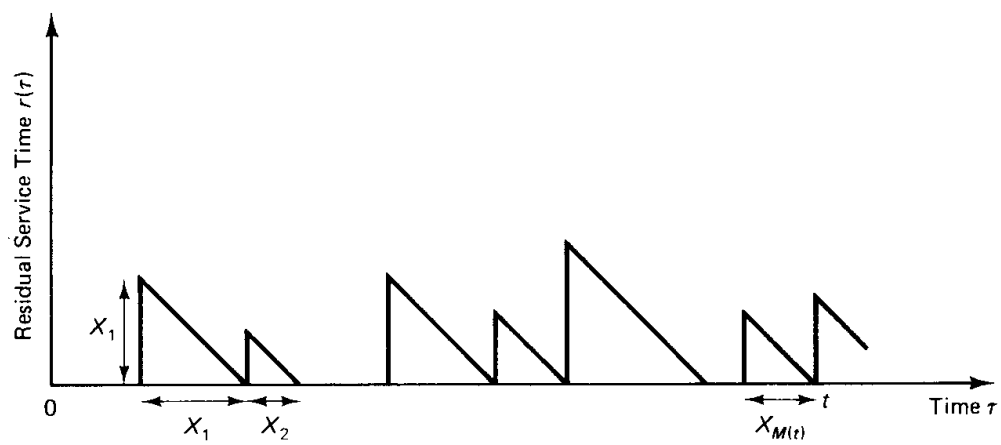



**Figure 3.16** Derivation of the mean residual service time. During period $[0, t]$, the time average of the residual service time $r(\tau)$ is

$$\frac{1}{t}\int_0^t r(\tau)\;d\tau = \frac{1}{t}\sum_{i=1}^{M(t)}\frac{1}{2}X_i^2 = \frac{1}{2}\frac{M(t)}{t}\frac{\sum_{i=1}^{M(t)} X_i^2}{M(t)}$$

where $X_i$ is the service time of the $i^{\text{th}}$ customer, and $M(t)$ is the number of service completions in $[0, t]$. Taking the limit as $t \to \infty$ and equating time and ensemble averages, we obtain the mean residual time $R = (1/2)\lambda\overline{X^2}$.

1. The existence of the steady-state averages $W$, $R$, and $N_Q$
2. The equality (with probability one) of the long-term time averages appearing in Eq. (3.51) with the corresponding ensemble averages

These assumptions can be justified by careful applications of the law of large numbers, but the details are beyond the scope of this book. However, these are natural assumptions for the systems of interest to us, and we will base similar derivations on graphical arguments and interchange of time averages with ensemble averages without further discussion.

One curious feature of the P-K formula (3.53) is that an $M/G/1$ queue can have $\rho < 1$ but infinite $W$ if the second moment $\overline{X^2}$ is $\infty$. What is happening in this case is that a small fraction of customers have incredibly long service times. When one of these customers is served, an incredible number of arrivals are queued and delayed by a significant fraction of that long service time. Thus, the contribution to $W$ is proportional to the square of the service time, leading to an infinite $W$ if $\overline{X^2}$ is infinite.

The derivation of the P-K formula above assumed that customers were served in order of arrival, that is, that the number of customers served between the $i^{\text{th}}$ arrival and service is just the number in queue at the $i^{\text{th}}$ arrival. It turns out, however, that this formula is valid for any order of servicing customers as long as the order is determined independently of the required service time. To see this, suppose that the $i^{\text{th}}$ and $j^{\text{th}}$ customers are both in the queue and that they exchange places. The expected queueing time of customer $i$ (over the service times of the customers in queue) will then be exchanged with that for customer $j$, but the average, over all customers, is unchanged. Since any service order can be considered as a sequence of reversals in queue position, the P-K formula remains valid (see also Problem 3.32).

To see why the P-K formula is invalid if the service order can depend on service time, consider a queue with two customers requiring 10 and 1 units of service time, respectively. Assuming that the server becomes available at time 0, serving the first customer first results in one customer starting service at time 0 and the other at time 10. Serving the second customer first results in one customer starting at time 0 and the other at time 1. Thus, the average queueing time over the two customers is 5 in the first case and 0.5 in the second case. Clearly, queueing time is reduced by serving customers with small service time first. For this situation, the derivation of the P-K formula breaks down at Eq. (3.46) since the customers that will be served before a newly arriving customer no longer have a mean service time equal to $1/\mu$.

**Example 3.15 Delay Analysis of an ARQ System**

Consider a go back $n$ ARQ system such as the one discussed in Section 2.4. Assume that packets are transmitted in frames that are one time unit long, and there is a maximum wait for an acknowledgment of $n-1$ frames before a packet is retransmitted (see Fig. 3.17). In this system packets are retransmitted for two reasons:

1. A given packet transmitted in frame $i$ might be rejected at the receiver due to errors, in which case the transmitter will transmit packets in frames $i+1, i+2, \ldots, i+n-1$, (if any are available), and then go back to retransmit the given packet in frame $i+n$.

2. A packet transmitted in frame $i$ might be accepted at the receiver, but the corresponding acknowledgment (in the form of the receive number) might not have arrived at the transmitter by the time the transmission of packet $i + n - 1$ is completed. This can happen due to errors in the return channel, large propagation delays, long return frames relative to the size of the goback number $n$, or a combination thereof.

We will assume (somewhat unrealistically) that retransmissions occur only due to reason 1, and that a packet is rejected at the receiver with probability $p$ independently of other packets.

Consider the case where packets arrive at the transmitter according to a Poisson process with rate $\lambda$. It follows that the time interval between start of the first transmission of a given packet after the last transmission of the previous packet and end of the last transmission of the given packet is $1 + kn$ time units with probability $(1 - p)p^k$. (This corresponds to $k$ retransmissions following the last transmission of the previous packet; see Fig. 3.17.) Thus, the transmitter's queue behaves like an $M/G/1$ queue with service time distribution given by

$$P\{X = 1 + kn\} = (1 - p)p^k, \quad k = 0, 1, \ldots$$

The first two moments of the service time are

$$\overline{X} = \sum_{k=0}^{\infty}(1 + kn)(1 - p)p^k = (1 - p)\left(\sum_{k=0}^{\infty} p^k + n\sum_{k=0}^{\infty} kp^k\right)$$

$$\overline{X^2} = \sum_{k=0}^{\infty}(1 + kn)^2(1 - p)p^k = (1 - p)\left(\sum_{k=0}^{\infty} p^k + 2n\sum_{k=0}^{\infty} kp^k + n^2\sum_{k=0}^{\infty} k^2p^k\right)$$

We now note that

$$\sum_{k=0}^{\infty} p^k = \frac{1}{1 - p}, \qquad \sum_{k=0}^{\infty} kp^k = \frac{p}{(1 - p)^2}, \qquad \sum_{k=0}^{\infty} k^2p^k = \frac{p + p^2}{(1 - p)^3}$$

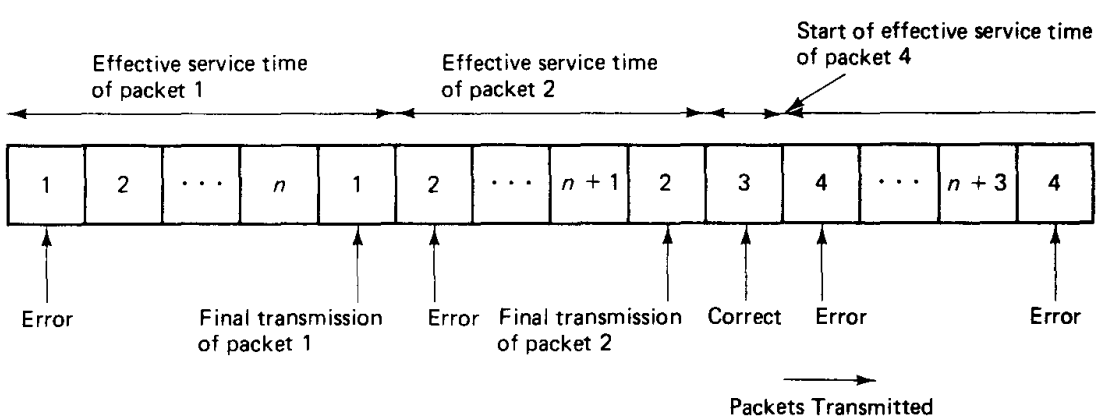



**Figure 3.17** Illustration of the effective service times of packets in the ARQ system of Example 3.15. For example, packet 2 has an effective service time of $n + 1$ because there was an error in the first attempt to transmit it following the last transmission of packet 1, but no error in the second attempt.

(The first sum is the usual geometric series sum, while the other two sums are obtained by differentiating the first sum twice.) Using these formulas in the equations for $\overline{X}$ and $\overline{X^2}$ above, we obtain

$$\overline{X} = 1 + \frac{np}{1-p}$$

$$\overline{X^2} = 1 + \frac{2np}{1-p} + \frac{n^2(p+p^2)}{(1-p)^2}$$

The P-K formula gives the average packet time in queue and in the system (up to the end of the last transmission):

$$W = \frac{\lambda\overline{X^2}}{2(1-\lambda\overline{X})}$$

$$T = \overline{X} + W$$

### 3.5.1 $M/G/1$ Queues with Vacations

Suppose that at the end of each busy period, the server goes on "vacation" for some random interval of time. Thus, a new arrival to an idle system, rather than going into service immediately, waits for the end of the vacation period (see Fig. 3.18). If the system is still idle at the completion of a vacation, a new vacation starts immediately. For data networks, vacations correspond to the transmission of various kinds of control and record-keeping packets when there is a lull in the data traffic; other applications will become apparent later.

Let $V_1, V_2, \ldots$ be the durations of the successive vacations taken by the server. We assume that $V_1, V_2, \ldots$ are independent and identically distributed (IID) random variables, also independent of the customer interarrival times and service times. As before, the arrivals are Poisson and the service times are IID with a general distribution. A new arrival to the system has to wait in the queue for the completion of the current service or vacation and then for the service of all the customers waiting before it. Thus, the waiting time formula $W = R/(1-\rho)$ is still valid [cf. Eq. (3.48)], where now $R$ is the mean residual time for completion of the service *or* vacation in process when the $i^{\text{th}}$ customer arrives.

The analysis of this new system is the same as that of the P-K formula except that vacations must be included in the graph of residual service times $r(\tau)$ (see Fig. 3.19). Let $M(t)$ be the number of services completed by time $t$ and $L(t)$ be the number of vacations completed by time $t$. Then [as in Eq. (3.49)], for any $t$ where a service or vacation is just completed, we have

$$\frac{1}{t}\int_0^t r(\tau)\,d\tau = \frac{1}{t}\sum_{i=1}^{M(t)}\frac{1}{2}X_i^2 + \frac{1}{t}\sum_{i=1}^{L(t)}\frac{1}{2}V_i^2 = \frac{M(t)}{t}\frac{\sum_{i=1}^{M(t)}\frac{1}{2}X_i^2}{M(t)} + \frac{L(t)}{t}\frac{\sum_{i=1}^{L(t)}\frac{1}{2}V_i^2}{L(t)} \tag{3.54}$$

As before, assuming that a steady-state exists, $M(t)/t$ approaches $\lambda$ with increasing $t$, and the first term on the right side of Eq. (3.54) approaches $\lambda\overline{X^2}/2$ as in the derivation

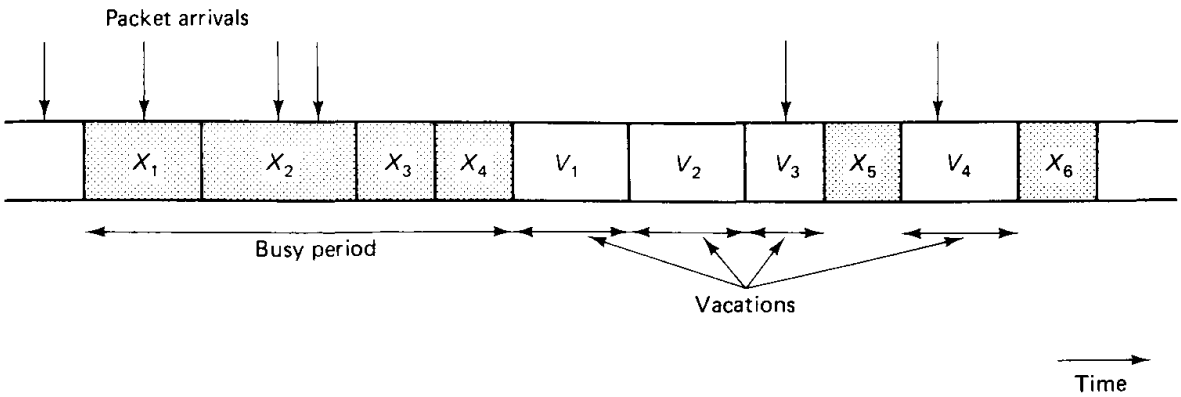



**Figure 3.18** $M/G/1$ system with vacations. At the end of a busy period, the server goes on vacation for time $V$ with first and second moments $\overline{V}$ and $\overline{V^2}$, respectively. If the system is empty at the end of a vacation, the server takes a new vacation. An arriving customer to an empty system must wait until the end of the current vacation to get service.

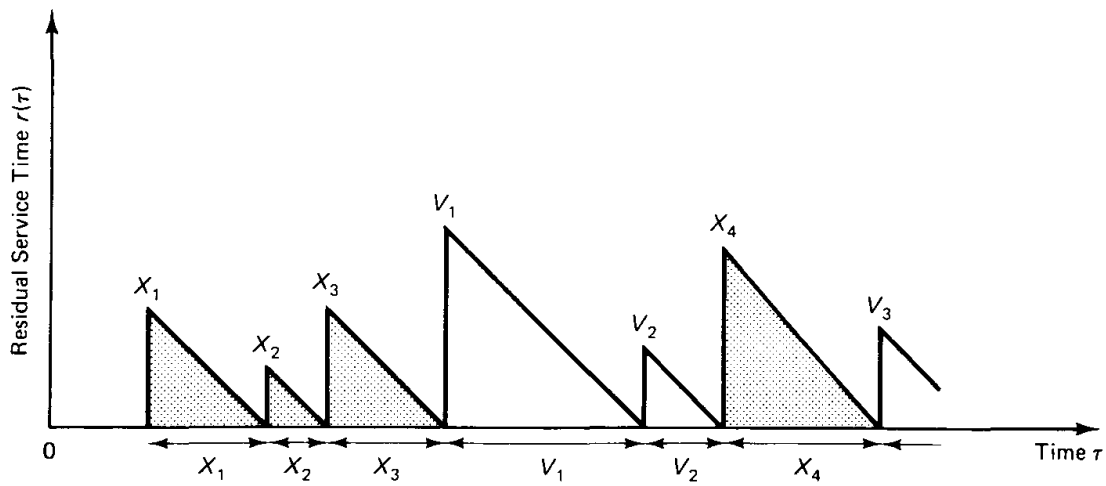



**Figure 3.19** Residual service times for an $M/G/1$ system with vacations. Busy periods alternate with vacation periods. If $M(t)$ and $L(t)$ are the numbers of services and vacations completed by time $t$, respectively, and $t$ is a time of completion of a service or a vacation, we have

$$\frac{1}{t}\int_0^t r(\tau)\,d\tau = \frac{1}{t}\sum_{i=1}^{M(t)}\frac{1}{2}X_i^2 + \frac{1}{t}\sum_{i=1}^{L(t)}\frac{1}{2}V_i^2 = \frac{M(t)}{t}\,\frac{\sum_{i=1}^{M(t)}\frac{1}{2}X_i^2}{M(t)} + \frac{L(t)}{t}\,\frac{\sum_{i=1}^{L(t)}\frac{1}{2}V_i^2}{L(t)}$$

Taking limit as $t \to \infty$ and arguing that $M(t)/t \to \lambda$ and $L(t)/t \to (1-\rho)/\overline{V}$, we obtain the mean residual time $R = \frac{\lambda\overline{X^2}}{2} + \frac{(1-\rho)\overline{V^2}}{2\overline{V}}$

of the P-K formula [cf. Eq. (3.52)]. For the second term, note that as $t \to \infty$, the fraction of time spent serving customers approaches $\rho$, and thus the fraction of time occupied with vacations is $1-\rho$. Assuming that time averages can be replaced by ensemble averages, we have $t(1-\rho)/L(t) \to \overline{V}$ with increasing $t$, and thus the second term in Eq. (3.54) approaches $(1-\rho)\overline{V^2}/(2\overline{V})$, where $\overline{V}$ and $\overline{V^2}$ are the first and second moments of the vacation interval, respectively. Combining this with $W = R/(1-\rho)$, and assuming

equality of the time and ensemble averages of $R$, we get

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho)} + \frac{\overline{V^2}}{2\overline{V}} \tag{3.55}$$

as the expected waiting time in queue for an $M/G/1$ system with vacations.

If we look carefully at the derivation of the preceding equation, we see that the mutual independence of the vacation intervals is not required (although the time and ensemble averages of the vacation intervals must still be equal) and the length of a vacation interval need not be independent of the customer arrival and service times. Naturally, with this kind of dependence, it becomes more difficult to calculate $\overline{V}$ and $\overline{V^2}$, as these quantities might be functions of the underlying $M/G/1$ process.

**Example 3.16 Frequency- and Time-Division Multiplexing on a Slot Basis**

We have $m$ traffic streams of equal-length packets arriving according to a Poisson process with rate $\lambda/m$ each. If the traffic streams are frequency-division multiplexed on $m$ subchannels of an available channel, the transmission time of each packet is $m$ time units. Then, each subchannel can be represented by an $M/D/1$ queueing system and the $M/D/1$ formula $W = \rho/(2\mu(1-\rho))$ [cf. Eq. (3.45)] with $\rho = \lambda$, $\mu = 1/m$, gives the average queueing delay per packet,

$$W_{\text{FDM}} = \frac{\lambda m}{2(1-\lambda)} \tag{3.56}$$

Consider the same FDM scheme with the difference that packet transmissions can start only at times $m, 2m, 3m, \ldots$ (*i.e.*, at the beginning of a slot of $m$ time units). We call this scheme *slotted frequency-division multiplexing* (SFDM), and note that it can be viewed as an $M/D/1$ queue with vacations. When there are no packets in the queue for a given stream at the beginning of a slot, the server takes a vacation for one slot, or $m$ time units. Thus, $\overline{V} = m$, $\overline{V^2} = m^2$, and the vacation system waiting time formula (3.55) becomes

$$W_{\text{SFDM}} = W_{\text{FDM}} + \frac{m}{2} \tag{3.57}$$

Finally, consider the case where the $m$ traffic streams are time-division multiplexed in a scheme whereby the time axis is divided in $m$-slot frames with one slot dedicated to each traffic stream (see Fig. 3.20). Each slot is one time unit long and can carry a single packet. Then, if we compare this TDM scheme with the SFDM scheme, we see that the queue for a given stream in TDM is precisely the same as the queue for SFDM, and

$$W_{\text{TDM}} = W_{\text{SFDM}} = W_{\text{FDM}} + \frac{m}{2} = \frac{m}{2(1-\lambda)} \tag{3.58}$$

If we now look at the total delay for TDM, we get a different picture, since the service time is 1 unit of time rather than $m$ units as in SFDM. By adding the service times to the queueing delays, we obtain

$$\begin{aligned} T_{\text{FDM}} &= m + \frac{\lambda m}{2(1-\lambda)} \\ T_{\text{SFDM}} &= T_{\text{FDM}} + \frac{m}{2} \\ T_{\text{TDM}} &= 1 + \frac{m}{2(1-\lambda)} = T_{\text{FDM}} - \left(\frac{m}{2} - 1\right) \end{aligned} \tag{3.59}$$

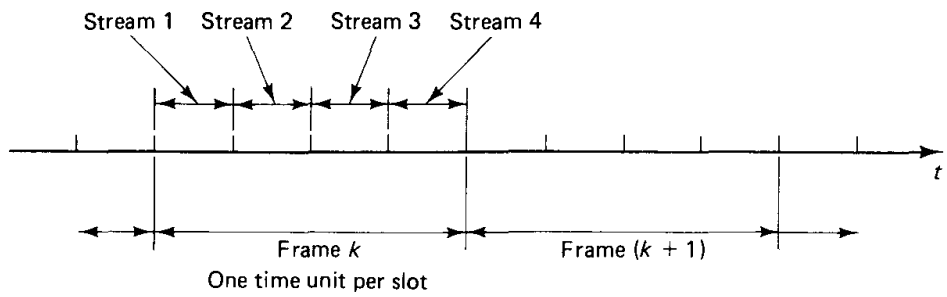



**Figure 3.20** TDM with $m = 4$ traffic streams.

Thus, the customer's average total delay is more favorable in TDM than in FDM (assuming that $m > 2$). The longer average waiting time in queue for TDM is more than compensated by the faster service time. Contrast this with the Example 3.9, which treats TDM with slots that are a very small portion of the packet size. Problem 3.33 outlines an alternative approach for deriving the TDM average delay.

### 3.5.2 Reservations and Polling

Organizing transmissions from several packet streams into a statistical multiplexing system requires some form of scheduling. In some cases, this scheduling is naturally and easily accomplished; in other cases, however, some form of reservation or polling system is required.

Situations of this type arise often in multiaccess channels, which will be treated extensively in Chapter 4. For a typical example, consider a communication channel that can be accessed by several spatially separated users; however, only one user can transmit successfully on the channel at any one time. The communication resource of the channel can be divided over time into a portion used for packet transmissions and another portion used for reservation or polling messages that coordinate the packet transmissions. In other words, the time axis is divided into *data intervals*, where actual data are transmitted, and *reservation intervals*, used for scheduling future data. For uniform presentation, we use the term "reservation" even though "polling" may be more appropriate to the practical situation.

We will consider $m$ traffic streams (also called users) and assume that each data interval contains packets of a *single* user. Reservations for these packets are made in the immediately preceding reservation interval. All users are taken up in cyclic order (see Fig. 3.21). There are several versions of this system differing in the rule for deciding which packets are transmitted during the data interval of each user. In the *gated* system, the rule is that only those packets that arrived prior to the user's preceding reservation interval are transmitted. By contrast, in the *exhaustive* system, the rule is that all available packets of a user are transmitted during the corresponding data interval, including those that arrived in this data interval or the preceding reservation interval. An intermediate version, which we call the *partially gated* system, results when the packets transmitted in a user's data interval are those that arrived up to the time this data interval began (and the corresponding reservation interval ended). A typical example of such reservation systems

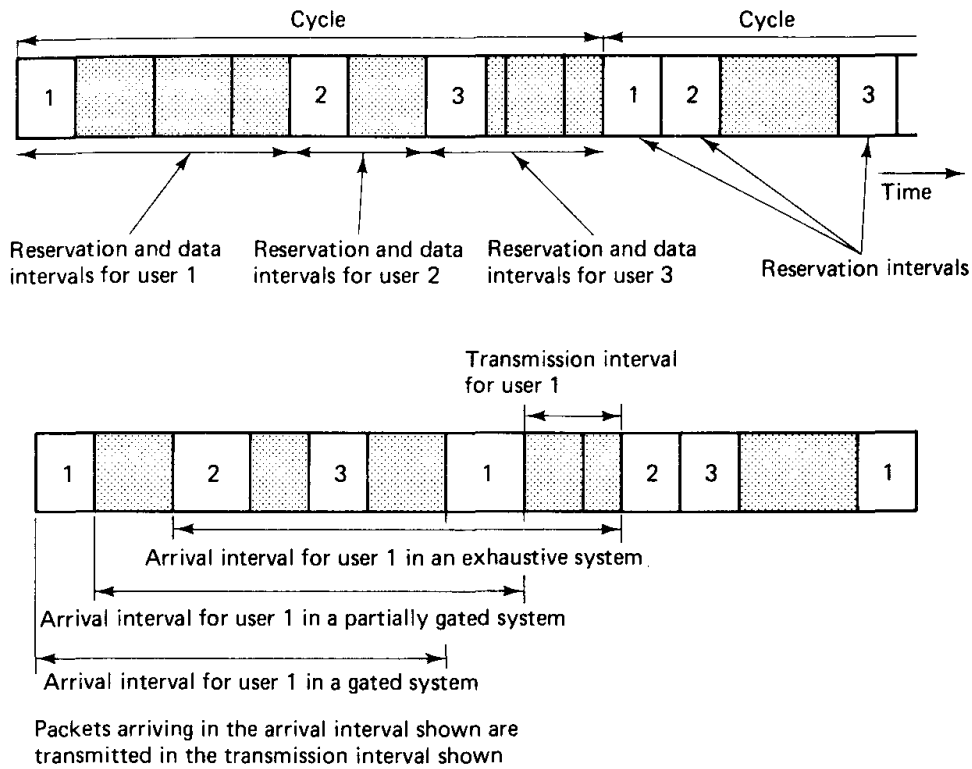



**Figure 3.21** Reservation or polling system with three users. In the exhaustive version, a packet of a user that arrives during the user's reservation or data interval is transmitted in the same data interval. In the partially gated version, a packet of a user arriving during the user's data interval must wait for an entire cycle and be transmitted during the next data interval of the user. In the fully gated version, packets arriving during the user's reservation interval must also wait for an entire cycle. The figure shows, for the three systems, the association between the interval in which a packet arrives and the interval in which the packet is transmitted.

is one of the most common local area networks, the token ring. The users are connected by cable in a unidirectional loop. Each user transmits the current packet backlog, then gives the opportunity to a neighbor to transmit, and the process is repeated. (A more detailed description of the token ring is given in Chapter 4.)

We assume that the arrival processes of all users are independent Poisson with rate $\lambda/m$, and that the first and second moments of the packet transmission times are $\overline{X} = 1/\mu$ and $\overline{X^2}$, respectively. The utilization factor is $\rho = \lambda/\mu$. Interarrival times and transmission times are, as usual, assumed independent. While we assume that all users have identical arrival and service statistics, we allow the reservation intervals of different users to have different statistics.

**Single-user system.** Our general line of analysis of reservation systems can be better understood in terms of the special case where $m = 1$, so we consider this case first. We may also view this as a system where all users share reservation and data intervals. Let $V_\ell$ be the duration of the $\ell^{\text{th}}$ reservation interval and assume that successive reservation intervals are independent and identically distributed random variables with first and second moments $\overline{V}$ and $\overline{V^2}$, respectively. We consider a gated system and

assume that the reservation intervals are statistically independent of the arrival times and service durations. Finally, for convenience of exposition, we assume that packets are transmitted in the order of their arrival. As in our derivation of the P-K formula, expected delays and queue lengths are independent of service order as long as service order is independent of service requirement (*i.e.*, packet length).

Consider the $i^{\text{th}}$ data packet arriving at the system. This packet must wait in queue for the residual time $R_i$ until the end of the current packet transmission or reservation interval. It must also wait for the transmission of the $N_i$ packets currently in the queue (this includes both packets for which reservations were already made in the last reservation interval and earlier arrivals waiting to make a reservation). Finally, the packet must wait during the next reservation interval $V_{\ell(i)}$, say, in which its reservation will be made (see Fig. 3.22). Thus, the expected queueing delay for the $i^{\text{th}}$ packet is given by

$$E\{W_i\} = E\{R_i\} + \frac{E\{N_i\}}{\mu} + E\{V_{\ell(i)}\} \tag{3.60}$$

The similarity of this reservation system to the $M/G/1$ queue with vacations should be noted. The only difference is that in the gated reservation system, a reservation interval starts when all packets that arrived prior to the start of the preceding reservation interval have been served, whereas in the vacation system, a vacation interval starts when all arrivals up to the current time have been served and the system is empty. (Thus in the gated reservation system, every packet has to wait in queue for a full reservation interval, while in the vacation system, only the packets that find the system empty upon arrival have to wait for part of a vacation interval. Note that the exhaustive version of this reservation system is equivalent to the vacation system.) The time-average mean residual time for the two systems is the same (the calculation based on Fig. 3.19 still applies) and is given by $\lambda\overline{X^2}/2 + (1-\rho)\overline{V^2}/2\overline{V}$. The value of $\lim_{i\to\infty} E\{N_i\}/\mu$ is $\rho W$ in both systems, and finally the value of $\lim_{i\to\infty} E\{V_{\ell(i)}\}$ is just $\overline{V}$. Thus, from

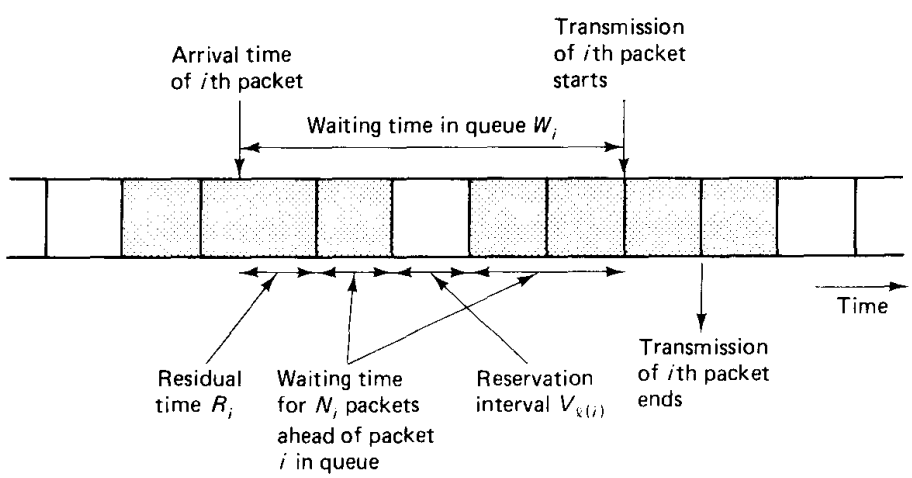



**Figure 3.22** Calculation of the average waiting time in the single-user gated system. The expected waiting time $E\{W_i\}$ of the $i^{\text{th}}$ packet is

$$E\{W_i\} = E\{R_i\} + \frac{E\{N_i\}}{\mu} + E\{V_{\ell(i)}\}$$

Eq. (3.60) the expected time in queue for the single-user reservation system is

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho)} + \frac{\overline{V^2}}{2\overline{V}} + \frac{\overline{V}}{1-\rho} \qquad \text{(single user, gated)} \tag{3.61}$$

In the common situation where the reservation interval is a constant $A$, this simplifies to

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho)} + \frac{A}{2}\left(\frac{3-\rho}{1-\rho}\right) \tag{3.62}$$

There is an interesting paradox associated with the waiting time formula (3.61). We have seen that a fraction $1-\rho$ of time is used on reservations. Since there is one reservation interval of mean duration $\overline{V}$ per cycle, we can conclude that the expected cycle length must be $\overline{V}/(1-\rho)$ (see also Example 3.6). The mean queueing delay in Eq. (3.61) can be an arbitrarily large multiple of this mean cycle length, which seems paradoxical since each packet is transmitted on the cycle following its arrival. The explanation of this is that more packets tend to arrive in long cycles than in short cycles, and thus mean cycle length is not representative of the cycle lengths seen by arriving packets; this is the same phenomenon that makes the mean residual service time used in the P-K formula derivation larger than one might think (see also Problem 3.31).

**Multiuser system.** Suppose that the system has $m$ users, each with independent Poisson arrivals of rate $\lambda/m$. Again $\overline{X}$ and $\overline{X^2}$ are the first two moments of the service time for each user's packets. We denote by $\overline{V_i}$ and $\overline{V_i^2}$, respectively, the first two moments of the reservation intervals of user $i$. The service times and reservation intervals are all independent. We number the users $0, 1, \ldots, m-1$ and assume that the $\ell^{\text{th}}$ reservation interval is used to make reservations for user $\ell \bmod m$ and the subsequent ($\ell^{\text{th}}$) data interval is used to send the packets corresponding to those reservations.

Consider the $i^{\text{th}}$ packet arrival into the system (counting packets in order of arrival, regardless of user). As before, the expected delay for this packet consists of three terms: first, the mean residual time for the packet or reservation in progress; second, the expected time to transmit the number $N_i$ of packets that must be transmitted before packet $i$; and third, the expected duration of reservation intervals (see Fig. 3.23). Thus,

$$E\{W_i\} = E\{R_i\} + \frac{E\{N_i\}}{\mu} + E\{Y_i\} \tag{3.63}$$

where $Y_i$ is the duration of all the whole reservation intervals during which packet $i$ must wait before being transmitted. The time average mean residual time is calculated as before, and is given by

$$R = \frac{\lambda \overline{X^2}}{2} + \frac{(1-\rho)\sum_{\ell=0}^{m-1} \overline{V_\ell^2}}{2\sum_{\ell=0}^{m-1} \overline{V_\ell}} \tag{3.64}$$

The number of packets $N_i$ that $i$ must wait for is not equal to the number already in queue, but the order of serving packets is independent of packet service time; thus, each packet served before $i$ still has a mean transmission time $1/\mu$ as indicated in Eq. (3.63) and by Little's formula, the value of $\lim_{i\to\infty} E\{N_i\}/\mu$ is $\rho W$ as before.

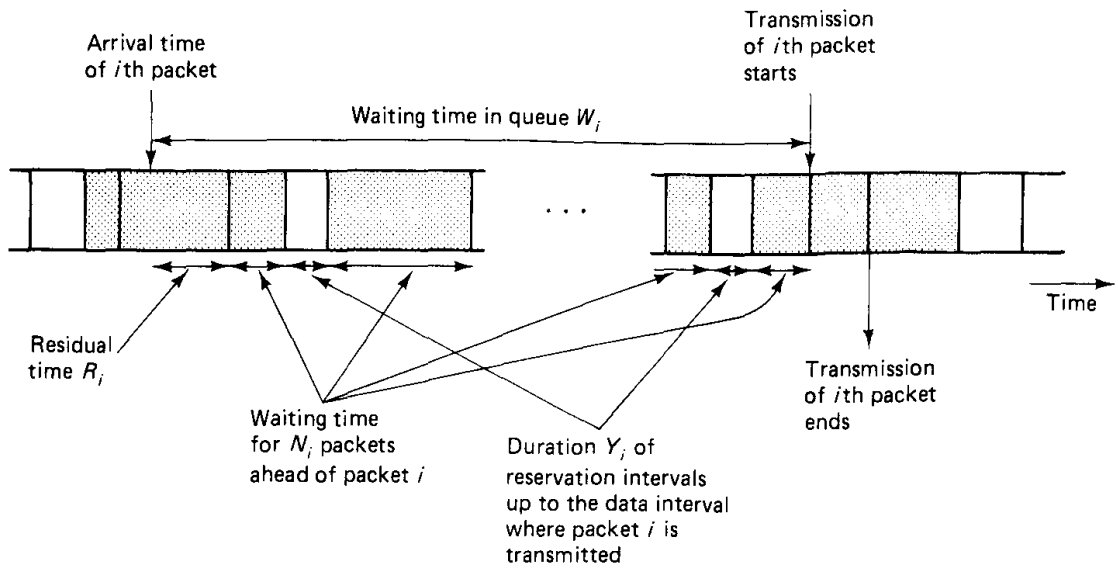



**Figure 3.23** Calculation of the average waiting time in the multiuser system. The expected waiting time $E\{W_i\}$ of the $i^{\text{th}}$ packet is

$$E\{W_i\} = E\{R_i\} + \frac{E\{N_i\}}{\mu} + E\{Y_i\}$$

Letting $Y = \lim_{i\to\infty} E\{Y_i\}$, we can thus write the steady-state version of Eq. (3.63):

$$W = R + \rho W + Y$$

or, equivalently,

$$W = \frac{R+Y}{1-\rho} \tag{3.65}$$

We first calculate $Y$ for an exhaustive system. Denote

$$\alpha_{\ell j} = E\{Y_i \mid \text{ packet } i \text{ arrives in user } \ell\text{'s reservation or data interval and belongs to user } (\ell + j) \bmod m\}$$

We have

$$\alpha_{\ell j} = \begin{cases} 0, & j = 0 \\ \overline{V}_{(\ell+1) \bmod m} + \cdots + \overline{V}_{(\ell+j) \bmod m}, & j > 0 \end{cases}$$

Since packet $i$ belongs to any user with equal probability $1/m$, we have

$$E\{Y_i \mid \text{ packet } i \text{ arrives in user } \ell\text{'s reservation or data interval}\}$$

$$= \frac{1}{m}\sum_{j=1}^{m-1} \alpha_{\ell j} = \sum_{j=1}^{m-1} \frac{m-j}{m} \overline{V}_{(\ell+j) \bmod m} \tag{3.66}$$

Since all users have equal data rate, the data intervals of all users have equal average length in steady-state. Therefore, in steady-state, a packet will arrive during user $\ell$'s data interval with probability $\rho/m$, and during user $\ell$'s reservation interval with probability

$(1-\rho)\overline{V}_\ell / \left(\sum_{k=0}^{m-1} \overline{V}_k\right)$. Using this fact in Eq. (3.66), we obtain the following equation for $Y = \lim_{i\to\infty} E\{Y_i\}$:

$$Y = \sum_{\ell=0}^{m-1} \left(\frac{\rho}{m} + \frac{(1-\rho)\overline{V}_\ell}{\sum_{k=0}^{m-1} \overline{V}_k}\right) \sum_{j=1}^{m-1} \frac{m-j}{m} \overline{V}_{(\ell+j) \bmod m}$$

$$= \frac{\rho}{m} \sum_{j=1}^{m-1} \frac{m-j}{m} \left(\sum_{\ell=0}^{m-1} \overline{V}_\ell\right) + \frac{1-\rho}{\sum_{k=0}^{m-1} \overline{V}_k} \sum_{\ell=0}^{m-1} \sum_{j=1}^{m-1} \frac{m-j}{m} \overline{V}_\ell \overline{V}_{(\ell+j) \bmod m} \tag{3.67}$$

The last sum above can be written

$$\sum_{\ell=0}^{m-1} \sum_{j=1}^{m-1} \frac{m-j}{m} \overline{V}_\ell \overline{V}_{(\ell+j) \bmod m} = \frac{1}{2} \left[\left(\sum_{\ell=0}^{m-1} \overline{V}_\ell\right)^2 - \sum_{\ell=0}^{m-1} \overline{V}_\ell^2\right]$$

(To see this, note that the right side above is the sum of all possible products $\overline{V}_\ell \overline{V}_{\ell'}$ for $\ell \neq \ell'$. The left side is the sum of all possible terms $(j/m)\overline{V}_\ell \overline{V}_{\ell'}$ and $[(m-j)/m]\overline{V}_\ell \overline{V}_{\ell'}$, where $j = |\ell - \ell'|$ and $\ell \neq \ell'$.) Using this expression, and denoting

$$\overline{V} = \frac{1}{m} \sum_{\ell=0}^{m-1} \overline{V}_\ell$$

as the reservation interval averaged over all users, we can write Eq. (3.67) as

$$Y = \frac{\rho \overline{V}(m-1)}{2} + \frac{(1-\rho)m\overline{V}}{2} - \frac{(1-\rho)\sum_{\ell=0}^{m-1} \overline{V}_\ell^2}{2m\overline{V}}$$

$$= \frac{(m-\rho)\overline{V}}{2} - \frac{(1-\rho)\sum_{\ell=0}^{m-1} \overline{V}_\ell^2}{2m\overline{V}} \tag{3.68}$$

Combining Eqs. (3.64), (3.65), and (3.68), we obtain

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho)} + \frac{(m-\rho)\overline{V}}{2(1-\rho)} + \frac{\sum_{\ell=0}^{m-1} \left(\overline{V_\ell^2} - \overline{V}_\ell^2\right)}{2m\overline{V}}$$

Denoting

$$\sigma_V^2 = \frac{\sum_{\ell=0}^{m-1} \left(\overline{V_\ell^2} - \overline{V}_\ell^2\right)}{m}$$

as the variance of the reservation intervals averaged over all users, we finally obtain

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho)} + \frac{(m-\rho)\overline{V}}{2(1-\rho)} + \frac{\sigma_V^2}{2\overline{V}} \qquad \text{(exhaustive)} \tag{3.69}$$

The partially gated system is the same as the exhaustive except that if a packet of a user arrives during a user's own data interval (an event of probability $\rho/m$ in steady-state), it is delayed by an additional $m\overline{V}$, the average sum of reservation intervals in a

cycle. Thus, $Y$ is increased by $\rho\overline{V}$ in the preceding calculation, and we obtain

$$W = \frac{\lambda\overline{X^2}}{2(1-\rho)} + \frac{(m+\rho)\overline{V}}{2(1-\rho)} + \frac{\sigma_V^2}{2\overline{V}} \qquad \text{(partially gated)} \tag{3.70}$$

Consider, finally, the fully gated system. This is the same as the partially gated system except that if a packet of a user arrives during a user's own reservation interval [an event of probability $(1-\rho)/m$ in steady-state], it is delayed by an additional $m\overline{V}$. This increases $Y$ by an additional $(1-\rho)\overline{V}$ and results in the equation

$$W = \frac{\lambda\overline{X^2}}{2(1-\rho)} + \frac{(m+2-\rho)\overline{V}}{2(1-\rho)} + \frac{\sigma_V^2}{2\overline{V}} \qquad \text{(gated)} \tag{3.71}$$

In comparing these results with the single-user system, consider the case where the reservation interval is a constant $A/m$. Thus, $A$ is the overhead or reservation time for an entire cycle of reservations for each user, which is usually the appropriate parameter to compare with $A$ in the single-user waiting-time formula (3.62). We then have ($\overline{V} = A/m$, $\sigma_V^2 = 0$)

$$W = \frac{\lambda\overline{X^2}}{2(1-\rho)} + \frac{A}{2}\left(\frac{1-\rho/m}{1-\rho}\right) \qquad \text{(exhaustive)} \tag{3.72}$$

$$W = \frac{\lambda\overline{X^2}}{2(1-\rho)} + \frac{A}{2}\left(\frac{1+\rho/m}{1-\rho}\right) \qquad \text{(partially gated)} \tag{3.73}$$

$$W = \frac{\lambda\overline{X^2}}{2(1-\rho)} + \frac{A}{2}\left(\frac{1+(2-\rho)/m}{1-\rho}\right) \qquad \text{(gated)} \tag{3.74}$$

It can be seen that delay is somewhat reduced in the multiuser case; essentially, packets are delayed by roughly the same amount until the reservation time in all cases, but delay is quite small after the reservation in the multiuser case.

**Limited service systems.** We now consider a variation of the multiuser system whereby, in each user's data interval, *only the first* packet of the user waiting in queue (if any) is transmitted (rather than *all* waiting packets). We concentrate on the gated and partially gated versions of this system, since an exhaustive version does not make sense. As before, we have

$$E\{W_i\} = E\{R_i\} + \frac{E\{N_i\}}{\mu} + E\{Y_i\}$$

and by taking the limit as $i \to \infty$, we obtain

$$W = R + \rho W + Y \tag{3.75}$$

Here $R$ is given by Eq. (3.64) as before. To calculate the new formula for $Y$ for the partially gated system, we argue as follows. A packet arriving during user $\ell$'s data or reservation interval will belong to any one of the users with equal probability $1/m$. Therefore, in steady-state, the expected number of packets waiting in the queue of the user that owns the arriving packet, averaged over all users, is $\lim_{i\to\infty} E\{N_i\}/m = \lambda W/m$.

Each of these packets causes an extra cycle of reservations $m\overline{V}$, so $Y$ is increased by an amount $\lambda W\overline{V}$. Using this fact in Eq. (3.75), we see that

$$W = \frac{R + \widetilde{Y}}{1 - \rho - \lambda\overline{V}}$$

where $\widetilde{Y}$ is the value of $Y$ obtained earlier for the partially gated system without the single-packet-per-data-interval restriction. Equivalently, we see from Eq. (3.65) that *the single-packet-per-data-interval restriction results in an increase of the average waiting time for the partially gated system by a factor*

$$\frac{1 - \rho}{1 - \rho - \lambda\overline{V}}$$

Using this fact in Eq. (3.70), we obtain

$$W = \frac{\lambda\overline{X^2}}{2(1 - \rho - \lambda\overline{V})} + \frac{(m + \rho)\overline{V}}{2(1 - \rho - \lambda\overline{V})} + \frac{\sigma_V^2(1 - \rho)}{2\overline{V}(1 - \rho - \lambda\overline{V})}$$

(limited service, partially gated) (3.76)

Consider now the gated version. $Y_i$ is the same as for the partially gated system except for an additional cycle of reservation intervals of average length $m\overline{V}$ associated with the event where packet $i$ arrives during the reservation interval of its owner, and the subsequent data interval is empty. It is easily verified (Problem 3.34) that the latter event occurs with steady-state probability $(1 - \rho - \lambda\overline{V})/m$. Therefore, for the gated system $Y$ equals the corresponding value for the partially gated system plus $(1 - \rho - \lambda\overline{V})\overline{V}$. This adds $\overline{V}$ to the value of $W$ for the partially gated system, and the average waiting time now is

$$W = \frac{\lambda\overline{X^2}}{2(1 - \rho - \lambda\overline{V})} + \frac{(m + 2 - \rho - 2\lambda\overline{V})\overline{V}}{2(1 - \rho - \lambda\overline{V})} + \frac{\sigma_V^2(1 - \rho)}{2\overline{V}(1 - \rho - \lambda\overline{V})}$$

(limited service, gated) (3.77)

Note that it is not enough that $\rho = \lambda/\mu < 1$ for $W$ to be bounded; rather, $\rho + \lambda\overline{V} < 1$ is required or, equivalently,

$$\lambda\left(\frac{1}{\mu} + \overline{V}\right) < 1$$

This is due to the fact that each packet requires a separate reservation interval of average length $\overline{V}$, thereby effectively increasing the average transmission time from $1/\mu$ to $1/\mu + \overline{V}$.

As a final remark, consider the case of a very large number of users $m$ and a very small average reservation interval $\overline{V}$. An examination of the equation given for the average waiting time $W$ of every multiuser system considered so far shows that as $m \to \infty$, $\overline{V} \to 0$, $\sigma_V^2/\overline{V} \to 0$, and $m\overline{V} \to A$, where $A$ is a constant, we have

$$W \to \frac{\lambda\overline{X^2}}{2(1 - \rho)} + \frac{A}{2(1 - \rho)}$$

It can be shown (see Example 3.6) that $A/(1-\rho)$ is the average length of a cycle ($m$ successive reservation and data intervals). Thus, $W$ approaches the $M/G/1$ average waiting time plus one half the average cycle length.

### 3.5.3 Priority Queueing

Consider the $M/G/1$ system with the difference that arriving customers are divided into $n$ different priority classes. Class 1 has the highest priority, class 2 has the second highest, and so on. The arrival rate and the first two moments of service time of each class $k$ are denoted $\lambda_k$, $\overline{X}_k = 1/\mu_k$, and $\overline{X_k^2}$, respectively. The arrival processes of all classes are assumed independent, Poisson, and independent of the service times.

**Nonpreemptive priority.** We first consider the nonpreemptive priority rule whereby a customer undergoing service is allowed to complete service without interruption even if a customer of higher priority arrives in the meantime. A separate queue is maintained for each priority class. When the server becomes free, the first customer of the highest nonempty priority queue enters service. This priority rule is one of the most appropriate for modeling packet transmission systems.

We will develop an equation for average delay for each priority class, which is similar to the P-K formula and admits a similar derivation. Denote

$$N_Q^k = \text{Average number in queue for priority } k$$

$$W_k = \text{Average queueing time for priority } k$$

$$\rho_k = \lambda_k/\mu_k = \text{System utilization for priority } k$$

$$R = \text{Mean residual service time}$$

We assume that the overall system utilization is less than 1, that is,

$$\rho_1 + \rho_2 + \cdots + \rho_n < 1$$

When this assumption is not satisfied, there will be some priority class $k$ such that the average delay of customers of priority $k$ and lower will be infinite while the average delay of customers of priority higher than $k$ will be finite. Problem 3.39 takes a closer look at this situation.

As in the derivation of the P-K formula given earlier, we have for the highest-priority class,

$$W_1 = R + \frac{1}{\mu_1} N_Q^1$$

Eliminating $N_Q^1$ from this equation using Little's Theorem,

$$N_Q^1 = \lambda_1 W_1$$

we obtain

$$W_1 = R + \rho_1 W_1$$

and finally,

$$W_1 = \frac{R}{1 - \rho_1} \tag{3.78}$$

For the second priority class, we have a similar expression for the queueing delay $W_2$ except that we have to count the additional queueing delay due to customers of higher priority that arrive while a customer is waiting in queue. This is the meaning of the last term in the formula

$$W_2 = R + \frac{1}{\mu_1} N_Q^1 + \frac{1}{\mu_2} N_Q^2 + \frac{1}{\mu_1} \lambda_1 W_2$$

Using Little's Theorem ($N_Q^k = \lambda_k W_k$), we obtain

$$W_2 = R + \rho_1 W_1 + \rho_2 W_2 + \rho_1 W_2$$

which yields

$$W_2 = \frac{R + \rho_1 W_1}{1 - \rho_1 - \rho_2}$$

Using the expression $W_1 = R/(1 - \rho_1)$ obtained earlier, we finally have

$$W_2 = \frac{R}{(1 - \rho_1)(1 - \rho_1 - \rho_2)}$$

The derivation is similar for all priority classes $k > 1$. The formula for the waiting time in queue is

$$W_k = \frac{R}{(1 - \rho_1 - \cdots - \rho_{k-1})(1 - \rho_1 - \cdots - \rho_k)} \tag{3.79}$$

The average delay per customer of class $k$ is

$$T_k = \frac{1}{\mu_k} + W_k \tag{3.80}$$

The mean residual service time $R$ can be derived as for the P-K formula (compare with Fig. 3.16). We have

$$R = \frac{1}{2} \sum_{i=1}^{n} \lambda_i \overline{X_i^2} \tag{3.81}$$

The average waiting time in queue and the average delay per customer for each class is obtained by combining Eqs. (3.79) to (3.81):

$$W_k = \frac{\sum_{i=1}^{n} \lambda_i \overline{X_i^2}}{2(1 - \rho_1 - \cdots - \rho_{k-1})(1 - \rho_1 - \cdots - \rho_k)} \tag{3.82}$$

$$T_k = \frac{1}{\mu_k} + W_k \tag{3.83}$$

The analysis given above does not extend easily to the case of multiple servers, primarily because there is no simple formula for the mean residual time $R$. If, however,

the service times of all priority classes are identically and exponentially distributed, there is a convenient characterization of $R$. Equation (3.79) then yields a closed-form expression for the average waiting times $W_k$ (see Problem 3.38).

Note that it is possible to affect the average delay per customer by choosing the priority classes appropriately. It is generally true that average delay tends to be reduced when customers with short service times are given higher priority. (For an example from common experience, consider the supermarket practice of having special checkout counters for customers with few items. A similar situation can be seen in copying machine waiting lines, where people often give priority to others who need to make just a few copies.) For an analytical substantiation, consider a nonpreemptive system and two customer classes $A$ and $B$, with respective arrival and service rates $\lambda_A$, $\mu_A$, and $\lambda_B$, $\mu_B$. A straightforward calculation using the formulas above shows that if $\mu_A > \mu_B$, then the average delay per customer (averaged over both classes)

$$T = \frac{\lambda_A T_A + \lambda_B T_B}{\lambda_A + \lambda_B}$$

is smaller when $A$ is given priority over $B$ than when $B$ is given priority over $A$. For related results, see Problem 3.40.

**Preemptive resume priority.** One of the features of the nonpreemptive priority rule is that the average delay of a priority class depends on the arrival rate of lower-priority classes. This is evident from Eq. (3.82), which gives the average waiting times $W_k$, and is due to the fact that a high-priority customer must wait for a lower-priority customer already in service. This dependence is not present in the *preemptive resume priority discipline*, whereby service of a customer is interrupted when a higher-priority customer arrives and is resumed from the point of interruption once all customers of higher priority have been served.

As an example of an (approximation to) such a system, consider a transmission line serving several Poisson packet streams of different priorities. The packets of each stream are subdivided into many small "subpackets" (*e.g.*, ATM cells), which in the absence of packets of higher priority, are contiguously transmitted on the line. The transmission of the subpackets of a given packet is halted when a packet of higher priority arrives and is resumed when no subpackets of higher priority packets are left in the system.

As we consider the calculation of $T_k$, the average time in the system of priority $k$ customers, we should keep in mind that the presence of customers of priorities $k+1$ through $n$ does not affect this calculation. Therefore, we can treat each priority class as if it were the lowest in the system. The system time $T_k$ consists of three terms:

1. The customer's average service time $1/\mu_k$.
2. The average time required, upon arrival of a priority $k$ customer, to service customers of priority 1 to $k$ already in the system (*i.e.*, the average unfinished work corresponding to priorities 1 through $k$). It can be seen that this time is equal to the average waiting time in the corresponding, ordinary $M/G/1$ system (without priorities), where the customers of priorities $k+1$ through $n$ are neglected, that is

[cf. Eq. (3.48)],

$$\frac{R_k}{1-\rho_1-\cdots-\rho_k}$$

where $R_k$ is the mean residual time

$$R_k = \frac{\sum_{i=1}^{k} \lambda_i \overline{X_i^2}}{2} \tag{3.84}$$

The reason is that at all times, the unfinished work (sum of remaining service times of all customers in the system) of an $M/G/1$-type system is independent of the priority discipline of the system. This is true for any system where the server is always busy while the system is nonempty, and customers leave the system only after receiving their required service. (An example of a system that does not have this property is the vacation system of Section 3.5.1.)

3. The average waiting time for customers of priorities 1 through $k-1$ who arrive while the customer of class $k$ is in the system. This term is

$$\sum_{i=1}^{k-1} \frac{1}{\mu_i}\lambda_i T_k = \sum_{i=1}^{k-1} \rho_i T_k$$

for $k > 1$, and is zero for $k = 1$.

Collecting the three terms above, we obtain the equation

$$T_k = \frac{1}{\mu_k} + \frac{R_k}{1-\rho_1-\cdots-\rho_k} + \left(\sum_{i=1}^{k-1} \rho_i\right) T_k \tag{3.85}$$

The final result is, for $k = 1$,

$$T_1 = \frac{(1/\mu_1)(1-\rho_1)+R_1}{1-\rho_1} \tag{3.86}$$

and for $k > 1$,

$$T_k = \frac{(1/\mu_k)(1-\rho_1-\cdots-\rho_k)+R_k}{(1-\rho_1-\cdots-\rho_{k-1})(1-\rho_1-\cdots-\rho_k)} \tag{3.87}$$

where $R_k$ is given by Eq. (3.84). As for the nonpreemptive system, there is no easy extension of this formula to the case of multiple servers unless the service times of all priority classes are identically and exponentially distributed (see Problem 3.38).

### 3.5.4 An Upper Bound for the $G/G/1$ System

Consider the $G/G/1$ system, which is the same as $M/G/1$ except that the interarrival times have a general rather than exponential distribution. We continue to assume that the interarrival times and service times are all independent. We want to show that the average waiting time in queue satisfies

$$W \le \frac{\lambda(\sigma_a^2+\sigma_b^2)}{2(1-\rho)}. \tag{3.88}$$

where

$$\sigma_a^2 = \text{Variance of the interarrival times}$$
$$\sigma_b^2 = \text{Variance of the service times}$$
$$\lambda = \text{Average interarrival time}$$
$$\rho = \text{Utilization factor } \lambda/\mu, \text{ where } 1/\mu \text{ is the average service time}$$

The upper bound (3.88) becomes exact asymptotically as $\rho \to 1$, that is, as the system becomes heavily loaded.

Let us denote

$$W_k = \text{Waiting time of the } k^{\text{th}} \text{ customer}$$
$$X_k = \text{Service time of the } k^{\text{th}} \text{ customer}$$
$$\tau_k = \text{Interarrival time between the } k^{\text{th}} \text{ and } (k+1)^{\text{st}} \text{ customer}$$

From Fig. 3.24 we see that

$$W_{k+1} = \max\{0, W_k + X_k - \tau_k\} \tag{3.89}$$

To simplify the analysis, we will use the following notation for any random variable $Y$:

$$Y^+ = \max\{0, Y\}, \qquad Y^- = -\min\{0, Y\}$$
$$\overline{Y} = E\{Y\}, \qquad \sigma_Y^2 = E\{Y^2 - \overline{Y}^2\}$$

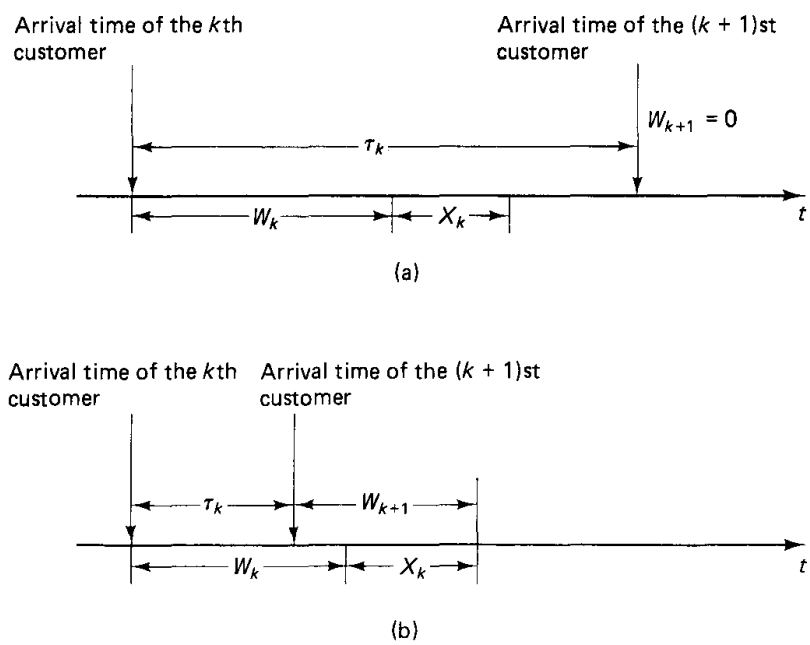



**Figure 3.24** Expressing the waiting time $W_{k+1}$ of the $(k+1)^{\text{st}}$ customer in terms of the waiting time $W_k$, the service time $X_k$, and the interarrival time $\tau_k$ of the $k^{\text{th}}$ customer. If the $k^{\text{th}}$ customer has departed before the $(k+1)^{\text{st}}$ customer's arrival, which is equivalent to $W_k + X_k - \tau_k \leq 0$, then $W_{k+1} = 0$ [case (a)]. Otherwise, we have $W_{k+1} = W_k + X_k - \tau_k$ [case (b)].

Note that we have

$$Y = Y^+ - Y^-, \qquad Y^+ \cdot Y^- = 0$$

from which we see that

$$\overline{Y} = \overline{Y^+} - \overline{Y^-}, \qquad \sigma_Y^2 = \sigma_{Y^+}^2 + \sigma_{Y^-}^2 + 2\overline{Y^+} \cdot \overline{Y^-} \tag{3.90}$$

Let us now write the expression (3.89) as

$$W_{k+1} = (W_k + V_k)^+ \tag{3.91}$$

where

$$V_k = X_k - \tau_k \tag{3.92}$$

Let us also denote

$$I_k = (W_k + V_k)^- \tag{3.93}$$

From Fig. 3.24 we see that $I_k$ is the length of the idle period between the arrival of the $k^{\text{th}}$ and the arrival of the $(k+1)^{\text{st}}$ customer.

We have, using Eq. (3.90),

$$\sigma_{(W_k+V_k)}^2 = \sigma_{(W_k+V_k)^+}^2 + \sigma_{(W_k+V_k)^-}^2 + 2\overline{(W_k + V_k)^+} \cdot \overline{(W_k + V_k)^-} \tag{3.94}$$

Since $W_k$ and $V_k$ are independent, we also have

$$\sigma_{(W_k+V_k)}^2 = \sigma_{W_k}^2 + \sigma_{V_k}^2 = \sigma_{W_k}^2 + \sigma_a^2 + \sigma_b^2 \tag{3.95}$$

Combining Eqs. (3.91) to (3.95), we obtain

$$\sigma_{W_k}^2 + \sigma_a^2 + \sigma_b^2 = \sigma_{W_{k+1}}^2 + \sigma_{I_k}^2 + 2\overline{W}_{k+1}\overline{I}_k$$

We now take the limit as $k \to \infty$, assuming that steady-state values exist, that is,

$$\overline{W}_k \to W, \qquad \sigma_{W_k}^2 \to \sigma_W^2, \qquad \overline{I}_k \to I, \qquad \sigma_{I_k}^2 \to \sigma_I^2$$

We obtain

$$W = \frac{\sigma_a^2 + \sigma_b^2}{2I} - \frac{\sigma_I^2}{2I} \tag{3.96}$$

The average idle time $I$ between two successive arrivals is equal to $(1-\rho)$ (the fraction of time the system is idle) multiplied by the average interarrival time $1/\lambda$, that is, $I = (1-\rho)/\lambda$. Thus, we can write Eq. (3.96) as

$$W = \frac{\lambda(\sigma_a^2 + \sigma_b^2)}{2(1-\rho)} - \frac{\lambda\sigma_I^2}{2(1-\rho)} \tag{3.97}$$

Since $\sigma_I^2 \geq 0$, we obtain the inequality

$$W \leq \frac{\lambda(\sigma_a^2 + \sigma_b^2)}{2(1-\rho)} \tag{3.98}$$

which is the desired result. Note that as the system becomes more heavily loaded, the average idle time $I_k$ tends to diminish and so does the variance $\sigma_I^2$, thereby making the upper bound increasingly accurate.

As an example, consider the $M/G/1$ queue. By the Pollatcheck–Khinchin formula we have

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho)} = \frac{\lambda(\sigma_b^2 + 1/\mu^2)}{2(1-\rho)} \tag{3.99}$$

Since for the Poisson arrival process with rate $\lambda$ we have $\sigma_a^2 = 1/\lambda^2$, by comparing Eqs. (3.97) and (3.99) we see that

$$\sigma_I^2 = \frac{1}{\lambda^2} - \frac{1}{\mu^2}$$

Thus the term that was neglected to derive the upper bound (3.98) for $W$ is equal to

$$\frac{\lambda \sigma_I^2}{2(1-\rho)} = \frac{\lambda}{2(1-\lambda/\mu)}\left(\frac{1}{\lambda^2} - \frac{1}{\mu^2}\right) = \frac{1}{2}\left(\frac{1}{\lambda} + \frac{1}{\mu}\right)$$

This term is always less than the average interarrival time $1/\lambda$ when $\rho < 1$. As $\rho \to 1$, it approaches $1/\mu$ and is negligible relative to the upper bound of Eq. (3.98).

We finally note that several other bounds and approximations for the $G/G/1$ queue have been obtained. A particularly simple improvement to the one we have given here is

$$W \le \frac{\lambda(\sigma_a^2 + \sigma_b^2)}{2(1-\rho)} - \frac{\lambda(1-\rho)\sigma_a^2}{2}$$

Its derivation is outlined in Problem 3.48.

## 3.6 NETWORKS OF TRANSMISSION LINES

In a data network, there are many transmission queues that interact in the sense that a traffic stream departing from one queue enters one or more other queues, perhaps after merging with portions of other traffic streams departing from yet other queues. Analytically, this has the unfortunate effect of complicating the character of the arrival processes at downstream queues. The difficulty is that the packet interarrival times become strongly correlated with packet lengths once packets have traveled beyond their entry queue. As a result it is impossible to carry out a precise and effective analysis comparable to the one for the $M/M/1$ and $M/G/1$ systems.

As an illustration of the phenomena that complicate the analysis, consider two transmission lines of equal capacity in tandem, as shown in Fig. 3.25. Assume that Poisson arrivals of rate $\lambda$ packets/sec enter the first queue, and that all packets have *equal* length. Therefore, the first queue is $M/D/1$ and the average packet delay there is given by the Pollaczek–Khinchin formula. However, at the second queue the interarrival times must be greater than or equal to $1/\mu$ (the packet transmission time). Furthermore, because the packet transmission times are equal at both queues, each packet arriving at

the second queue will complete transmission at or before the time the next packet arrives, so there is *no waiting at the second queue*. Therefore, a delay model based on Poisson assumptions is totally inappropriate for the second queue.

Consider next the case of the two tandem transmission lines where packet lengths are exponentially distributed and are independent of each other as well as of the interarrival times at the first queue. Then the first queue is $M/M/1$. The second queue, however, *cannot* be modeled as $M/M/1$. The reason is, again, that *the interarrival times at the second queue are strongly correlated with the packet lengths*. In particular, the interarrival time of two packets at the second queue is greater than or equal to the transmission time of the second packet at the first queue (see Fig. 3.26). As a result, long packets will typically wait less time at the second queue than short packets, since their transmission at the first queue takes longer, thereby giving the second queue more time to empty out. For a traffic analogy, consider a slow truck traveling on a busy narrow street together with several faster cars. The truck will typically see empty space ahead of it while being closely followed by the faster cars.

As an indication of the difficulty of analyzing queueing network problems involving dependent interarrival and service times, no analytical solution is known for even the simple tandem queueing problem of Fig. 3.25 involving Poisson arrivals and exponentially distributed service times. In the real situation where packet lengths and interarrival times are correlated, a simulation has shown that under heavy traffic conditions, average delay per packet is smaller than in the idealized situation where there is no such correlation. The reverse is true under light traffic conditions. It is not known whether and in what form this result can be extended to more general networks.

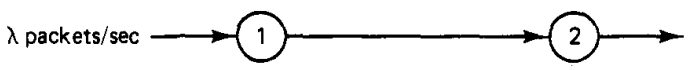



**Figure 3.25** Two equal-capacity transmission lines in tandem. If all packets have equal length, there is no queueing delay in the second queue.

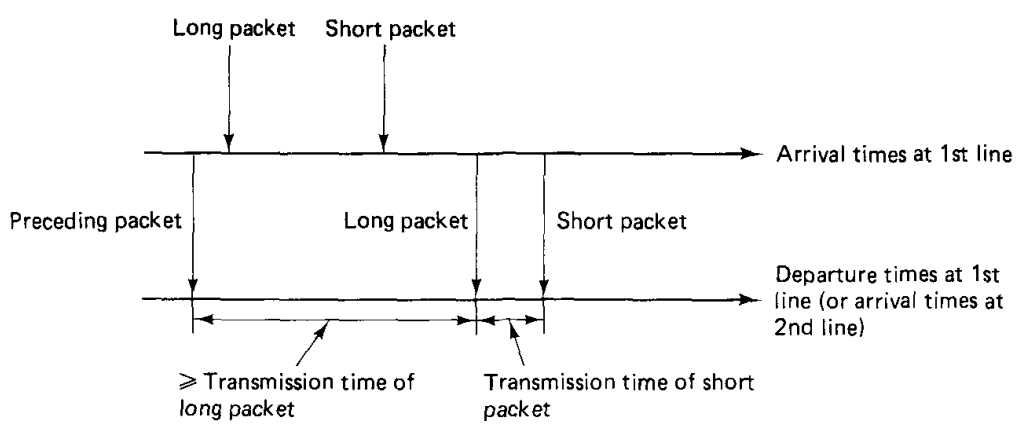



**Figure 3.26** Timing diagram of packet arrivals and departures completions in a system of two transmission lines in tandem. The interarrival time of two packets at the second queue is greater or equal to the transmission time of the second packet. (It is greater if and only if the second packet finds the first queue empty upon arrival.) Hence the interarrival times at the second queue are correlated with the packet lengths.

### 3.6.1 The Kleinrock Independence Approximation

We now formulate a framework for approximation of average delay per packet in data networks. Consider a network of communication links as shown in Fig. 3.27. Assume that there are several packet streams, each following a unique path that consists of a sequence of links through the network. Let $x_s$, in packets/sec, be the arrival rate of the packet stream $s$. Then the total arrival rate at link $(i, j)$ is

$$\lambda_{ij} = \sum_{\substack{\text{all packet streams } s \\ \text{crossing link } (i,j)}} x_s$$

The preceding network model is well suited for virtual circuit networks, with each packet stream modeling a separate virtual circuit. For datagram networks, it is sometimes necessary to use a more general model that allows bifurcation of the traffic of a packet stream. Here there are again several packet streams, each having a unique origin and destination. However, there may be several paths followed by the packets of a stream (see Fig. 3.28). Assume that no packets travel in a loop, let $x_s$ denote the arrival rate of packet stream $s$, and let $f_{ij}(s)$ denote the fraction of the packets of stream $s$ that go through link $(i, j)$. Then the total arrival rate at link $(i, j)$ is

$$\lambda_{ij} = \sum_{\substack{\text{all packet streams } s \\ \text{crossing link } (i,j)}} f_{ij}(s) x_s$$

We have seen from the special case of two tandem queues that even if the packet streams are Poisson with independent packet lengths at their point of entry into the network, this property is lost after the first transmission line. To resolve the dilemma, it was suggested by Kleinrock [Kle64] that merging several packet streams on a transmission line has an effect akin to restoring the independence of interarrival times and packet lengths. For example, if the second transmission line in the preceding tandem queue case were to receive a substantial amount of additional external Poisson traffic,

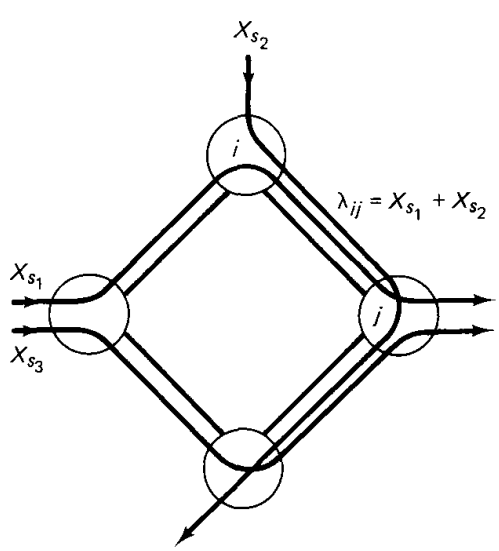



**Figure 3.27** Model suitable for virtual circuit networks. There are several packet streams, each using a single path. The total arrival rate $\lambda_{ij}$ at a link $(i, j)$ is equal to the sum of the arrival rates $x_s$ of all packet streams $s$ traversing the link.

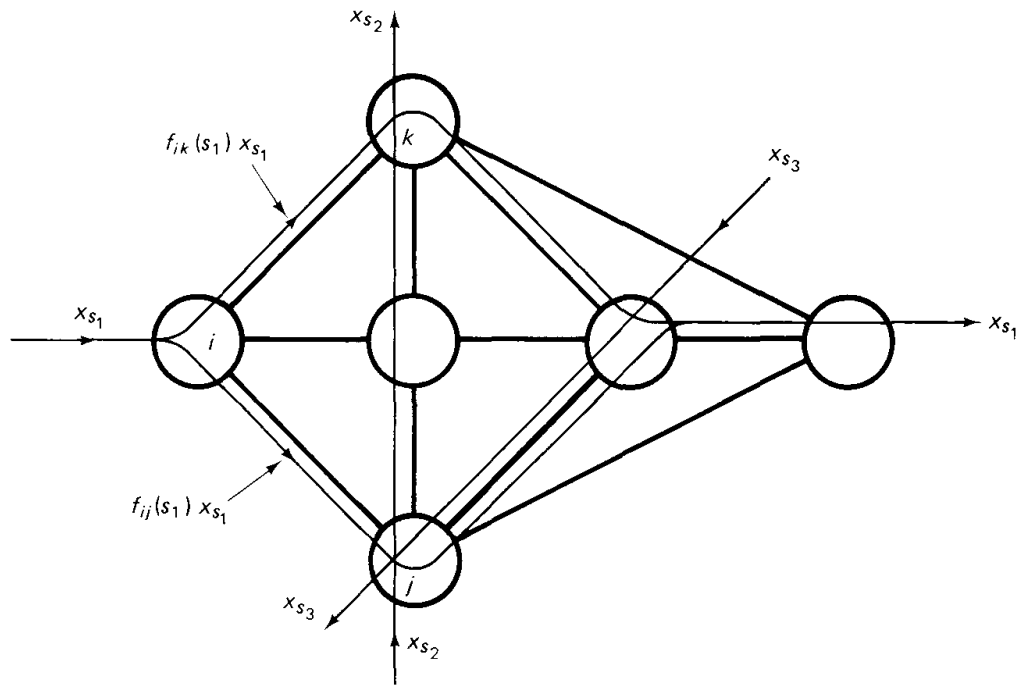



**Figure 3.28** Model suitable for datagram networks. There are several packet streams, each associated with a unique origin–destination pair. However, packets of the same stream may follow one of several paths. The total arrival rate $\lambda_{ij}$ at a link $(i, j)$ is equal to the sum of the fractions $f_{ij}(s)x_s$ of the arrival rates of all packet streams $s$ traversing the link.

the dependence of interarrival and service times displayed in Fig. 3.26 would be weakened considerably. It was concluded that it is often appropriate to adopt an $M/M/1$ queueing model for each communication link regardless of the interaction of traffic on this link with traffic on other links. (See also the discussion preceding Jackson's theorem in Section 3.8.) This is known as the *Kleinrock independence approximation* and seems to be a reasonably good approximation for systems involving Poisson stream arrivals at the entry points, packet lengths that are nearly exponentially distributed, a densely connected network, and moderate-to-heavy traffic loads. Based on this $M/M/1$ model, the average number of packets in queue or service at $(i, j)$ is

$$N_{ij} = \frac{\lambda_{ij}}{\mu_{ij} - \lambda_{ij}} \tag{3.100}$$

where $1/\mu_{ij}$ is the average packet transmission time on link $(i, j)$. The average number of packets summed over all queues is

$$N = \sum_{(i,j)} \frac{\lambda_{ij}}{\mu_{ij} - \lambda_{ij}} \tag{3.101}$$

so by Little's Theorem, the average delay per packet (neglecting processing and propagation delays) is

$$T = \frac{1}{\gamma} \sum_{(i,j)} \frac{\lambda_{ij}}{\mu_{ij} - \lambda_{ij}} \tag{3.102}$$

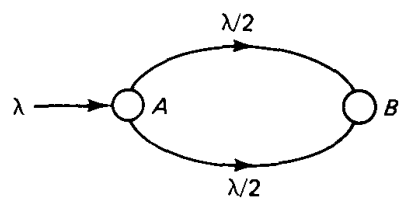



**Figure 3.29** Poisson process with rate $\lambda$ divided among two links. If division is done by randomization, each link behaves like an $M/M/1$ queue. If division is done by metering, the whole system behaves like an $M/M/2$ queue.

where $\gamma = \sum_s x_s$ is the total arrival rate in the system. If the average processing and propagation delay $d_{ij}$ at link $(i,j)$ is not negligible, this formula should be adjusted to

$$T = \frac{1}{\gamma} \sum_{(i,j)} \left( \frac{\lambda_{ij}}{\mu_{ij} - \lambda_{ij}} + \lambda_{ij} d_{ij} \right) \tag{3.103}$$

Finally, the average delay per packet of a traffic stream traversing a path $p$ is given by

$$T_p = \sum_{\substack{\text{all } (i,j) \\ \text{on path } p}} \left( \frac{\lambda_{ij}}{\mu_{ij}(\mu_{ij} - \lambda_{ij})} + \frac{1}{\mu_{ij}} + d_{ij} \right) \tag{3.104}$$

where the three terms in the sum above represent average waiting time in queue, average transmission time, and processing and propagation delay, respectively.

In many networks, the assumption of exponentially distributed packet lengths is not appropriate. Given a different type of probability distribution of the packet lengths, one may keep the approximation of independence between queues but use the P-K formula for average number in the system in place of the $M/M/1$ formula (3.100). Equations (3.101) to (3.104) for average delay would then be modified in an obvious way.

For virtual circuit networks (cf. Fig. 3.27), the main approximation involved in the $M/M/1$ formula (3.101) is due to the correlation of the packet lengths and the packet interarrival times at the various queues in the network. If somehow this correlation was not present (*e.g.*, if a packet upon departure from a transmission line was assigned a new length drawn from an exponential distribution), then the average number of packets in the system would be given indeed by the formula

$$N = \sum_{(i,j)} \frac{\lambda_{ij}}{\mu_{ij} - \lambda_{ij}}$$

This fact (by no means obvious) is a consequence of Jackson's Theorem, which will be discussed in Section 3.8.

In datagram networks that involve multiple path routing for some origin–destination pairs (cf. Fig. 3.28), the accuracy of the $M/M/1$ approximation deteriorates for another reason, which is best illustrated by an example.

**Example 3.17**

Suppose that node $A$ sends traffic to node $B$ along two links with service rate $\mu$ in the network of Fig. 3.29. Packets arrive at $A$ according to a Poisson process with rate $\lambda$ packets/sec. Packet transmission times are exponentially distributed and independent of interarrival times as in the $M/M/1$ system. Assume that the arriving traffic is to be divided equally among the two links. However, how should this division be implemented? Consider the following possibilities.

1. *Randomization.* Here each packet is assigned upon arrival at $A$ to one of the two links based on the outcome of a fair coin flip. It is then possible to show that the arrival process on each of the two queues is Poisson and independent of the packet lengths (see Problem 3.11). Therefore, each of the two queues behaves like an $M/M/1$ queue with arrival rate $\lambda/2$ and average delay per packet

$$T_R = \frac{1}{\mu - \lambda/2} = \frac{2}{2\mu - \lambda} \tag{3.105}$$

which is consistent with the Kleinrock independence approximation.

2. *Metering.* Here each arriving packet is assigned to a queue that currently has the smallest total backlog in bits and will therefore empty out first. An equivalent system maintains a common queue for the two links and routes the packet at the head of the queue to the link that becomes idle first. This works like an $M/M/2$ system with arrival rate $\lambda$ and with each link playing the role of a server. Using the result of Section 3.4.1, the average delay per packet can be calculated to be

$$T_M = \frac{2}{(2\mu - \lambda)(1 + \rho)} \tag{3.106}$$

where $\rho = \lambda/2\mu$.

Comparing the average delay expressions (3.105) and (3.106), we see that metering performs better than randomization in terms of delay by a factor $1/(1+\rho)$. This is basically the same advantage that statistical multiplexing with multiple channels holds over time-division multiplexing as discussed in Example 3.10. Generally, it is preferable to use some form of metering rather than randomization when dividing traffic among alternative routes. However, in contrast with randomization, metering destroys the Poisson character of the arrival process at the point of division. In our example, when metering is used, the interarrival times at each link are neither exponentially distributed nor independent of preceding packet lengths. Therefore, the use of metering (which is recommended for performance reasons) tends to degrade the accuracy of the $M/M/1$ approximation.

We finally mention an alternative approach for approximating average delay in a network of transmission lines. This approach uses $G/G/1$ approximations in place of $M/M/1$ or $M/G/1$ approximations. The key idea is that given the first two moments of the interarrival and service times of each of the external packet streams, one may approximate reasonably well the first two moments of the interarrival and service times of the total packet arrival stream at each queue (see [Whi83a], [Whi83b], and the references quoted there). Then the average delay at each queue can be estimated using $G/G/1$ bounds and approximations of the type discussed in Section 3.5.4.

## 3.7 TIME REVERSIBILITY—BURKE'S THEOREM

The analysis of the $M/M/1$, $M/M/m$, $M/M/\infty$, and $M/M/m/m$ systems was based on the equality of the steady-state frequency of transitions from $j$ to $j+1$, that is, $p_j P_{j(j+1)}$, with the steady-state frequency of transitions from $j+1$ to $j$, that is, $p_{j+1} P_{(j+1)j}$. These relations, called *detailed balance equations*, are valid for any Markov

chain with integer states in which transitions can occur only between neighboring states (*i.e.*, from $j$ to $j-1$, $j$, or $j+1$); these Markov chains are called *birth–death* processes. The detailed balance equations lead to an important property called time reversibility, as we now explain.

Consider an irreducible, aperiodic, discrete-time Markov chain $X_n, X_{n+1}, \ldots$ having transition probabilities $P_{ij}$ and stationary distribution $\{p_j \mid j \geq 0\}$ with $p_j > 0$ for all $j$. Suppose that the chain is in steady-state, that is,

$$P\{X_n = j\} = p_j\,, \qquad \text{for all } n$$

(This occurs if the initial state is chosen according to the stationary distribution, and is equivalent to imagining that the process began at time $-\infty$.)

Suppose that we trace the sequence of states going backward in time. That is, starting at some $n$, consider the sequence of states $X_n, X_{n-1}, \ldots$ . This sequence is itself a Markov chain, as seen by the following calculation:

$$\begin{aligned}
&P\{X_m = j \mid X_{m+1} = i,\ X_{m+2} = i_2, \ldots, X_{m+k} = i_k\} \\
&= \frac{P\{X_m = j, X_{m+1} = i,\ X_{m+2} = i_2, \ldots, X_{m+k} = i_k\}}{P\{X_{m+1} = i,\ X_{m+2} = i_2, \ldots, X_{m+k} = i_k\}} \\
&= \frac{P\{X_m = j, X_{m+1} = i\}\, P\{X_{m+2} = i_2, \ldots, X_{m+k} = i_k \mid X_m = j, X_{m+1} = i\}}{P\{X_{m+1} = i\} P\{X_{m+2} = i_2, \ldots, X_{m+k} = i_k \mid X_{m+1} = i\}} \\
&= \frac{P\{X_m = j, X_{m+1} = i\}}{P\{X_{m+1} = i\}} \\
&= \frac{P\{X_m = j\}\, P\{X_{m+1} = i \mid X_m = j\}}{P\{X_{m+1} = i\}} \\
&= \frac{p_j P_{ji}}{p_i}
\end{aligned}$$

where the third equality follows from the Markov property of the chain $X_n, X_{n+1}, \ldots$ Thus, conditional on the state at time $m+1$, the state at time $m$ is independent of that at times $m+2$, $m+3, \ldots$ . The backward transition probabilities are given by

$$P^*_{ij} = P\{X_m = j \mid X_{m+1} = i\} = \frac{p_j P_{ji}}{p_i}, \qquad i, j \geq 0 \tag{3.107}$$

If $P^*_{ij} = P_{ij}$ for all $i, j$ (*i.e.*, the transition probabilities of the forward and reversed chain are identical), we say that the chain is *time reversible*.

We list some properties of the reversed chain:

1. The reversed chain is irreducible, aperiodic, and has the same stationary distribution as the forward chain. [This property can be shown either by elementary reasoning using the definition of the reversed chain, or by verifying the equality $p_j = \sum_{i=0}^{\infty} p_i P^*_{ij}$ using Eq. (3.107).] The intuitive idea here is that the reversed chain corresponds to the same process, looked at in the reversed time direction. Thus, if the steady-state probabilities are viewed as proportions of time the process

visits the states, then the steady-state occupancy distributions of the forward and the reverse chains are equal. Note that in view of this equality, the form of the transition probabilities of the reversed chain $P^*_{ij} = p_j P_{ji}/p_i$ [cf. Eq. (3.107)] can be intuitively explained. It expresses the fact that (with probability 1) the proportion of transitions from $j$ to $i$ out of all transitions in the forward chain (which is $p_j P_{ji}$) equals the proportion of transitions from $i$ to $j$ out of all transitions in the reversed chain (which is $p_i P^*_{ij}$).

2. If we can find positive numbers $p_i$, $i \geq 0$, summing to unity and such that the scalars

$$P^*_{ij} = \frac{p_j P_{ji}}{p_i}, \qquad i, j \geq 0 \tag{3.108}$$

form a transition probability matrix, that is,

$$\sum_{j=0}^{\infty} P^*_{ij} = 1, \qquad i = 0, 1, \ldots$$

then $\{p_i \mid i \geq 0\}$ is the stationary distribution and $P^*_{ij}$ are the transition probabilities of the reversed chain. [To see this, note that by multiplying with $p_i$ Eq. (3.108) and adding over $j$, we obtain

$$\sum_{j=0}^{\infty} p_j P_{ji} = p_i \sum_{j=0}^{\infty} P^*_{ij} = p_i$$

which is the global balance equation and implies that $\{p_i \mid i \geq 0\}$ is the stationary distribution.] This property, which holds regardless of whether the chain is time reversible, is useful if through an intelligent guess, we can verify Eq. (3.108), thereby obtaining both the $p_j$ and $P^*_{ij}$; for examples of such applications, see Section 3.8.

3. A chain is time reversible if and only if the detailed balance equations hold:

$$p_i P_{ij} = p_j P_{ji}, \qquad i, j \geq 0$$

This follows from the equality $p_i P^*_{ij} = p_j P_{ji}$ [cf. Eq. (3.107)] and the definition of time reversibility. In other words, a system is time reversible if in a typical system history, transitions from $i$ to $j$ occur with the same frequency as transitions from $j$ to $i$ (and therefore also with the same frequency as transitions from $i$ to $j$ when this system history is reversed in time). In particular, the chains corresponding to the queueing systems $M/M/1$, $M/M/m$, $M/M/\infty$, and $M/M/m/m$ discussed in Sections 3.3 and 3.4 are time reversible (in the limit as $\delta \to 0$). More generally, chains corresponding to birth–death processes ($P_{ij} = 0$ if $|i - j| > 1$) are time reversible. Figure 3.30 gives some additional examples of reversible and nonreversible systems.

The idea of time reversibility extends in a straightforward manner to irreducible continuous-time Markov chains. The corresponding analysis can be carried out either directly or by discretizing time in intervals of length $\delta$, considering the corresponding

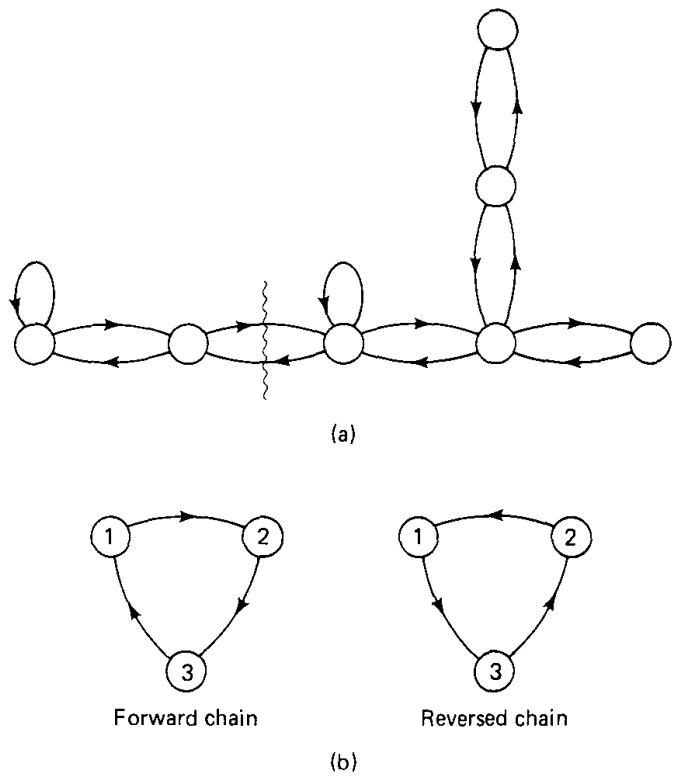



**Figure 3.30** (a) Example of a time reversible chain. To see this, note that by splitting the state space in two subsets as shown we obtain global balance equations which are identical with the detailed balance equations. (b) Example of a chain which is not time reversible. The states in the forward and the reversed systems move in the clockwise and counterclockwise directions, respectively.

discrete-time chain, and passing back to the continuous chain by taking the limit as $\delta \to 0$. All results regarding the reversed chain carry over almost verbatim from their discrete-time counterparts by replacing transition probabilities with transition rates. In particular, if the continuous-time chain has transition rates $q_{ij}$ and a stationary distribution $\{p_j \mid j \geq 0\}$ with $p_j > 0$ for all $j$, then:

1. The reversed chain is a continuous-time Markov chain with the same stationary distribution as the forward chain and with transition rates

$$q_{ij}^* = \frac{p_j q_{ji}}{p_i}, \qquad i, j \geq 0 \tag{3.109}$$

2. If we can find positive numbers $p_i$, $i \geq 0$, summing to unity and such that the scalars

$$q_{ij}^* = \frac{p_j q_{ji}}{p_i}, \qquad i, j \geq 0 \tag{3.110}$$

satisfy for all $i \geq 0$

$$\sum_{j=0}^{\infty} q_{ij} = \sum_{j=0}^{\infty} q_{ij}^* \tag{3.111}$$

then $\{p_i \mid i \geq 0\}$ is the stationary distribution of both the forward and the reversed chain, and $q_{ij}^*$ are the transition rates of the reversed chain. The relation $\sum_{j=0}^{\infty} q_{ij} = \sum_{j=0}^{\infty} q_{ij}^*$ equates, for every state $i$, the total rate out of $i$ in the forward and the reversed chains, and by taking into account also the relation $q_{ij}^* = p_j q_{ji}/p_i$, it can

be seen to be equivalent to the global balance equation

$$p_i \sum_{j=0}^{\infty} q_{ij} = \sum_{j=0}^{\infty} p_j q_{ji}$$

[cf. Eq. (3A.10) of Appendix A].

3. The forward chain is time reversible if and only if its stationary distribution and transition rates satisfy the detailed balanced equations

$$p_i q_{ij} = p_j q_{ji}\,, \qquad i, j \geq 0$$

Consider now the $M/M/1$, $M/M/m$, and $M/M/\infty$ queueing systems. We assume that the initial state is chosen according to the stationary distribution so that the queueing systems are in steady-state at all times. The reversed process can be represented by another queueing system where departures correspond to arrivals of the original system and arrivals correspond to departures of the original system (see Fig. 3.31). Because time reversibility holds for all these systems as discussed above, the forward and reversed systems are statistically indistinguishable in steady-state. In particular by using the fact that the departure process of the forward system corresponds to the arrival process of the reversed system, we obtain the following result:

**Burke's Theorem.** Consider an $M/M/1$, $M/M/m$, or $M/M/\infty$ system with arrival rate $\lambda$. Suppose that the system starts in steady-state. Then the following hold true:

**(a)** The departure process is Poisson with rate $\lambda$.

**(b)** At each time $t$, the number of customers in the system is independent of the sequence of departure times prior to $t$.

*Proof:* (a) This follows from the fact that the forward and reversed systems are statistically indistinguishable in steady-state, and the departure process in the forward system is the arrival process in the reversed system.

(b) As shown in Fig. 3.32, for a fixed time $t$, the departures prior to $t$ in the forward process are also the arrivals after $t$ in the reversed process. The arrival process in the reversed system is independent Poisson, so the future arrival process does not depend on the current number in the system, which in forward system terms means that the past departure process does not depend on the current number in the system. **Q.E.D.**

Note that part (b) of Burke's Theorem is quite counterintuitive. One would expect that a recent stream of closely spaced departures suggests a busy system with an atypically large number of customers in queue. Yet Burke's Theorem shows that this is not so. Note, however, that Burke's Theorem says nothing about the state of the system *before* a stream of closely spaced departures. Such a state would tend to have abnormally many customers in queue, in accordance with intuition.

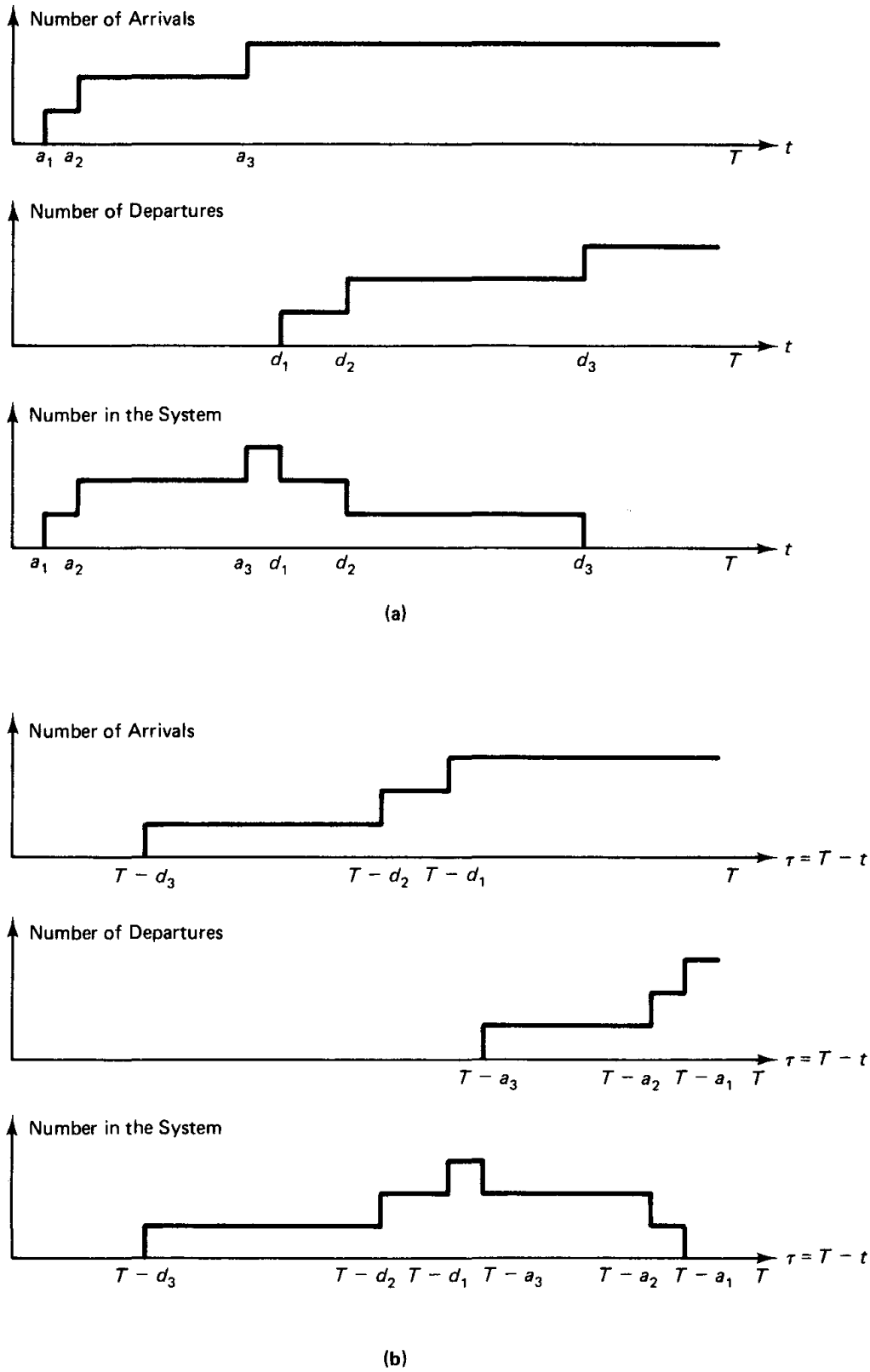



**Figure 3.31** (a) Forward system number of arrivals, number of departures, and occupancy during $[0, T]$. (b) Reversed system number of arrivals, number of departures, and occupancy during $[0, T]$.

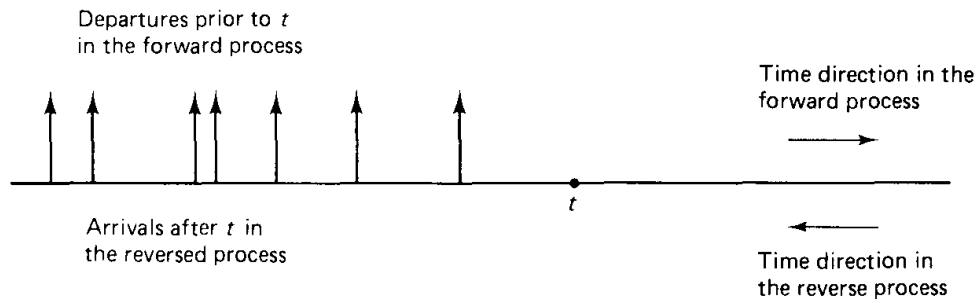



**Figure 3.32** Customer departures *prior* to time $t$ in the forward system become customer arrivals *after* time $t$ in the reversed system.

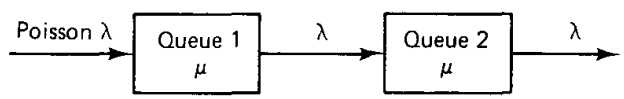



**Figure 3.33** Two queues in tandem. The service times at the two queues are exponentially distributed and mutually independent. Using Burke's Theorem, we can show that the number of customers in queues 1 and 2 are independent at a given time and

$$P\{n \text{ at queue } 1,\ m \text{ at queue } 2\} = \rho_1^n(1-\rho_1)\rho_2^m(1-\rho_2)$$

that is, the two queues behave as if they are independent $M/M/1$ queues in isolation.

### Example 3.18 Two $M/M/1$ Queues in Tandem

Consider a queueing network involving Poisson arrivals and two queues in tandem with exponential service times (see Fig. 3.33). There is a major difference between this system and the one discussed in Section 3.6 in that here we assume that the service times of a customer at the first and second queues are mutually independent as well as independent of the arrival process. As a result of this assumption, we will see that the occupancy distribution in the two queues is the same as if they were independent $M/M/1$ queues in isolation. This fact will also be shown in a more general context in the next section.

Let the rate of the Poisson arrival process be $\lambda$, and let the mean service times at queues 1 and 2 be $1/\mu_1$ and $1/\mu_2$, respectively. Let $\rho_1 = \lambda/\mu_1$ and $\rho_2 = \lambda/\mu_2$ be the corresponding utilization factors, and assume that $\rho_1 < 1$ and $\rho_2 < 1$. We will show that under steady-state conditions the number of customers at queue 1 and at queue 2 at any given time are independent. Furthermore,

$$P\{n \text{ at queue } 1,\ m \text{ at queue } 2\} = \rho_1^n(1-\rho_1)\rho_2^m(1-\rho_2) \tag{3.112}$$

To prove this we first note that queue 1 is an $M/M/1$ queue, so by part (a) of Burke's Theorem, the departure process from queue 1 is Poisson. By assumption, it is also independent of the service times at queue 2. Therefore, queue 2, viewed in isolation, is an $M/M/1$ queue. Thus, from the results of Section 3.1,

$$\begin{aligned} P\{n \text{ at queue } 1\} &= \rho_1^n(1-\rho_1) \\ P\{m \text{ at queue } 2\} &= \rho_2^m(1-\rho_2) \end{aligned} \tag{3.113}$$

From part (b) of Burke's Theorem it follows that the number of customers presently in queue 1 is independent of the sequence of earlier arrivals at queue 2 and therefore also of

the number of customers presently in queue 2. This implies that

$$P\{n \text{ at queue } 1,\ m \text{ at queue } 2\} = P\{n \text{ at queue } 1\} \cdot P\{m \text{ at queue } 2\}$$

and using Eq. (3.113) we obtain the desired product form (3.112).

We note that, by part (a) of Burke's Theorem, the arrival and the departure processes at both queues of the preceding example are Poisson. This fact can be similarly shown for a much broader class of queueing networks with Poisson arrivals and independent, exponentially distributed service times. We call such networks *acyclic* and define them as follows. We say that queue $j$ is a *downstream neighbor* of queue $i$ if there is a positive probability that a departing customer from queue $i$ will next enter queue $j$. We say that queue $j$ *lies downstream* of queue $i$ if there is a sequence of queues starting from $i$ and ending at $j$ such that each queue after $i$ in the sequence is a downstream neighbor of its predecessor. A queueing network is called acyclic if it is impossible to find two queues $i$ and $j$ such that $j$ lies downstream of $i$, and $i$ lies downstream of $j$. Having an acyclic network is essential for the Poisson character of the arrival and departure processes at each queue to be maintained (see Section 3.8). However, the product form (3.112) of the occupancy distribution generalizes in a natural way to networks that are not acyclic, as we show in the next section.

## 3.8 NETWORKS OF QUEUES—JACKSON'S THEOREM

As discussed in Section 3.6, the main difficulty with analysis of networks of transmission lines is that the packet interarrival times after traversing the first queue are correlated with their lengths. It turns out that if somehow this correlation were eliminated (which is the premise of the Kleinrock independence approximation) and randomization is used to divide traffic among different routes, then the average number of packets in the system can be derived as if each queue in the network were $M/M/1$. This is an important result known as Jackson's Theorem. In this section we derive a simple version of this theorem and some of its extensions.

Consider a network of $K$ first-come first-serve, single-server queues in which customers arrive from outside the network at each queue $i$ in accordance with independent Poisson processes at rate $r_i$. We allow the possibility that $r_i = 0$, in which case there are no external arrivals at queue $i$, but we require that $r_i > 0$ for at least one $i$. Once a customer is served at queue $i$, it proceeds to join each queue $j$ with probability $P_{ij}$ or to exit the network with probability $1 - \sum_{j=1}^{K} P_{ij}$.

The routing probabilities $P_{ij}$ together with the external input rates $r_j$ can be used to determine the total arrival rate of customers $\lambda_j$ at each queue $j$, that is, the sum of $r_j$ and the arrival rate of customers coming from other queues. Calculating $\lambda_j$ is fairly easy when the network is of the acyclic type discussed at the end of Section 3.7. If there is a positive probability that a customer may visit the same queue twice, a more complex

computation is necessary, based on the equations

$$\lambda_j = r_j + \sum_{i=1}^{K} \lambda_i P_{ij}, \qquad j = 1, \ldots, K \tag{3.114}$$

These equations represent a linear system in which the rates $\lambda_j$, $j = 1, \ldots, K$, constitute a set of $K$ unknowns. To guarantee that they can be solved uniquely to yield $\lambda_j$, $j = 1, \ldots, K$ in terms of $r_j$, $P_{ij}$, $i$, $j = 1, \ldots, K$, we make a fairly natural assumption that essentially asserts that each customer will eventually exit the system with probability 1. This assumption is that for every queue $i_1$, there is a queue $i$ with $1 - \sum_{j=1}^{K} P_{ij} > 0$ and a sequence $i_1, i_2, \ldots, i_k, i$ such that $P_{i_1 i_2} > 0, \ldots, P_{i_k i} > 0$.*

The service times of customers at the $j^{\text{th}}$ queue are assumed exponentially distributed with mean $1/\mu_j$ and are assumed mutually independent and independent of the arrival process at the queue. The utilization factor of each queue is denoted

$$\rho_j = \frac{\lambda_j}{\mu_j}, \qquad j = 1, \ldots, K \tag{3.115}$$

and we assume that $\rho_j < 1$ for all $j$.

In order to model a packet network such as the one considered in Section 3.6 within the framework described above, it is necessary to accept several simplifying conditions in addition to assuming Poisson arrivals and exponentially distributed packet lengths. The first is the independence of packet lengths and interarrival times discussed earlier. The second is relevant to datagram networks, and has to do with the assumption that bifurcation of traffic at a network node can be modeled reasonably well by a randomization process whereby each departing packet from queue $i$ joins queue $j$ with probability $P_{ij}$—this need not be true, as discussed in Section 3.6. Still a packet network differs from the model of this section because it involves several traffic streams which may have different routing probabilities at each node, and which maintain their identity as they travel along different routes (see the virtual circuit and datagram network models of Figs. 3.27 and 3.28). This difficulty can be partially addressed by using an extension of Jackson's Theorem that applies to a network with multiple classes of customers. Within this more general framework, we can model traffic streams corresponding to different origin–destination pairs as different classes of customers. If all traffic streams have the same average packet length, it turns out that Jackson's Theorem as stated below is valid assuming the simplifying conditions mentioned earlier; see the analysis in the next subsection.

*For a brief explanation aimed at the advanced reader, consider the Markov chain with states $0, 1, \ldots, K$ and transition probabilities from states $i \neq 0$ to states $j \neq 0$ equal to $P_{ij}$, and transition probabilities to state 0 equal to $P_{00} = 1$, $P_{i0} = 1 - \sum_{j=1}^{K} P_{ij}$ for $i \neq 0$. (Thus state 0 is an absorbing state that corresponds to exit of a customer from the system.) Let $P$ be the $K \times K$ matrix with elements $P_{ij}$. The sum of the $i^{\text{th}}$ row elements of the matrix $P^m$ ($P$ to the $m^{\text{th}}$ power) is the probability that the Markov chain has not arrived at state 0 after $m$ transitions starting from state $i$. Our hypothesis on $P_{ij}$ implies that the chain will eventually (with probability 1) arrive at state 0 regardless of the initial state. It follows that $\lim_{m\to\infty} P^m = 0$, so unity is not an eigenvalue of $P$. Therefore, $I - P$ is nonsingular, where $I$ is the identity matrix, from which it can be seen that the system of equations (3.114) has a unique solution.

For an analysis, we view the system as a continuous-time Markov chain in which the state $n$ is the vector $(n_1, n_2, \ldots, n_K)$, where $n_i$ denotes the number of customers at queue $i$. At a given state

$$n = (n_1, n_2, \ldots, \ldots, n_K)$$

the possible successor states correspond to a single customer arrival and/or departure. In particular, the transition from $n$ to state

$$n(j^+) = (n_1, \ldots, n_{j-1}, n_j + 1, n_{j+1}, \ldots, n_K)$$

corresponding to an external arrival at queue $j$, has transition rate

$$q_{nn(j^+)} = r_j$$

The transition from $n$ to state

$$n(j^-) = (n_1, \ldots, n_{j-1}, n_j - 1, n_{j+1}, \ldots, n_K)$$

corresponding to a departure from queue $j$ to the outside, has transition rate

$$q_{nn(j^-)} = \mu_j \left(1 - \sum_i P_{ji}\right)$$

The transition from $n$ to state

$$n(i^+, j^-) = (n_1, \ldots, n_{i-1}, n_i + 1, n_{i+1}, \ldots, n_{j-1}, n_j - 1, n_{j+1}, \ldots, n_K)$$

corresponding to a customer moving from queue $j$ to queue $i$, has transition rate

$$q_{nn(i^+, j^-)} = \mu_j P_{ji}$$

Let $P(n_1, \ldots, n_K)$ denote the stationary distribution of the chain. We have:

**Jackson's Theorem.** Assuming that $\rho_j < 1$, $j = 1, \ldots, K$, we have for all $n_1, \ldots, n_K \geq 0$,

$$P(n) = P_1(n_1) P_2(n_2) \cdots P_K(n_K) \tag{3.116}$$

where $n = (n_1, \ldots, n_K)$ and

$$P_j(n_j) = \rho_j^{n_j}(1 - \rho_j), \qquad n_j \geq 0 \tag{3.117}$$

*Proof:* In our proof we will assume that $\lambda_j > 0$ for all $j$. There is no loss of generality in doing so because every queue $j$ with $\lambda_j = 0$ is empty in steady-state, so we have $P_j(0) = 1$ and $P_j(n_j) = 0$ for $n_j > 0$, and queue $j$ can be ignored in deriving the stationary distribution of Eqs. (3.116) and (3.117). It can be verified that the condition $\lambda_j > 0$ for all $j$ together with the assumption made earlier to guarantee the uniqueness of solution of Eq. (3.114) imply that the Markov chain with states $n = (n_1, \ldots, n_K)$ describing the system is irreducible; we leave the proof of this for the reader. We will use a technique outlined in Section 3.7 whereby we guess at the transition rates of the reversed process and verify that, together with the probability distribution of Eqs. (3.116) and (3.117), they satisfy the total departure rate equation (3.111). (The Markov chain is

not time reversible here. Nonetheless, the use of the reversed process is both analytically convenient and conceptually useful.)

For any two state vectors $n$ and $n'$, let $q_{nn'}$ be the corresponding transition rate. Jackson's Theorem will be proved if the rates $q^*_{nn'}$ defined for all $n, n'$ by the equation

$$q^*_{nn'} = \frac{P(n')q_{n'n}}{P(n)} \tag{3.118}$$

satisfy, for all $n$, the total rate equation

$$\sum_m q_{nm} = \sum_m q^*_{nm} \tag{3.119}$$

which as mentioned in Section 3.7, is equivalent to the global balance equations.

For transitions between states $n$, $n(j^+)$, and $n(j^-)$, we have

$$q_{nn(j^+)} = r_j \tag{3.120}$$

$$q_{nn(j^-)} = \mu_j \left(1 - \sum_i P_{ji}\right) \tag{3.121}$$

The rates $q^*_{nn(j^+)}$ and $q^*_{nn(j^-)}$ are defined by Eqs. (3.118), (3.120), and (3.121). Using the fact $P\big(n(j^+)\big) = \rho_j P(n) = \lambda_j P(n)/\mu_j$ [cf. Eqs. (3.115)–(3.117)], we obtain

$$q^*_{nn(j^+)} = \lambda_j \left(1 - \textstyle\sum_i P_{ji}\right) \tag{3.122}$$

$$q^*_{nn(j^-)} = \frac{\mu_j r_j}{\lambda_j} \tag{3.123}$$

Next consider transitions between states $n$ and $n(i^+, j^-)$ corresponding to a customer moving from queue $j$ to queue $i$. We have

$$q_{nn(i^+,j^-)} = \mu_j P_{ji} \tag{3.124}$$

and using the fact that $P\big(n(i^+, j^-)\big) = \rho_i P(n)/\rho_j = \lambda_i \mu_j P(n)/(\lambda_j \mu_i)$, we obtain $q^*_{n(i^+,j^-)n}$ from Eq. (3.118) as

$$q^*_{n(i^+,j^-)n} = \frac{\mu_i \lambda_j P_{ji}}{\lambda_i} \tag{3.125}$$

Since for all other types of pairs of state vectors $n$, $n'$, we have

$$q_{nn'} = 0 \tag{3.126}$$

it follows from Eq. (3.118) that

$$q^*_{n'n} = 0 \tag{3.127}$$

There remains to verify that the rates $q_{nm}$ and $q^*_{nm}$ satisfy the total rate equation $\sum_m q_{nm} = \sum_m q^*_{nm}$. We have for the forward system, using Eqs. (3.120), (3.121), and (3.124),

$$\sum_m q_{nm} = \sum_{j=1}^{K} q_{nn(j^+)} + \sum_{\{(j,i)|n_j>0\}} q_{nn(i^+,j^-)} + \sum_{\{j|n_j>0\}} q_{nn(j^-)}$$

$$= \sum_{j=1}^{K} r_j + \sum_{\{(j,i)|n_j>0\}} \mu_j P_{ji} + \sum_{\{j|n_j>0\}} \mu_j \left(1 - \sum_{i=1}^{K} P_{ji}\right)$$

$$= \sum_{j=1}^{K} r_j + \sum_{\{j|n_j>0\}} \mu_j \tag{3.128}$$

Similarly, using Eqs. (3.122), (3.123), (3.125), and (3.114), we obtain for the reversed system

$$\sum_m q^*_{nm} = \sum_{j=1}^{K} q^*_{nn(j^+)} + \sum_{\{(j,i)|n_j>0\}} q^*_{nn(i^+,j^-)} + \sum_{\{j|n_j>0\}} q^*_{nn(j^-)}$$

$$= \sum_{j=1}^{K} \lambda_j \left(1 - \sum_{i=1}^{K} P_{ji}\right) + \sum_{\{(j,i)|n_j>0\}} \frac{\mu_j \lambda_i P_{ij}}{\lambda_j} + \sum_{\{j|n_j>0\}} \frac{\mu_j r_j}{\lambda_j}$$

$$= \sum_{j=1}^{K} \lambda_j \left(1 - \sum_{i=1}^{K} P_{ji}\right) + \sum_{\{j|n_j>0\}} \frac{\mu_j (r_j + \sum_{i=1}^{K} \lambda_i P_{ij})}{\lambda_j}$$

$$= \sum_{j=1}^{K} \lambda_j \left(1 - \sum_{i=1}^{K} P_{ji}\right) + \sum_{\{j|n_j>0\}} \mu_j \tag{3.129}$$

By writing Eq. (3.114) as $r_j = \lambda_j - \sum_{i=1}^{K} \lambda_i P_{ij}$ and adding over $j = 1, \ldots, K$, we obtain

$$\sum_{j=1}^{K} r_j = \sum_{j=1}^{K} \lambda_j \left(1 - \sum_{i=1}^{K} P_{ji}\right) \tag{3.130}$$

By combining the last three equations, we see that the total rate equation $\sum_m q_{nm} = \sum_m q^*_{nm}$ is satisfied. **Q.E.D.**

Note that the transition rates $q^*_{nn'}$ defined by Eqs. (3.122), (3.123), (3.125), and (3.127) are those of the reversed process. It can be seen that the reversed process corresponds to a network of queues where traffic arrives at queue $i$ from outside the network according to a Poisson process with rate $\lambda_i \left(1 - \sum_j P_{ij}\right)$ [cf. Eq. (3.122)]. The routing probability from queue $i$ to queue $j$ in the reversed process is

$$\frac{\lambda_j P_{ji}}{r_i + \sum_k \lambda_k P_{ki}}$$

[cf. Eqs. (3.123) and (3.125)]. This is also the probability that an arriving customer at queue $i$ just departed from queue $j$ in the forward process. Note that the processes of departure out of the forward system are the exogenous arrival processes of the reversed system, which suggests that the processes of departure out of the system are independent

Poisson. Indeed, this can be proved by observing that the interarrival times in the reversed system are independent and exponentially distributed.

**Example 3.19 Computer System with Feedback Loop for I/O**

Consider a model of a computer CPU connected to an I/O device as shown in Fig. 3.34(a). Jobs enter the system according to a Poisson process with rate $\lambda$, and use the CPU for an exponentially distributed time interval with mean $1/\mu_1$. Upon exiting the CPU, a job with probability $p_1$ exits the system, and with probability $p_2$ $(= 1 - p_1)$ uses the I/O device for a time which is exponentially distributed with mean $1/\mu_2$. Upon exit from the I/O device, a job again joins the CPU queue. We assume that all service times, including successive service times of the same job at the CPU or the I/O device, are independent.

We first calculate the arrival rates $\lambda_1$ and $\lambda_2$ at the CPU and I/O device queues, respectively. We have (cf. Fig. 3.34)

$$\lambda_1 = \lambda + \lambda_2, \qquad \lambda_2 = p_2\lambda_1$$

[These are Eqs. (3.114) specialized to this example.] By solving for $\lambda_1$ and $\lambda_2$ we obtain

$$\lambda_1 = \frac{\lambda}{p_1}, \qquad \lambda_2 = \frac{\lambda p_2}{p_1} \tag{3.131}$$

Let

$$\rho_1 = \frac{\lambda_1}{\mu_1}, \qquad \rho_2 = \frac{\lambda_2}{\mu_2} \tag{3.132}$$

The steady-state probability distribution of the system is given by Jackson's Theorem

$$P(n_1, n_2) = \rho_1^{n_1}(1 - \rho_1)\rho_2^{n_2}(1 - \rho_2)$$

The average number of jobs $N_i$ in the $i^{\text{th}}$ queue is the same as for an $M/M/1$ system with utilization factor $\rho_i$, that is,

$$N_1 = \frac{\rho_1}{1 - \rho_1}, \qquad N_2 = \frac{\rho_2}{1 - \rho_2}$$

The total number in the system is

$$N = N_1 + N_2 = \frac{\rho_1}{1 - \rho_1} + \frac{\rho_2}{1 - \rho_2}$$

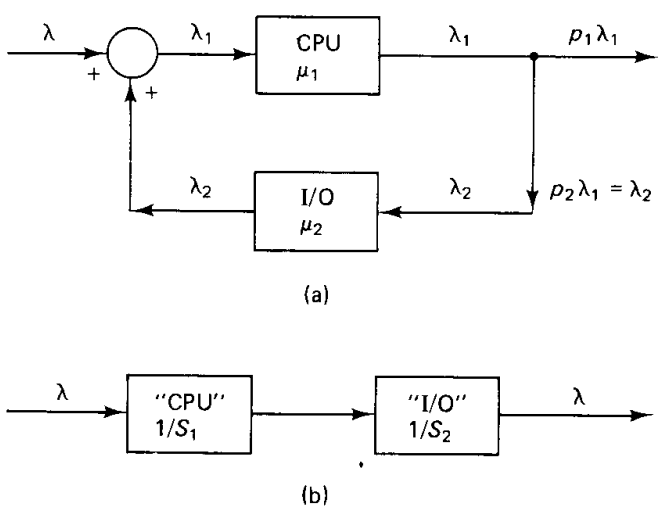



**Figure 3.34** (a) Feedback model of a CPU and an I/O device (cf. Example 3.19). (b) "Equivalent" tandem model of a CPU and an I/O queue which has the same occupancy distribution as the feedback model.

and the average time in the system is

$$T = \frac{N}{\lambda} = \frac{\rho_1}{\lambda(1-\rho_1)} + \frac{\rho_2}{\lambda(1-\rho_2)}$$

Using Eqs. (3.131) and (3.132) we can write this relation as

$$T = \frac{\lambda_1/\mu_1}{\lambda(1-\lambda_1/\mu_1)} + \frac{\lambda_2/\mu_2}{\lambda(1-\lambda_2/\mu_2)} = \frac{\lambda/(\mu_1 p_1)}{\lambda\big(1-\lambda/(\mu_1 p_1)\big)} + \frac{\lambda p_2/(\mu_2 p_1)}{\lambda\big(1-\lambda p_2/(\mu_2 p_1)\big)}$$

$$= \frac{S_1}{1-\lambda S_1} + \frac{S_2}{1-\lambda S_2} \tag{3.133}$$

where

$$S_1 = \frac{1}{\mu_1 p_1}, \qquad S_2 = \frac{p_2}{\mu_2 p_1} \tag{3.134}$$

Since the utilization factor of the CPU queue is $\rho_1 = \lambda_1/\mu_1 = \lambda/(\mu_1 p_1)$, while the arrival rate of *new* job arrivals at the CPU (as opposed to feedback arrivals) is $\lambda$, we see from Little's Theorem that $S_1$ is the total CPU time a job requires on the average (this includes all visits of the job to the CPU). Similarly, $S_2$ is the total I/O time a job requires on the average.

An interesting interpretation of Eqs. (3.133) and (3.134) is that the average number of jobs and time in the system are the same as in an "equivalent" tandem model of CPU and I/O queues with service rates $1/S_1$ and $1/S_2$, respectively, as shown in Fig. 3.34(b). However, the probability density function of the time in the system is not the same in the feedback and tandem systems. To get some idea of this fact, suppose that $p_1 = p_2 = 1/2$ and that the CPU service rate is much faster than the I/O service rate ($\mu_1 >> \mu_2$). Then half the jobs in the feedback system do not require any I/O service and their average time in the system is much smaller than the average time of the other half. This is not so in the tandem system where the average job time in the "CPU" queue is very small and the system time is distributed approximately as in the "I/O" queue, that is, as in an $M/M/1$ queue with Poisson rate $\lambda$ and service rate $1/S_1$.

Jackson's Theorem says in effect that the numbers of customers in the system's queues are distributed as if each queue is $M/M/1$ and is independent of the other queues [compare Eq. (3.117) and the corresponding equations in Section 3.3]. Despite this fact, *the total arrival process at each queue need not be Poisson*. As an example (see Fig. 3.35), suppose that there is a single queue with a service rate which is very large relative to the arrival rate from the outside. Suppose also that with probability $p$ near unity, a customer upon completion of service is fed back into the queue. Hence, when an arrival occurs at the queue, there is a large probability of another arrival at the queue in a short time (namely, the feedback arrival), whereas at an arbitrary time point, there will be only a very slight chance of an arrival occurring shortly since $\lambda$ is small. In other words, queue arrivals tend to occur in bursts triggered by the arrival of a single customer from the outside. Hence, the queue arrival process does not have independent interarrival times and cannot be Poisson.

**Heuristic explanation of Jackson's Theorem.** Our proof of Jackson's theorem is based on algebraic manipulation and gives little insight as to why this remarkable result holds. For this reason we provide a heuristic explanation for the case of the feed-

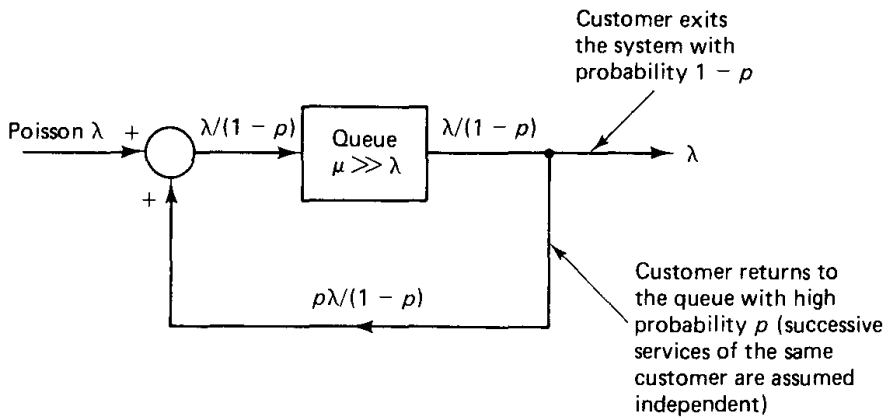



**Figure 3.35** Example of a queue within a network where the external arrival process is Poisson but the total arrival process at the queue is not Poisson. An external arrival is typically processed fast (since $\mu$ is much larger than $\lambda$) and with high probability returns to the queue through the feedback loop. As a result, the total queue arrival process typically consists of bursts of arrivals, with each burst triggered by the arrival of a single customer from the outside.

back network of Fig. 3.35. This explanation can be generalized and made rigorous albeit at the expense of a great deal of technical complications (see [Wal83]).

Suppose that we introduce a delay $\Delta$ in the feedback loop of the single-queue network discussed above (see Fig. 3.36). Let us denote by $n(t)$ the number in the queue at time $t$, and by $f_\Delta(t)$ the content of the delay line at time $t$. The interpretation here is that $f_\Delta(t)$ is a function of time that specifies the customer output of the delay line in the subsequent $\Delta$ interval $(t, t+\Delta]$. Suppose that the initial distribution $n(0)$ of the queue state at time 0, is equal to the steady-state distribution of an $M/M/1$ queue, that is,

$$P\{n(0) = n\} = \rho^n(1-\rho) \tag{3.135}$$

where $\rho = \lambda/\big(\mu(1-p)\big)$ is the utilization factor. Suppose also that $f_\Delta(0)$ is a portion of a Poisson arrival process with rate $\lambda$. The customers in $f_\Delta(0)$ have service times that are independent, exponentially distributed with parameter $\mu$. We assume that $n(0)$ and $f_\Delta(0)$ are independent. Then, the input to the queue over the interval $[0, \Delta)$ will

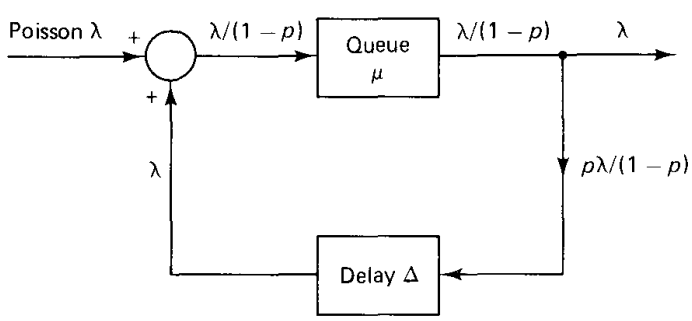



**Figure 3.36** Heuristic explanation of Jackson's Theorem. Consider the introduction of an arbitrarily small positive delay $\Delta$ in the feedback loop of the network of Fig. 3.35. An occupancy distribution of the queue that equals the $M/M/1$ equilibrium, and a content of the delay line that is an independent $\Delta$ segment of a Poisson process form an equilibrium distribution of the overall system. Therefore, the $M/M/1$ equilibrium distribution is an equilibrium for the queue as suggested by Jackson's Theorem even though the total arrival process to the queue is not Poisson.

be the sum of two independent Poisson streams which are independent of the number in queue at time 0. It follows that the queue will behave in the interval $[0, \Delta)$ like an $M/M/1$ queue in equilibrium. Therefore, $n(\Delta)$ will be distributed according to the $M/M/1$ steady-state distribution of Eq. (3.135), and by part (b) of Burke's theorem, $n(\Delta)$ will be independent of the departure process from the queue in the interval $[0, \Delta)$, or, equivalently, of $f_\Delta(\Delta)$ —the delay line content at time $\Delta$. Furthermore, by part (a) of Burke's Theorem, $f_\Delta(\Delta)$ will be Poisson. Thus, to summarize, we started out with independent initial conditions $n(0)$ and $f_\Delta(0)$ which had the equilibrium distribution of an $M/M/1$ queue and the statistics of a Poisson process, respectively, and $\Delta$ seconds later we obtained corresponding quantities $n(\Delta)$ and $f_\Delta(\Delta)$ with the same properties. Using the same reasoning, we can show that for all $t$ which are multiples of $\Delta$, $n(t)$ and $f_\Delta(t)$ have the same properties. It follows that the $M/M/1$ steady-state distribution of Eq. (3.135) is an equilibrium distribution for the queueing system for an arbitrary positive value of the feedback delay $\Delta$, and this strongly suggests the validity of Jackson's Theorem. Note that this argument does not suggest that the feedback process, and therefore also the total arrival process to the queue, are Poisson. Indeed, it can be seen that successive $\Delta$ portions of the feedback arrival stream are correlated since, with probability $p$, a departing customer from the queue appears as an arrival $\Delta$ seconds later. Therefore, over the interval $[0, \infty)$, the feedback process is not Poisson. This is consistent with our earlier observations regarding the example of Fig. 3.35.

### 3.8.1 Extensions of Jackson's Theorem

There are a number of interesting extensions and variations of Jackson's Theorem, and in this and the next subsections we will describe a few of them.

**State-dependent service rates.** The model for Jackson's Theorem assumed so far requires that all queues have a single server. An extension to the multiserver case can be obtained by allowing the service rate at each queue to depend on the number of customers at that queue. Thus the model is the same as before but the service time at the $j^{\text{th}}$ queue is exponentially distributed with mean $1/\mu_j(m)$, where $m$ is the number of customers in the $j^{\text{th}}$ queue just before the customer's departure ($m$ includes the customer). The single-queue version of this model includes as special cases the $M/M/m$ and $M/M/\infty$ queues, and can be analyzed by means of a Markov chain (see Problem 3.16). The corresponding network of queues model can also be analyzed by means of a Markov chain, and is characterized by a product form structure for the stationary distribution.

Let us define

$$\rho_j(m) = \frac{\lambda_j}{\mu_j(m)}, \qquad j = 1, \ldots, K, \quad m = 1, 2, \ldots \tag{3.136}$$

where $\lambda_j$ is the total arrival rate at the $j^{\text{th}}$ queue determined by Eq. (3.114). Let us also define

$$\hat{P}_j(n_j) = \begin{cases} 1, & \text{if } n_j = 0 \\ \rho_j(1)\rho_j(2)\cdots\rho_j(n_j), & \text{if } n_j > 0 \end{cases} \tag{3.137}$$

We have:

**Jackson's Theorem for State-Dependent Service Rates.** We have for all states $n = (n_1, \ldots, n_K)$

$$P(n) = \frac{\hat{P}_1(n_1) \cdots \hat{P}_K(n_K)}{G} \tag{3.138}$$

assuming that $0 < G < \infty$, where the normalizing constant $G$ is given by

$$G = \sum_{n_1=0}^{\infty} \cdots \sum_{n_K=0}^{\infty} \hat{P}_1(n_1) \cdots \hat{P}_K(n_K) \tag{3.139}$$

*Proof:* Note that the formula for $G$ guarantees that $P(n)$ is a probability distribution, that is, the sum of all $P(n)$ is unity. Using this fact, the proof is obtained by repeating the steps of the earlier proof of Jackson's Theorem, substituting the state-dependent service rates $\mu_j(m)$ in place of the rates $\mu_j$ at the appropriate points, and is left for the reader. **Q.E.D.**

**Multiple classes of customers.** In many interesting networks of queues the routing probabilities $P_{ij}$ are not the same for all customers. Typical examples arise in data networks where the transmission queue joined by a packet at each intermediate node depends on the packet's destination and possibly its origin. It is therefore necessary to distinguish between customers of different types or classes. We will show that the product form expressions derived so far remain valid provided that the service time distribution at each queue is the same for all customer classes.

Let the customer classes be $c = 1, 2, \ldots, C$, let $r_j(c)$ be the rate of the external Poisson arrival process of class $c$ at queue $j$, and let $P_{ij}(c)$ be the routing probabilities of class $c$. The assumptions made for an open Jackson network with a single customer class are replicated for each customer class, so that the equations

$$\lambda_j(c) = r_j(c) + \sum_{i=1}^{K} \lambda_i(c) P_{ij}(c), \qquad j = 1, \ldots, K, \quad c = 1, 2, \ldots, C \tag{3.140}$$

can be solved uniquely to give the total arrival rate $\lambda_j(c)$ at each queue $j$ and for each customer class $c$. We assume that the service times at queue $j$ are exponentially distributed with a common mean $1/\mu_j(m)$ for all customer classes, which depends on $m$, the total number of customers in the queue. As earlier, customers are served on a first-come first-serve basis.

The state of each queue is characterized not just by the total number of customers present in the queue, but also by the class of the customers and the relative order of arrival of the customers of different classes. Thus, we define the *composition of the* $j^{\text{th}}$ *queue* at a given time as

$$z_j = (c_1, c_2, \ldots, c_{n_j})$$

where $n_j$ is the total number of customers in the queue and $c_i$ is the class of the customer in the $i^{\text{th}}$ queue position.

The state of the queueing network at a given time is

$$z = (z_1, z_2, \ldots, z_K)$$

where $z_j$ is the composition of the $j^{\text{th}}$ queue at that time. It can be viewed as the state of a Markov chain the transition probabilities of which can be described in terms of the given quantities $\lambda_j(c)$, $\mu_j(m)$, and $P_{ij}(c)$. To state the appropriate form of Jackson's Theorem, define

$$\hat{\rho}_j(c, m) = \frac{\lambda_j(c)}{\mu_j(m)}, \qquad j = 1, \ldots, K, \quad c = 1, 2, \ldots, C \tag{3.141}$$

$$\hat{P}_j(z_j) = \begin{cases} 1, & \text{if } n_j = 0 \\ \hat{\rho}_j(c_1, 1)\hat{\rho}_j(c_2, 2) \cdots \hat{\rho}_j(c_{n_j}, n_j), & \text{if } n_j > 0 \end{cases} \tag{3.142}$$

$$G = \sum_{(z_1, \ldots, z_K)} \prod_{j=1}^{K} \hat{P}_j(z_j) \tag{3.143}$$

The proof of the following theorem follows the same pattern as the corresponding proof for the single customer class case, and is left for the reader.

**Jackson's Theorem for Multiple Classes of Customers.** Assuming that $0 < G < \infty$, the steady-state probability $\hat{P}(z)$ of state $z = (z_1, z_2, \ldots, z_K)$ is given by

$$\hat{P}(z) = \frac{\hat{P}_1(z_1) \cdots \hat{P}_K(z_K)}{G} \tag{3.144}$$

The steady-state probability $P(n) = P(n_1, \ldots, n_K)$ of having a total of $n_j$ customers at queue $j = 1, \ldots, K$ (irrespective of class) is given by

$$P(n) = \sum_{z \in Z(n)} \hat{P}(z)$$

where $Z(n)$ is the set of states for which there is a total of $n_j$ customers in queue $j$. By adding the expression (3.144) over $z \in Z(n)$, it is straighforward to verify that when the service rate at each queue is the same for all customer classes and is independent of the queue size, we have

$$P(n) = \prod_{j=1}^{K} \rho_j^{n_j}(1 - \rho_j) \tag{3.145}$$

where

$$\rho_j = \frac{\sum_{c=1}^{C} \lambda_j(c)}{\mu_j} \tag{3.146}$$

and $\mu_j$ is the service rate at queue $j$. In other words, the expression for $P(n)$ is the same as when there is a single customer class with total arrival rate at each queue $j$ equal to the sum of the arrival rates of all customer classes $\sum_{j=1}^{C} \lambda_j(c)$.

We note that when the service rates at the queues are state dependent (but identical for all classes) the steady-state probabilities $P(n)$ can be shown to be given by the (single class) formulas (3.136) to (3.139). (See the references cited at the end of the chapter.)

The following example addresses the first data network model discussed in Section 3.6 (cf. Fig. 3.27). A similar analysis can be used for the datagram network model of Fig. 3.28.

**Example 3.20 Virtual Circuit Network**

Consider the network of communication links discussed in Section 3.5 (cf. Fig. 3.27). There are several traffic streams (or virtual circuits) denoted $c = 1, 2, \ldots, C$. Virtual circuit $c$ uses a path $p_c$ and has a Poisson arrival rate $x_c$. The total arrival rate of each link $(i, j)$ is

$$\lambda_{ij} = \sum_{\{c \mid (i,j) \text{ lies on the path } p_c\}} x_c$$

Assume that the transmission times of all packets at link $(i, j)$ are exponentially distributed with mean $1/\mu_{ij}$, which is the same for all virtual circuits. Assume also that the transmission times of all packets are independent, including the transmission times of the same packet at two different links (this is the essence of the Kleinrock independence approximation). Then the multiple-class model of this subsection applies and based on Eq. (3.145), the average number of packets in the system, $N$, is the same as if each link were an $M/M/1$ queue in isolation, that is,

$$N = \sum_{(i,j)} \frac{\lambda_{ij}}{\mu_{ij} - \lambda_{ij}}$$

Example 3.20 shows how multiple customer classes can be used to model data network situations where the route used by a packet depends on its origin and destination. There is still an unrealistic assumption in this example, namely that the transmission times of the same packet at two different links are independent. Furthermore, the assumption that all packet transmission times are exponentially distributed with common mean is often violated in practice. For a more realistic model, we would like to be able to assume more general transmission time distributions (*e.g.*, deterministic transmission times). It turns out that the product form of Eqs. (3.141) to (3.144) holds even when the service time distributions belong to a broad class of "phase-type" distributions, which can approximate arbitrarily closely deterministic service times (see [GeP87] and [Wal88]). For this, however, we need to assume that the service discipline at each queue is either *processor sharing* or *last-come first-serve* instead of first-come first-serve. Processor sharing refers to a situation where all customers in the queue are simultaneously served at the same rate (which is $\mu/n$ when $\mu$ is the total service rate and $n$ is the number of customers). Last-come first-serve refers to the situation where upon arrival at a queue, a customer goes immediately into service, replacing the customer who is in service at the time (if any) on a preemptive-resume basis. While processor sharing or last-come first-serve may not be reasonable models for most data networks, the validity of the product form expression (3.141) to (3.144) under a variety of different assumptions is reassuring. It suggests that product forms provide a good first approximation in many practical situations where their use cannot be rigorously justified. Current practice and

experience seems to be supporting this view. We note, however, that for special types of priority disciplines, there are queueing networks that are unstable (some queue lengths grow indefinitely) even though the arrival rate is smaller than the service rate at each queue [KuS89]. We refer to the sources given at the end of the chapter for more details and discussion on the subject.

### 3.8.2 Closed Queueing Networks

Many interesting queueing problems involve a network of queues where the total number of customers is fixed because no customers are allowed to arrive or depart. Networks of this type are called *closed*, emphasizing the distinction from the earlier networks in this section which are called *open*. Examples 3.5 and 3.7 illustrate applications of closed networks. In both examples the fixed number of customers in the network depends on some limited resource, and the main purpose of analysis is to understand how the availability of this resource affects performance characteristics such as system throughput.

Closed networks can also be analyzed using Markov chains and it can be shown that the steady-state occupancy distribution has a product form under assumptions similar to those used earlier for open networks. For simplicity, we assume a single customer class, but extensions involving multiple customer classes are possible. Let $M$ be the fixed number of customers in the system and let $P_{ij}$ be the routing probability that a customer that departs from queue $i$ will next visit queue $j$. Note that because no customer can exit the system, we have

$$\sum_{j=1}^{K} P_{ij} = 1, \qquad i = 1, \ldots, K$$

Let also $\mu_j(m)$ be the service rate at the $j^{\text{th}}$ queue when the number of customers at that queue is $m$.

An important difference from the open network case is that the total arrival rates, denoted $\lambda_j(M)$, at the queues $j = 1, \ldots, K$ are not easily determined. We still have the equations

$$\lambda_j = \sum_{i=1}^{K} \lambda_i P_{ij}, \qquad j = 1, \ldots, K \tag{3.147}$$

obtained by setting to zero the external arrival rates $r_j$ in Eq. (3.114). These equations do not have a unique solution anymore, but under some fairly natural assumptions, they determine the arrival rates $\lambda_j(M)$ up to a multiplicative constant. In particular, let us assume that the Markov chain with states $1, \ldots, K$ and transition probabilities $P_{ij}$ is irreducible (see Appendix A). Then it can be shown that all solutions $\lambda_j$, $j = 1, \ldots, K$, of Eq. (3.147) are of the form*

*For a brief explanation, fix $\lambda_1$ at some positive value $a$ and consider the system of equations $\lambda_j = aP_{1j} + \sum_{i=2}^{K} \lambda_i P_{ij}$, $j = 2, \ldots, K$. Because of the irreducibility assumption, this system has a unique solution. [See the explanation given in connection with the uniqueness of solution of the corresponding open network equation (3.114).] This unique solution is proportional to $a$ and it can be shown to have positive elements.

$$\lambda_j = \alpha\overline{\lambda}_j, \qquad j = 1, \ldots, K$$

where $\alpha$ is a scalar and $\overline{\lambda}_j$, $j = 1, \ldots, K$ is a particular solution with $\overline{\lambda}_j > 0$ for all $j$. Thus the true arrival rates are given by

$$\lambda_j(M) = \alpha(M)\overline{\lambda}_j, \qquad j = 1, \ldots, K \tag{3.148}$$

where $\alpha(M)$ is the constant of proportionality corresponding to $M$. Note that while $\overline{\lambda}_j$ can be chosen to be independent of $M$, both $\alpha(M)$ and the true total arrival rates $\lambda_j(M)$ increase with $M$. In the case where the queue service rates $\mu_j$ are independent of the number of customers, $\alpha(M)$ tends asymptotically to the value that makes the maximum utilization factor $\max\{\lambda_1(M)/\mu_1, \ldots, \lambda_K(M)/\mu_K\}$ equal to one.

We now describe the form of Jackson's Theorem for closed networks. Let

$$\rho_j(m) = \frac{\overline{\lambda}_j}{\mu_j(m)}, \qquad j = 1, \ldots, K, \quad m = 1, 2, \ldots \tag{3.149}$$

where $\{\overline{\lambda}_j \mid j = 1, \ldots, K\}$ is some positive solution of the system of equations (3.147). Denote

$$\hat{P}_j(n_j) = \begin{cases} 1, & \text{if } n_j = 0 \\ \rho_j(1)\rho_j(2)\cdots\rho_j(n_j), & \text{if } n_j > 0 \end{cases} \tag{3.150}$$

$$G(M) = \sum_{\{(n_1,\ldots,n_K) \mid n_1+\cdots+n_K=M\}} \hat{P}_1(n_1)\cdots\hat{P}_K(n_K) \tag{3.151}$$

We have:

**Jackson's Theorem for Closed Networks.** Under the preceding assumptions, we have for all states $n = (n_1, \ldots, n_K)$ with $n_1 + \cdots + n_K = M$

$$P(n) = \frac{\hat{P}_1(n_1)\cdots\hat{P}_K(n_K)}{G(M)} \tag{3.152}$$

[Note that because all solutions of Eq. (3.147) are scalar multiples of each other, the expression (3.152) for the probabilities $P(n)$ is not affected by the choice of the solution as long as this solution is nonzero. Note also that $G(M)$ is a normalization constant that ensures that $P(n)$ is a probability distribution.]

*Proof:* The proof is similar to the proof of Jackson's Theorem for open networks. We consider state vectors $n$ and $n'$ of the form

$$n = (n_1, \ldots, n_i, \ldots, n_j, \ldots, n_K)$$

$$n' = (n_1, \ldots, n_{i-1}, n_i + 1, n_{i+1}, \ldots, n_{j-1}, n_j - 1, n_{j+1}, \ldots, n_K)$$

Let $q_{nn'}$ be the corresponding transition rate. Jackson's Theorem will be proved if the rates $q^*_{nn'}$ defined for all $n, n'$ by the equation

$$q^*_{nn'} = \frac{P(n')q_{n'n}}{P(n)} \tag{3.153}$$

satisfy, for all states $n$, the total rate equation

$$\sum_m q_{nm} = \sum_m q^*_{nm} \tag{3.154}$$

Indeed, let us assume for the purpose of the proof that the particular solutions $\overline{\lambda}_j$ are taken to be equal to the true arrival rates $\lambda_j(M)$, and for convenience let us denote both $\overline{\lambda}_j$ and $\lambda_j(M)$ as $\lambda_j$. Then we have [cf. Eqs. (3.124) and (3.125)]

$$q_{nn'} = \mu_j(n_j)P_{ji}, \qquad q^*_{nn'} = \frac{\mu_j(n_j)\lambda_i P_{ij}}{\lambda_j}$$

and the total rate equation (3.154) is written as

$$\sum_{\{(j,i)|n_j>0\}} \mu_j(n_j)P_{ji} = \sum_{\{(j,i)|n_j>0\}} \frac{\mu_j(n_j)\lambda_i P_{ij}}{\lambda_j}$$

We have

$$\begin{aligned}\sum_{\{(j,i)|n_j>0\}} \mu_j(n_j)P_{ji} &= \sum_{i=1}^{K} \sum_{\{j|n_j>0\}} \mu_j(n_j)P_{ji} = \sum_{\{j|n_j>0\}} \mu_j(n_j) \sum_{i=1}^{K} P_{ji} \\ &= \sum_{\{j|n_j>0\}} \mu_j(n_j)\end{aligned} \tag{3.155}$$

We also have

$$\begin{aligned}\sum_{\{(j,i)|n_j>0\}} \frac{\mu_j(n_j)\lambda_i P_{ij}}{\lambda_j} &= \sum_{i=1}^{K} \sum_{\{j|n_j>0\}} \frac{\mu_j(n_j)\lambda_i P_{ij}}{\lambda_j} = \sum_{\{j|n_j>0\}} \frac{\mu_j(n_j)}{\lambda_j} \sum_{i=1}^{K} \lambda_i P_{ij} \\ &= \sum_{\{j|n_j>0\}} \mu_j(n_j)\end{aligned} \tag{3.156}$$

From Eqs. (3.155) and (3.156) we see that the total rate equation (3.154) holds. **Q.E.D.**

**Example 3.21 Closed Computer System with Feedback Loop for I/O**

Consider a model of a computer CPU connected to an I/O device as shown in Fig. 3.37. This is a similar model to the one discussed in Example 3.19. The difference is that here

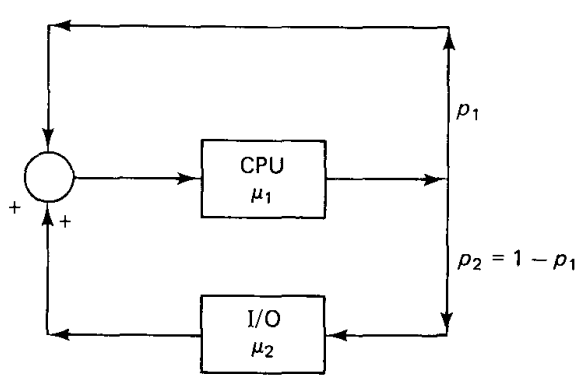



**Figure 3.37** Closed network model of a feedback system of a CPU and an I/O device.

we have a closed network with each job reentering the CPU directly (with probability $p_1$) or after using the I/O device (with probability $p_2 = 1 - p_1$). There are $M$ jobs in the system. We select

$$\overline{\lambda}_1 = \mu_1, \qquad \overline{\lambda}_2 = p_2\mu_1$$

as the particular solution of the system $\lambda_j = \sum_{i=1}^{2} \lambda_i P_{ij}$, $j = 1, 2$. With this choice we have

$$\rho_1 = 1, \qquad \rho_2 = \frac{p_2\mu_1}{\mu_2}$$

and the steady-state distribution of the system is given by [cf. Eqs. (3.149) to (3.151)]

$$P(M - n, n) = \frac{\rho_2^n}{G(M)}, \qquad n = 0, 1, \ldots, M$$

where the normalizing constant $G(M)$ is given by

$$G(M) = \sum_{n=0}^{M} \rho_2^n$$

The CPU utilization factor is given by

$$U(M) = 1 - P(0, M) = 1 - \frac{\rho_2^M}{G(M)} = \frac{G(M-1)}{G(M)}$$

and from Little's Theorem we obtain the arrival rate at the CPU as $\lambda_1(M) = U(M)\mu_1$. The expression above for the utilization factor $U(M)$ is a special case of a more general formula (see Problem 3.65).

**Example 3.22 Throughput of a Time-Sharing System**

Consider the time-sharing computer system with $N$ terminals discussed in Example 3.7 [cf. Fig. 3.38(a)]. We will make detailed statistical assumptions on the times spent by jobs at the terminals and the CPU. We will consequently be able to obtain a closed-form expression for the throughput of the system in place of the upper and lower bounds obtained in Section 3.2.

In particular, let us assume that the reflection time of a job at a terminal is exponentially distributed with mean $R$ and the processing time of a job at the CPU is exponentially distributed with mean $P$. All reflection and processing times are assumed independent. Then the terminal and CPU queues, viewed in isolation, can be modeled as an $M/M/\infty$ and an $M/M/1$ queue, respectively [see Fig. 3.38(b)]. Let $\overline{\lambda} = 1/P$ be the particular solution of the arrival rate equation for the system [cf. Eq. (3.147)]. With this choice we have

$$\rho_1 = \frac{R}{P}, \qquad \rho_2 = 1$$

The steady-state probability distribution is given by [cf. Eqs. (3.149) to (3.151)]

$$P(n, N - n) = \frac{(R/P)^n}{n!G(N)}$$

where the normalizing constant $G(N)$ is given by

$$G(N) = 1 + (R/P) + \frac{(R/P)^2}{2!} + \cdots + \frac{(R/P)^N}{N!}$$

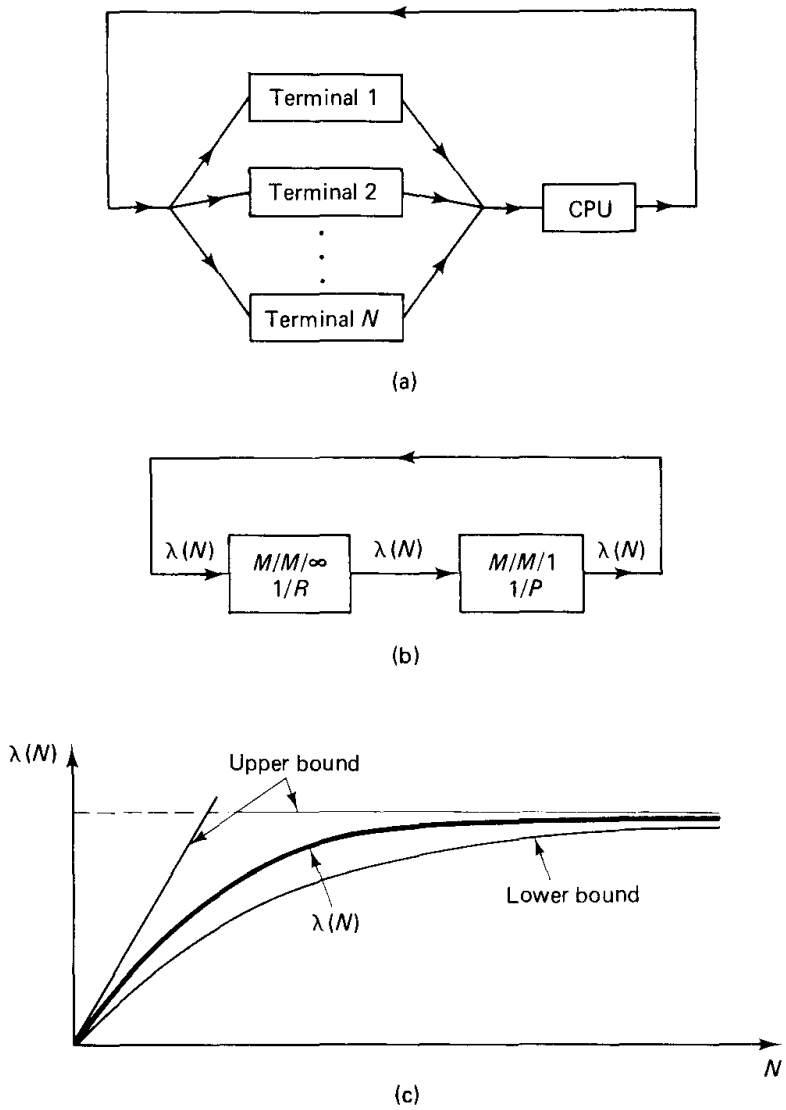



**Figure 3.38** (a) Closed network model of a time-sharing system consisting of $N$ terminals and a CPU. (b) Network of queues model of the system. There are at $N$ jobs in the system at all times. (c) Throughput $\lambda(N)$ as a function of the number of terminals compared with the upper and lower bounds

$$\frac{N}{R+NP} \le \lambda(N) \le \min\left\{\frac{1}{P}, \frac{N}{R+P}\right\}$$

derived in Example 3.7.

The CPU utilization factor is

$$U(N) = 1 - \frac{P(N,0)}{G(N)} = \frac{(R/P)^N}{N!G(N)} = \frac{G(N-1)}{G(N)}$$

and by Little's Theorem, it is also equal to $\lambda(N)P$, where $\lambda(N)$ is the system throughput. Therefore, we have $\lambda(N) = U(N)/P$, or

$$\lambda(N) = \frac{1}{P}\frac{G(N-1)}{G(N)}$$

This expression for $\lambda(N)$ is shown in Fig. 3.38(c) and is contrasted with the upper and lower bounds

$$\frac{N}{R+NP} \leq \lambda(N) \leq \min\left\{\frac{1}{P}, \frac{N}{R+P}\right\} \tag{3.157}$$

obtained in Section 3.2.

### 3.8.3 Computational Aspects—Mean Value Analysis

Given a closed queueing network with $M$ customers, one is typically interested in calculating

$$N_j(M) = \text{ Average number of customers in the } j^{\text{th}} \text{ queue}$$

$$T_j(M) = \text{ Average customer time spent per visit in the } j^{\text{th}} \text{ queue}$$

From these one can obtain the arrival rate at the $j^{\text{th}}$ queue given by Little's Theorem as

$$\lambda_j(M) = \frac{N_j(M)}{T_j(M)} \tag{3.158}$$

One possibility is to calculate first the normalizing constant $G(M)$ of Eq. (3.151) and then to use the steady-state distribution $P(n)$ of Eq. (3.152) to obtain all quantities of interest. Several different algorithms can be used for this computation, which is often nontrivial when $M$ is large. We will describe an alternative approach, known as *mean value analysis*, which calculates $N_j(M)$ and $T_j(M)$ directly. The normalizing constant $G(M)$ can then be obtained from these quantities and the arrival rates of Eq. (3.158). [See Problem 3.65 for the case where the service rates $\mu_j(m)$ do not depend on the number of customers $m$.]

Let us assume for simplicity that the service rate at the $j^{\text{th}}$ queue is $\mu_j$ and does not depend on the number of customers in the queue. The main idea in mean value analysis is to start with the known quantities

$$T_j(0) = N_j(0) = 0, \qquad j = 1, \ldots, K \tag{3.159}$$

(corresponding to an empty system) and then calculate $T_j(1)$ and $N_j(1)$ (corresponding to one customer in the system), then calculate $T_j(2)$ and $N_j(2)$, and so on until the desired quantities $T_j(M)$ and $N_j(M)$ are obtained. This calculation is based on the equation (to be justified shortly)

$$T_j(s) = \frac{1}{\mu_j}\big(1 + N_j(s-1)\big), \qquad j = 1, \ldots, K, \qquad s = 1, \ldots, M \tag{3.160}$$

which obtains $T_j(s)$ from $N_j(s-1)$ for all $j$. Then $N_j(s)$ is calculated from $T_j(s)$ for all $j$, using the equation (which is in effect Little's Theorem, as will be seen shortly)

$$N_j(s) = s\frac{\overline{\lambda}_j T_j(s)}{\sum_{i=1}^{K} \overline{\lambda}_i T_i(s)}, \qquad j = 1, \ldots, K, \qquad s = 1, \ldots, M \tag{3.161}$$

where $\overline{\lambda}_j$, $j = 1, \ldots, K$ is a positive solution of the system of equations $\lambda_j = \sum_{i=1}^{K} \lambda_i P_{ij}$, $j = 1, \ldots, K$ [cf. Eq. (3.147)].

We proceed to derive Eqs. (3.160) and (3.161). Since we have for all $j$, $\lambda_j(s) = \alpha(s)\overline{\lambda}_j$ for some scalar $\alpha(s) > 0$, Eq. (3.161) can be written as

$$N_j(s) = s\frac{\lambda_j(s)T_j(s)}{\sum_{i=1}^{K}\lambda_i(s)T_i(s)}$$

and becomes evident once we observe that we have $\lambda_i(s)T_i(s) = N_i(s)$ for all $i$ (by Little's Theorem) and $s = \sum_{i=1}^{K} N_i(s)$ [by the definition of $N_i(s)$].

To derive Eq. (3.160), we need an important result known as the *Arrival Theorem.* It states that the occupancy distribution found by a customer upon arrival at the $j^{\text{th}}$ queue is the same as the steady-state distribution of the $j^{\text{th}}$ queue in a closed network with the arriving customer removed. Thus, in an $s$-customer closed network, the average number of customers found upon arrival by a customer at the $j^{\text{th}}$ queue is equal to $N_j(s-1)$, the average number seen by a random observer in the $(s-1)$-customer closed network. This explains the form of Eq. (3.160).

An intuitive explanation of the Arrival Theorem is given in Problem 3.59. For an analytical justification, assume that the $s$-customer closed network is in steady-state at time $t$ and let $x(t)$ denote the state at that time. For each state $n = (n_1, \ldots, n_K)$ with $n_i > 0$, we want to calculate

$$\alpha_{ij}(n) = P\big\{x(t) = n \mid \text{a customer moved from queue } i \text{ to queue } j \text{ just after time } t\big\} \tag{3.162}$$

Let us denote by $M_{ij}(t)$ the event of a customer move from queue $i$ to queue $j$ just after time $t$, and let us denote by $M_i(t)$ the event of a customer move from queue $i$ just after time $t$. Then Eq. (3.162) can be written as

$$\begin{aligned}\alpha_{ij}(n) &= \frac{P\big\{x(t) = n, M_{ij}(t) \mid M_i(t)\big\}}{P\big\{M_{ij}(t) \mid M_i(t)\big\}}\\ &= \frac{P\big\{x(t) = n \mid M_i(t)\big\}P\big\{M_{ij}(t) \mid x(t) = n,\ M_i(t)\big\}}{P\big\{M_{ij}(t) \mid M_i(t)\big\}}\\ &= \frac{P(n)P_{ij}}{\sum_{\{n'=(n'_1,\ldots,n'_K)\mid n'_i>0\}} P(n')P_{ij}}\end{aligned}$$

and finally, using Eqs. (3.149) to (3.152) for the steady-state probabilities $P(n)$,

$$\alpha_{ij}(n) = \frac{\hat{P}_1(n_1)\cdots\hat{P}_K(n_K)}{\sum_{\{(n'_1,\ldots,n'_K)\mid n'_1+\cdots+n'_K=s,\ n'_i>0\}} \hat{P}_1(n'_1)\cdots\hat{P}_K(n'_K)} \tag{3.163}$$

The numerator and the denominator of this equation contain a common factor $\rho_i$ because $n_i > 0$ in the numerator and $n'_i > 0$ in each term of the denominator. By dividing with $\rho_i$, and by using the expression (3.150) for $\hat{P}_j(n_j)$, we obtain

$$\alpha_{ij}(n) = \frac{\hat{P}_1(n_1)\cdots\hat{P}_{i-1}(n_{i-1})\hat{P}_i(n_i-1)\hat{P}_{i+1}(n_{i+1})\cdots\hat{P}_K(n_K)}{\sum_{\{(n'_1,\ldots,n'_K)\mid n'_1+\cdots+n'_K=s-1\}} \hat{P}_1(n'_1)\cdots\hat{P}_K(n'_K)}$$

Therefore, $\alpha_{ij}(n)$ is equal to the steady-state probability of state $(n_1, \ldots, n_{i-1}, n_i - 1, n_{i+1}, \ldots, n_K)$ in the $(s-1)$-customer closed network, as stated by the Arrival Theorem.

We note that the Arrival Theorem holds also in some cases where there are multiple classes of customers and where the queues have multiple servers. Mean value analysis can also be used in these cases, with Eq. (3.160) replaced by the appropriate formula.

Finally, a number of approximate methods based on mean value analysis have been proposed. As an example, suppose that an approximate relation of the form

$$N_j(M-1) = f_j\big(N_j(M)\big)$$

is hypothesized; for large $M$, one reasonable possibility is

$$N_j(M-1) = \frac{M-1}{M} N_j(M)$$

Then Eqs. (3.160) and (3.161) yield the system of nonlinear equations

$$T_j(M) = \frac{1}{\mu_j}\Big(1 + f_j\big(N_j(M)\big)\Big), \qquad j = 1, \ldots, K$$

$$N_j(M) = M \frac{\overline{\lambda}_j T_j(M)}{\sum_{i=1}^{K} \overline{\lambda}_i T_i(M)}, \qquad j = 1, \ldots, K$$

which can be solved by iterative methods to yield approximate values for $T_j(M)$ and $N_j(M)$.

## SUMMARY

Queueing models provide qualitative insights on the performance of data networks, and quantitative predictions of average packet delay. An example of the former is the comparison of time-division and statistical multiplexing, while an example of the latter is the delay analysis of reservation systems.

To obtain tractable queueing models for data networks, it is frequently necessary to make simplifying assumptions. A prime example is the Kleinrock independence approximation discussed in Section 3.6. Delay predictions based on this approximation are adequate for many uses. A more accurate alternative is simulation, which, however, can be slow, expensive, and lacking in insight.

Little's Theorem is a simple but extremely useful result since it holds under very general conditions. To proceed beyond this theorem we assumed Poisson arrivals and independent interarrival and service times. This led to the $M/G/1$ system and its extensions in reservation and priority queueing systems. We analyzed a surprisingly large number of important delay models using simple graphical arguments. An alternative analysis was based on the use of Markov chain models and led to the derivation of the occupancy probability distribution of the $M/M/1$, $M/M/m$, and related systems.

Reversibility is an important notion that helps to prove and understand Jackson's Theorem and provides a taste of advanced queueing topics.

## NOTES, SOURCES, AND SUGGESTED READING

**Section 3.2.** Little's Theorem was formalized in [Lit61]. Rigorous proofs under various assumptions are given in [Sti72] and [Sti74]. Several applications in finding performance bounds of computer systems are described in [StA85].

**Section 3.3.** For a general background on the Poisson process, Markov chains, and related topics, see [Ros80], [Ros83], and [KaT75]. Standard texts on queueing theory include [Coo81], [GrH85], [HeS82], [Kle75], and [Wol89]. A reference for the fact that Poisson arrivals see a typical occupancy distribution (Section 3.3.2) is [Wol82a].

**Section 3.4.** Queueing systems that admit analysis via Markov chain theory include those where the service times have an Erlang distribution; see [Kle76], Chap. 4. For extensions to more general models and computational methods, see [Kei79], [Neu81], [Haj82], and [Twe82]. For methods to calculate the blocking probability in circuit switching networks (Example 3.14), see [Kau81], [Kel86], [LeG89], [RoT90], and [TsR90].

**Section 3.5.** The P-K formula is often derived by using $z$-transforms; see [Kle75]. This derivation is not very insightful, but gives the probability distribution of the system occupancy (not just the mean that we obtained via our much simpler analysis). For more on delay analysis of ARQ systems, see [AnP86] and [ToW79].

The results on polling and reservation systems are fairly recent; see [Coo70], [Eis79], [FeA85], [FuC85], [IEE86], and [Kue79]. The original references that are closest to our analysis are [Has72] for unlimited service systems, [NoT78] for limited service systems, and [Hum78] for nonsymmetric polling systems. Reference [Tak86] is a monograph devoted to polling. There are two main reservation and polling systems considered in the literature: the symmetric case, where all users have identical arrival and reservation interval statistics, and the nonsymmetric case, where these statistics are user dependent. The former case admits simple expressions for the mean waiting times while the latter does not. We have considered the partially symmetric case, where all users have identical arrival statistics but different reservation interval statistics. The fact that simple expressions hold for this case has not been known earlier, and in this respect, our formulas are original. Our treatment in terms of simple graphical arguments is also original. Approximate formulas for nonsymmetric polling systems are given in [BoM86] and [IbC89]. The result of Problem 3.35 on limited service systems with shared reservation and data intervals is new.

An extensive treatment of priority queueing systems is [Jai68]. A simpler, less comprehensive exposition is given in [Kle75].

The material on the $G/G/1$ queue is due to [Kin62]. For further material, see Chapter 11 of [Wol89], [Whi83a], [Whi83b], and the references quoted there.

**Section 3.6.** Delay analysis for data networks in terms of $M/M/1$ approximations was introduced in [Kle64]. References [Wol82b] and [PiW82] study via analysis and simulation the behavior of two queues in tandem when the service times of a customer at the two queues are dependent. The special issue [IEE86] provides a view of recent work on the subject.

**Section 3.7.** The notion of reversibility was used in Markov chain analysis by Kolmogorov [Kol36], and was explored in depth in [Kel79] and [Wal88].

**Section 3.8.** There is an extensive literature on product form solutions of queueing networks following Jackson's original paper [Jac57]. The survey [DiK85] lists 314 references. There are also several books on the subject: [Kel79], [BrB80], [GeP87], [Wal88], and [CoG89]. The heuristic explanation of Jackson's theorem is due to [Wal83].

## PROBLEMS

**3.1** Customers arrive at a fast-food restaurant at a rate of five per minute and wait to receive their order for an average of 5 minutes. Customers eat in the restaurant with probability 0.5 and carry out their order without eating with probability 0.5. A meal requires an average of 20 minutes. What is the average number of customers in the restaurant? (Answer: 75.)

**3.2** Two communication nodes 1 and 2 send files to another node 3. Files from 1 and 2 require on the average $R_1$ and $R_2$ time units for transmission, respectively. Node 3 processes a file of node $i$ ($i = 1, 2$) in an average of $P_i$ time units and then requests another file from either node 1 or node 2 (the rule of choice is left unspecified). If $\lambda_i$ is the throughput of node $i$ in files sent per unit time, what is the region of all feasible throughput pairs $(\lambda_1, \lambda_2)$ for this system?

**3.3** A machine shop consists of $N$ machines that occasionally fail and get repaired by one of the shop's $m$ repairpersons. A machine will fail after an average of $R$ time units following its previous repair and requires an average of $P$ time units to get repaired. Obtain upper and lower bounds (functions of $R$, $N$, $P$, and $m$) on the number of machine failures per unit time and on the average time between repairs of the same machine.

**3.4** The average time $T$ a car spends in a certain traffic system is related to the average number of cars $N$ in the system by a relation of the form $T = \alpha + \beta N^2$, where $\alpha > 0$, $\beta > 0$ are given scalars.

**(a)** What is the maximal car arrival rate $\lambda^*$ that the system can sustain?

**(b)** When the car arrival rate is less than $\lambda^*$, what is the average time a car spends in the system assuming that the system reaches a statistical steady state? Is there a unique answer? Try to argue against the validity of the statistical steady-state assumption.

**3.5** An absent-minded professor schedules two student appointments for the same time. The appointment durations are independent and exponentially distributed with mean 30 minutes. The first student arrives on time, but the second student arrives 5 minutes late. What is the expected time between the arrival of the first student and the departure of the second student? (Answer: 60.394 minutes.)

**3.6** A person enters a bank and finds all of the four clerks busy serving customers. There are no other customers in the bank, so the person will start service as soon as one of the customers in service leaves. Customers have independent, identical, exponential distribution of service time.

**(a)** What is the probability that the person will be the last to leave the bank assuming that no other customers arrive?

**(b)** If the average service time is 1 minute, what is the average time the person will spend in the bank?

**(c)** Will the answer in part (a) change if there are some additional customers waiting in a common queue and customers begin service in the order of their arrival?

**3.7** A communication line is divided in two identical channels each of which will serve a packet traffic stream where all packets have equal transmission time $T$ and equal interarrival time $R > T$. Consider, alternatively, statistical multiplexing of the two traffic streams by combining the two channels into a single channel with transmission time $T/2$ for each packet. Show that the average system time of a packet will be decreased from $T$ to something between $T/2$ and $3T/4$, while the variance of waiting time in queue will be increased from 0 to as much as $T^2/16$.

**3.8** Consider a packet stream whereby packets arrive according to a Poisson process with rate 10 packets/sec. If the interarrival time between any two packets is less than the transmission time of the first to arrive, the two packets are said to collide. (This notion will be made more meaningful in Chapter 4 when we discuss multiaccess schemes.) Find the probabilities that a packet does not collide with either its predecessor or its successor, and that a packet does not collide with another packet assuming:

**(a)** All packets have a transmission time of 20 msec. (Answer: Both probabilities are equal to 0.67.)

**(b)** Packets have independent, exponentially distributed transmission times with mean 20 msec. (This part requires the $M/M/\infty$ results.) (Answer: The probability of no collision with predecessor or successor is 0.694. The probability of no collision is 0.682.)

**3.9** A communication line capable of transmitting at a rate of 50 Kbits/sec will be used to accommodate 10 sessions each generating Poisson traffic at a rate 150 packets/min. Packet lengths are exponentially distributed with mean 1000 bits.

**(a)** For each session, find the average number of packets in queue, the average number in the system, and the average delay per packet when the line is allocated to the sessions by using:

**(1)** 10 equal-capacity time-division multiplexed channels. (Answer: $N_Q = 5$, $N = 10$, $T = 0.4$ sec.)

**(2)** Statistical multiplexing. (Answer: $N_Q = 0.5$, $N = 1$, $T = 0.04$ sec.)

**(b)** Repeat part (a) for the case where five of the sessions transmit at a rate of 250 packets/min while the other five transmit at a rate of 50 packets/min. (Answer: $N_Q = 21$, $N = 26$, $T = 1.038$ sec.)

**3.10** This problem deals with some of the basic properties of the Poisson process.

**(a)** Derive Eqs. (3.11) to (3.14).

**(b)** Show that if the arrivals in two disjoint time intervals are independent and Poisson distributed with parameters $\lambda\tau_1$, $\lambda\tau_2$, then the number of arrivals in the union of the intervals is Poisson distributed with parameter $\lambda(\tau_1 + \tau_2)$. (This shows in particular that the Poisson distribution of the number of arrivals in any interval [cf. Eq. (3.10)] is consistent with the independence requirement in the definition of the Poisson process.) *Hint*: Verify the correctness of the following calculation, where $N_1$ and $N_2$ are the number of arrivals in the two disjoint intervals:

$$P\{N_1 + N_2 = n\} = \sum_{k=0}^{n} P\{N_1 = k\}P\{N_2 = n-k\}$$
$$= e^{-\lambda(\tau_1+\tau_2)} \sum_{k=0}^{n} \frac{(\lambda\tau_1)^k(\lambda\tau_2)^{n-k}}{k!(n-k)!}$$
$$= e^{-\lambda(\tau_1+\tau_2)} \frac{(\lambda\tau_1 + \lambda\tau_2)^n}{n!}$$

**(c)** Show that if $k$ independent Poisson processes $A_1, \ldots, A_k$ are combined into a single process $A = A_1 + A_2 + \cdots + A_k$, then $A$ is Poisson with rate $\lambda$ equal to the sum of the rates $\lambda_1, \ldots \lambda_k$ of $A_1, \ldots, A_k$. Show also that the probability that the first arrival of the combined process comes from $A_1$ is $\lambda_1/\lambda$ independently of the time of arrival. *Hint*: For $k = 2$ write

$$P\{A_1(t+\tau) + A_2(t+\tau) - A_1(t) - A_2(t) = n\}$$
$$= \sum_{m=0}^{n} P\{A_1(t+\tau) - A_1(t) = m\}P\{A_2(t+\tau) - A_2(t) = n-m\}$$

and continue as in the hint for part (b). Also write for any $t$

$$P\{1 \text{ arrival from } A_1 \text{ prior to } t \mid 1 \text{ occurred}\}$$
$$= \frac{P\{1 \text{ arrival from } A_1 \text{ prior to } t,\ 0 \text{ from } A_2\}}{P\{1 \text{ occurred}\}}$$
$$= \frac{\lambda_1 t e^{-\lambda_1 t} e^{-\lambda_2 t}}{\lambda t e^{-\lambda t}} = \frac{\lambda_1}{\lambda}$$

**(d)** Suppose we know that in an interval $[t_1, t_2]$ only one arrival of a Poisson process has occurred. Show that, conditional on this knowledge, the time of this arrival is uniformly distributed in $[t_1, t_2]$. *Hint*: Verify that if $t$ is the time of arrival, we have for all $s \in [t_1, t_2]$,

$$P\{t < s \mid 1 \text{ arrival occurred in } [t_1, t_2]\}$$
$$= \frac{P\{1 \text{ arrival occurred in } [t_1, s),\ 0 \text{ arrivals occurred in } [s, t_2]\}}{P\{1 \text{ arrival occurred}\}}$$
$$= \frac{s - t_1}{t_2 - t_1}$$

**3.11** Packets arrive at a transmission facility according to a Poisson process with rate $\lambda$. Each packet is independently routed with probability $p$ to one of two transmission lines and with probability $(1-p)$ to the other.

**(a)** Show that the arrival processes at the two transmission lines are Poisson with rates $\lambda p$ and $\lambda(1-p)$, respectively. Furthermore, the two processes are independent. *Hint*: Let $N_1(t)$ and $N_2(t)$ be the number of arrivals in $[0, t]$ in lines 1 and 2, respectively. Verify the correctness of the following calculation:

$$
\begin{aligned}
&P\{N_1(t) = n, N_2(t) = m\} \\
&\qquad = P\{N_1(t) = n, N_2(t) = m \mid N(t) = n + m\} \frac{e^{-\lambda t}(\lambda t)^{n+m}}{(n+m)!} \\
&\qquad = \binom{n+m}{n} p^n (1-p)^m \frac{e^{-\lambda t}(\lambda t)^{n+m}}{(n+m)!} \\
&\qquad = \frac{e^{-\lambda t p}(\lambda t p)^n}{n!} \frac{e^{-\lambda t(1-p)}(\lambda t(1-p))^m}{m!}
\end{aligned}
$$

$$
P\{N_1(t) = n\} = \sum_{m=0}^{\infty} P\{N_1(t) = n, N_2(t) = m\} = \frac{e^{-\lambda t p}(\lambda t p)^n}{n!}
$$

**(b)** Use the result of part (a) to show that the probability distribution of the customer delay in a (first-come first-serve) $M/M/1$ queue with arrival rate $\lambda$ and service rate $\mu$ is exponential, that is, in steady-state we have

$$
P\{T_i \geq \tau\} = e^{-(\mu-\lambda)\tau}
$$

where $T_i$ is the delay of the $i^{\text{th}}$ customer. *Hint*: Consider a Poisson process $A$ with arrival rate $\mu$, which is split into two processes, $A_1$ and $A_2$, by randomization according to a probability $\rho = \lambda/\mu$; that is, each arrival of $A$ is an arrival of $A_1$ with probability $\rho$ and an arrival of $A_2$ with probability $(1-\rho)$, independently of other arrivals. Show that the interarrival times of $A_2$ have the same distribution as $T_i$.

**3.12** Let $\tau_1$ and $\tau_2$ be two exponentially distributed, independent random variables with means $1/\lambda_1$ and $1/\lambda_2$. Show that the random variable $\min\{\tau_1, \tau_2\}$ is exponentially distributed with mean $1/(\lambda_1 + \lambda_2)$ and that $P\{\tau_1 < \tau_2\} = \lambda_1/(\lambda_1 + \lambda_2)$. Use these facts to show that the $M/M/1$ queue can be described by a continuous-time Markov chain with transition rates $q_{n(n+1)} = \lambda$, $q_{(n+1)n} = \mu$, $n = 0, 1, \ldots$. (See Appendix A for material on continuous-time Markov chains.)

**3.13** Persons arrive at a taxi stand with room for $W$ taxis according to a Poisson process with rate $\lambda$. A person boards a taxi upon arrival if one is available and otherwise waits in a line. Taxis arrive at the stand according to a Poisson process with rate $\mu$. An arriving taxi that finds the stand full departs immediately; otherwise, it picks up a customer if at least one is waiting, or else joins the queue of waiting taxis.

**(a)** Use an $M/M/1$ queue formulation to obtain the steady-state distribution of the person's queue. What is the steady-state probability distribution of the taxi queue size when $W = 5$ and $\lambda$ and $\mu$ are equal to 1 and 2 per minute, respectively? (Answer: Let $p_i$ = Probability of $i$ taxis waiting. Then $p_0 = 1/32$, $p_1 = 1/32$, $p_2 = 1/16$, $p_3 = 1/8$, $p_4 = 1/4$, $p_5 = 1/2$.)

**(b)** In the leaky bucket flow control scheme to be discussed in Chapter 6, packets arrive at a network entry point and must wait in a queue to obtain a permit before entering the network. Assume that permits are generated by a Poisson process with given rate and can be stored up to a given maximum number; permits generated while the maximum number of permits is available are discarded. Assume also that packets arrive according to a Poisson process with given rate. Show how to obtain the occupancy distribution of the queue of packets waiting for permits. *Hint*: This is the same system as the one of part (a).

(c) Consider the flow control system of part (b) with the difference that permits are not generated according to a Poisson process but are instead generated periodically at a given rate. (This is a more realistic assumption.) Formulate the problem of finding the occupancy distribution of the packet queue as an $M/D/1$ problem.

type of

**3.14** A communication node $A$ receives Poisson packet traffic from two other nodes, 1 and 2, at rates $\lambda_1$ and $\lambda_2$, respectively, and transmits it, on a first-come first-serve basis, using a link with capacity $C$ bits/sec. The two input streams are assumed independent and their packet lengths are identically and exponentially distributed with mean $L$ bits. A packet from node 1 is always accepted by $A$. A packet from node 2 is accepted only if the number of packets in $A$ (in queue or under transmission) is less than a given number $K > 0$; otherwise, it is assumed lost.

(a) What is the range of values of $\lambda_1$ and $\lambda_2$ for which the expected number of packets in $A$ will stay bounded as time increases?

(b) For $\lambda_1$ and $\lambda_2$ in the range of part (a) find the steady-state probability of having $n$ packets in $A$ $(0 \leq n < \infty)$. Find the average time needed by a packet from source 1 to clear $A$ once it enters $A$, and the average number of packets in $A$ from source 1. Repeat for packets from source 2.

**3.15** Consider a system that is identical to $M/M/1$ except that when the system empties out, service does not begin again until $k$ customers are present in the system ($k$ is given). Once service begins it proceeds normally until the system becomes empty again. Find the steady-state probabilities of the number in the system, the average number in the system, and the average delay per customer. [Answer: The average number in the system is $N = \rho/(1-\rho) + (k-1)/2$.]

**3.16** *M/M/1-Like System with State-Dependent Arrival and Service Rate.* Consider a system which is the same as $M/M/1$ except that the rate $\lambda_n$ and service rate $\mu_n$ when there are $n$ customers in the system depend on $n$. Show that

$$p_{n+1} = (\rho_0 \cdots \rho_n) p_0$$

where $\rho_k = \lambda_k/\mu_{k+1}$ and

$$p_0 = \left[1 + \sum_{k=0}^{\infty} (\rho_0 \cdots \rho_k)\right]^{-1}$$

**3.17** *Discrete-Time Version of the M/M/1 System.* Consider a queueing system where interarrival and service times are integer valued, so customer arrivals and departures occur at integer times. Let $\lambda$ be the probability that an arrival occurs at any time $k$, and assume that at most one arrival can occur. Also let $\mu$ be the probability that a customer who was in service at time $k$ will complete service at time $k+1$. Find the occupancy distribution $p_n$ in terms of $\lambda$ and $\mu$.

**3.18** Empty taxis pass by a street corner at a Poisson rate of 2 per minute and pick up a passenger if one is waiting there. Passengers arrive at the street corner at a Poisson rate of 1 per minute and wait for a taxi only if there are fewer than four persons waiting; otherwise, they leave and never return. Find the average waiting time of a passenger who joins the queue. (Answer: 13/15 min.)

**3.19** A telephone company establishes a direct connection between two cities expecting Poisson traffic with rate 30 calls/min. The durations of calls are independent and exponentially distributed with mean 3 min. Interarrival times are independent of call durations. How many circuits should the company provide to ensure that an attempted call is blocked (because all

circuits are busy) with probability less than 0.01? It is assumed that blocked calls are lost (*i.e.*, a blocked call is not attempted again).

**3.20** A mail-order company receives calls at a Poisson rate of one per 2 min and the duration of the calls is exponentially distributed with mean 3 min. A caller who finds all telephone operators busy patiently waits until one becomes available. Write a computer program to determine how many operators the company should use so that the average waiting time of a customer is half a minute or less?

**3.21** Consider the $M/M/1/m$ system which is the same as $M/M/1$ except that there can be no more than $m$ customers in the system and customers arriving when the system is full are lost. Show that the steady-state occupancy probabilities are given by

$$p_n = \frac{\rho^n(1-\rho)}{1-\rho^{m+1}}, \qquad 0 \le n \le m$$

**3.22** An athletic facility has five tennis courts. Players arrive at the courts at a Poisson rate of one pair per 10 min and use a court for an exponentially distributed time with mean 40 min.

**(a)** Suppose that a pair of players arrives and finds all courts busy and $k$ other pairs waiting in queue. How long will they have to wait to get a court on the average?

**(b)** What is the average waiting time in queue for players who find all courts busy on arrival?

**3.23** Consider an $M/M/\infty$ queue with servers numbered $1, 2, \ldots$ There is an additional restriction that upon arrival a customer will choose the lowest-numbered server that is idle at the time. Find the fraction of time that each server is busy. Will the answer change if the number of servers is finite? *Hint*: Argue that in steady-state the probability that all of the first $m$ servers are busy is given by the Erlang B formula of the $M/M/m/m$ system. Find the total arrival rate to servers $(m+1)$ and higher, and from this, the arrival rate to each server.

**3.24** $M/M/1$ *Shared Service System.* Consider a system which is the same as $M/M/1$ except that whenever there are $n$ customers in the system they are all served simultaneously at an equal rate $1/n$ per unit time. Argue that the steady-state occupancy distribution is the same as for the $M/M/1$ system. *Note*: It can be shown that the steady-state occupancy distribution is the same as for $M/M/1$ even if the service time distribution is not exponential (*i.e.*, for an $M/G/1$ type of system) ([Ros83], p. 171).

**3.25** *Blocking Probability for Single-Cell Radio Systems ([BaA81] and [BaA82]).* A cellular radiotelephone system serves a given geographical area with $m$ radiotelephone channels connected to a single switching center. There are two types of calls: radio-to-radio calls, which occur with a Poisson rate $\lambda_1$ and require two radiochannels per call, and radio-to-nonradio calls, which occur with a Poisson rate $\lambda_2$ and require one radiochannel per call. The duration of all calls is exponentially distributed with mean $1/\mu$. Calls that cannot be accommodated by the system are blocked. Give formulas for the blocking probability of the two types of calls.

**3.26** A facility of $m$ identical machines is sharing a single repairperson. The time to repair a failed machine is exponentially distributed with mean $1/\lambda$. A machine, once operational, fails after a time that is exponentially distributed with mean $1/\mu$. All failure and repair times are independent. What is the steady-state proportion of time where there is no operational machine?

**3.27** $M/M/2$ *System with Heterogeneous Servers.* Derive the stationary distribution of an $M/M/2$ system where the two servers have different service rates. A customer that arrives when the system is empty is routed to the faster server.

**3.28** In Example 3.11, verify the formula $\sigma_f = (\lambda/\mu)^{1/2} s_\gamma$. *Hint*: Write

$$E\{f^2\} = E\left\{\left(\sum_{i=1}^{n} \gamma_i\right)^2\right\} = E\left\{E\left\{\left(\sum_{i=1}^{n} \gamma_i\right)^2 \mid n\right\}\right\},$$

and use the fact that $n$ is Poisson distributed.

**3.29** Customers arrive at a grocery store's checkout counter according to a Poisson process with rate 1 per minute. Each customer carries a number of items that is uniformly distibuted between 1 and 40. The store has two checkout counters, each capable of processing items at a rate of 15 per minute. To reduce the customer waiting time in queue, the store manager considers dedicating one of the two counters to customers with $x$ items or less and dedicating the other counter to customers with more than $x$ items. Write a small computer program to find the value of $x$ that minimizes the average customer waiting time.

**3.30** In the $M/G/1$ system, show that

$$P\{\text{the system is empty}\} = 1 - \lambda\overline{X}$$

$$\text{Average length of time between busy periods} = \frac{1}{\lambda}$$

$$\text{Average length of busy period} = \frac{\overline{X}}{1 - \lambda\overline{X}}$$

$$\text{Average number of customers served in a busy period} = \frac{1}{1 - \lambda\overline{X}}$$

**3.31** Consider the following argument in the $M/G/1$ system: When a customer arrives, the probability that another customer is being served is $\lambda\overline{X}$. Since the served customer has mean service time $\overline{X}$, the average time to complete the service is $\overline{X}/2$. Therefore, the mean residual service time is $\lambda\overline{X}^2/2$. What is wrong with this argument?

**3.32** *$M/G/1$ System with Arbitrary Order of Service.* Consider the $M/G/1$ system with the difference that customers are not served in the order they arrive. Instead, upon completion of a customer's service, one of the waiting customers in queue is chosen according to some rule, and is served next. Show that the P-K formula for the average waiting time in queue $W$ remains valid provided that the relative order of arrival of the customer chosen is independent of the service times of the customers waiting in queue. *Hint*: Argue that the independence hypothesis above implies that at any time $t$, the number $N_Q(t)$ of customers waiting in queue is independent of the service times of these customers. Show that this in turn implies that $U = R + \rho W$, where $R$ is the mean residual time and $U$ is the average steady-state unfinished work in the system (total remaining service time of the customers in the system). Argue that $U$ and $R$ are independent of the order of customer service.

**3.33** Show that Eq. (3.59) for the average delay of time-division multiplexing on a slot basis can be obtained as a special case of the results for the limited service reservation system. *Hint*: Consider the gated system with zero packet length.

**3.34** Consider the limited service reservation system. Show that for both the gated and the partially gated versions:

**(a)** The steady-state probability of arrival of a packet during a reservation interval is $1 - \rho$.

**(b)** The steady-state probability of a reservation interval being followed by an empty data interval is $(1 - \rho - \lambda\overline{V})/(1 - \rho)$. *Hint:* If $p$ is the required probability, argue that the ratio of the times used for data intervals and for reservation intervals is $(1 - p)\overline{X}/\overline{V}$.

**3.35** *Limited Service Reservation System with Shared Reservation and Data Intervals.* Consider the gated version of the limited service reservation system with the difference that the $m$ users share reservation and data intervals, (*i.e.*, all users make reservations in the same interval and transmit at most one packet each in the subsequent data interval). Show that

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho-\lambda\overline{V}/m)} + \frac{(1-\rho)\overline{V^2}}{2(1-\rho-\lambda\overline{V}/m)\overline{V}} + \frac{(1-\rho\alpha-\lambda\overline{V}/m)\overline{V}}{1-\rho-\lambda\overline{V}/m}$$

where $\overline{V}$ and $\overline{V^2}$ are the first two moments of the reservation interval, and $\alpha$ satisfies

$$\frac{\overline{K}+(\hat{K}-1)(2\overline{K}-\hat{K})}{2m\overline{K}} - \frac{1}{2m} \le \alpha \le \frac{1}{2} - \frac{1}{2m}$$

where

$$\overline{K} = \frac{\lambda\overline{V}}{1-\rho}$$

is the average number of packets per data interval, and $\hat{K}$ is the smallest integer which is larger than $\overline{K}$. Verify that the formula for $W$ becomes exact as $\rho \to 0$ (light load) and as $\rho \to 1 - \lambda\overline{V}/m$ (heavy load). *Hint:* Verify that

$$W = R + \lambda W + \left(1 + \frac{\lambda W}{m} - S\right)\overline{V}$$

where $S = \lim_{i\to\infty} E\{S_i\}$ and $S_i$ is the number (0 or 1) of packets of the owner of packet $i$ that will start transmission between the time of arrival of packet $i$ and the end of the cycle in which packet $i$ arrives. Try to obtain bounds for $S$ by considering separately the cases where packet $i$ arrives in a reservation and in a data interval.

**3.36** Repeat part (a) of Problem 3.9 for the case where packet lengths are not exponentially distributed, but 10% of the packets are 100 bits long and the rest are 1500 bits long. Repeat the problem for the case where the short packets are given nonpreemptive priority over the long packets. (Answer: $N_Q = 0.791$, $N = 1.47$, $T = 0.588$ sec.)

**3.37** Persons arrive at a Xerox machine according to a Poisson process with rate one per minute. The number of copies to be made by each person is uniformly distributed between 1 and 10. Each copy requires 3 sec. Find the average waiting time in queue when:

**(a)** Each person uses the machine on a first-come first-serve basis. (Answer: $W = 3.98$.)

**(b)** Persons with no more than 2 copies to make are given nonpreemptive priority over other persons.

**3.38** *Priority Systems with Multiple Servers.* Consider the priority systems of Section 3.5.3 assuming that there are $m$ servers and that all priority classes have exponentially distributed service times with common mean $1/\mu$.

**(a)** Consider the nonpreemptive system. Show that Eq. (3.79) yields the average queueing times with the mean residual time $R$ given by

$$R = \frac{P_Q}{m\mu}$$

where $P_Q$ is the steady-state probability of queueing given by the Erlang C formula of Eq. (3.36). [Here $\rho_i = \lambda_i/(m\mu)$ and $\rho = \sum_{i=1}^{n} \rho_i$.]

**(b)** Consider the preemptive resume system. Argue that $W_{(k)}$, defined as the average time in queue averaged over the first $k$ priority classes, is the same as for an $M/M/m$ system with arrival rate $\lambda_1 + \cdots + \lambda_k$ and mean service time $1/\mu$. Use Little's Theorem to

show that the average time in queue of a $k^{\text{th}}$ priority class customer can be obtained recursively from

$$W_1 = W_{(1)}$$

$$W_k = \frac{1}{\lambda_k}\left[W_{(k)}\sum_{i=1}^{k}\lambda_i - W_{(k-1)}\sum_{i=1}^{k-1}\lambda_i\right], \qquad k = 2, 3, \ldots, n$$

**3.39** Consider the nonpreemptive priority queueing system of Section 3.5.3 for the case where the available capacity is sufficient to handle the highest-priority traffic but cannot handle the traffic of all priorities, that is,

$$\rho_1 < 1 < \rho_1 + \rho_2 + \cdots + \rho_n$$

Find the average delay per customer of each priority class. *Hint*: Determine the departure rate of the highest-priority class that will experience infinite average delay and the mean residual service time.

**3.40** *Optimization of Class Ordering in a Nonpreemptive System.* Consider an $n$-class, nonpreemptive priority system:

**(a)** Show that the sum $\sum_{k=1}^{n}\rho_k W_k$ is independent of the priority order of classes, and in fact

$$\sum_{k=1}^{n}\rho_k W_k = \frac{R\rho}{1-\rho}$$

where $\rho = \rho_1 + \rho_2 + \cdots + \rho_n$. (This is known as the $M/G/1$ conservation law [Kle64].) *Hint:* Use Eq. (3.79). Alternatively, argue that $U = R + \sum_{k=1}^{n}\rho_k W_k$, where $U$ is the average steady-state unfinished work in the system (total remaining service time of customers in the system), and $U$ and $R$ are independent of the priority order of the classes.

**(b)** Suppose that there is a cost $c_k$ per unit time for each class $k$ customer that waits in queue. Show that cost is minimized when classes are ordered so that

$$\frac{\overline{X_1}}{c_1} \le \frac{\overline{X_2}}{c_2} \le \cdots \le \frac{\overline{X_n}}{c_n}$$

*Hint*: Express the cost as $\sum_{k=1}^{n}(c_k/\overline{X}_k)(\rho_k W_k)$ and use part (a). Also use the fact that interchanging the order of any two adjacent classes leaves the waiting time of all other classes unchanged.

**3.41** *Little's Theorem for Arbitrary Order of Service; Analytical Proof [Sti74].* Consider the analysis of Little's Theorem in Section 3.2 and the notation introduced there. We allow the possibility that the initial number in the system is positive [*i.e.*, $N(0) > 0$]. Assume that the time-average arrival and departure rates exist and are equal:

$$\lambda = \lim_{t\to\infty}\frac{\alpha(t)}{t} = \lim_{t\to\infty}\frac{\beta(t)}{t}$$

and that the following limit defining the time-average system time exists:

$$T = \lim_{k\to\infty}\frac{1}{N(0)+\alpha(t)}\left(\sum_{i\in D(t)} T_i + \sum_{i\in\overline{D}(t)}(t - t_i)\right)$$

where $D(t)$ is the set of customers departed by time $t$ and $\overline{D}(t)$ is the set of customers that are in the system at time $t$. (For all customers that are initially in the system, the time $T_i$ is counted starting at time 0.) Show that regardless of the order in which customers are served, Little's Theorem ($N = \lambda T$) holds with

$$N = \lim_{t \to \infty} \frac{1}{t} \int_0^t N(\tau)\, d\tau$$

Show also that

$$T = \lim_{k \to \infty} \frac{1}{k} \sum_{i=1}^{k} T_i$$

*Hint:* Take $t \to \infty$ below:

$$\frac{1}{t} \sum_{i \in D(t)} T_i \le \frac{1}{t} \int_0^t N(\tau)\, d\tau \le \frac{1}{t} \sum_{i \in D(t) \cup \overline{D}(t)} T_i$$

**3.42** *A Generalization of Little's Theorem.* Consider an arrival–departure system with arrival rate $\lambda$, where entering customers are forced to pay money to the system according to some rule.

**(a)** Argue that the following identity holds:

$$\text{Average rate at which the system earns } = \lambda \cdot (\text{Average amount a customer pays})$$

**(b)** Show that Little's Theorem is a special case.

**(c)** Consider the $M/G/1$ system and the following cost rule: Each customer pays at a rate of $y$ per unit time when its remaining service time is $y$, whether in queue or in service. Show that the formula in (a) can be written as

$$W = \lambda \left( \overline{X} W + \frac{\overline{X^2}}{2} \right)$$

which is the Pollaczek–Khinchin formula.

**3.43** *$M/G/1$ Queue with Random-Sized Batch Arrivals.* Consider the $M/G/1$ system with the difference that customers are arriving in batches according to a Poisson process with rate $\lambda$. Each batch has $n$ customers, where $n$ has a given distribution and is independent of customer service times. Adapt the proof of Section 3.5 to show that the waiting time in queue is given by

$$W = \frac{\lambda \overline{n} \overline{X^2}}{2(1-\rho)} + \frac{\overline{X}(\overline{n^2} - \overline{n})}{2\overline{n}(1-\rho)}$$

*Hint*: Use the equation $W = R + \rho W + W_B$, where $W_B$ is the average waiting time of a customer for other customers who arrived in the same batch.

**3.44** *$M/G/1$ Queue with Overhead for Each Busy Period.* Consider the $M/G/1$ queue with the difference that the service of the first customer in each busy period requires an increment $\Delta$ over the ordinary service time of the customer. We assume that $\Delta$ has a given distribution and is independent of all other random variables in the model. Let $\rho = \lambda \overline{X}$ be the utilization factor. Show that

**(a)** $p_0 = P\{\text{the system is empty}\} = (1-\rho)/(1+\lambda\overline{\Delta})$.

**(b)** Average length of busy period $= (\overline{X} + \overline{\Delta})/(1-\rho)$.

**(c)** The average waiting time in queue is

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho)} + \frac{\lambda[\overline{(X+\Delta)^2} - \overline{X^2}]}{2(1+\lambda\overline{\Delta})}$$

**(d)** Parts (a), (b), and (c) also hold in the case where $\Delta$ may depend on the interarrival and service time of the first customer in the corresponding busy period.

**3.45** Consider a system that is identical to $M/G/1$ except that when the system empties out, service does not begin again until $k$ customers are present in the system ($k$ is given). Once service begins, it proceeds normally until the system becomes empty again. Show that:

**(a)** In steady-state:

$$P\{\text{system empty}\} = \frac{1-\rho}{k}$$

$$P\{\text{system nonempty and waiting}\} = \frac{(k-1)(1-\rho)}{k}$$

$$P\{\text{system nonempty and serving}\} = \rho$$

**(b)** The average length of a busy period is

$$\frac{\rho + k - 1}{\lambda(1-\rho)}$$

Verify that this average length is equal to the average time between arrival and start of service of the first customer in a busy period, plus $k$ times the average length of a busy period for the corresponding $M/G/1$ system ($k = 1$).

**(c)** Suppose that we divide a busy period into a busy/waiting portion and a busy/serving portion. Show that the average number in the system during a busy/waiting portion is $k/2$ and the average number in the system during a busy/serving portion is

$$\frac{N_{M/G/1}}{\rho} + \frac{k-1}{2}$$

where $N_{M/G/1}$ is the average number in the system for the corresponding $M/G/1$ system ($k = 1$). *Hint*: Relate a busy/serving portion of a busy period with $k$ independent busy periods of the corresponding $M/G/1$ system where $k = 1$.

**(d)** The average number in the system is

$$N_{M/G/1} + \frac{k-1}{2}$$

**3.46** *Single-Vacation $M/G/1$ System.* Consider the $M/G/1$ system with the difference that each busy period is followed by a single vacation interval. Once this vacation is over, an arriving customer to an empty system starts service immediately. Assume that vacation intervals are independent, identically distributed, and independent of the customer interarrival and service times. Prove that the average waiting time in queue is

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho)} + \frac{\overline{V^2}}{2I}$$

where $I$ is the average length of an idle period, and show how to calculate $I$.

**3.47** *The $M/G/\infty$ System.* Consider a queueing system with Poisson arrivals at rate $\lambda$. There are an infinite number of servers, so that each arrival starts service at an idle server immediately on arrival. Each server has a general service time distribution and $F_X(x) = P\{X \le x\}$

denotes the probability that a service starting at any given time $\tau$ is completed by time $\tau + x$ [$F_X(x) = 0$ for $x \leq 0$]. The servers have independent and identical service time distributions.

**(a)** For $x$ and $\delta$ ($0 < \delta < x$) very small, find the probability that there was an arrival in the interval $[\tau - x, \tau - x + \delta]$ *and* that this arrival is still being served at time $\tau$.

**(b)** Show that the mean service time for any arrival is given by

$$\overline{X} = \int_0^\infty [1 - F_X(x)]\, dx$$

*Hint*: Use a graphical argument.

**(c)** Use parts (a) and (b) to verify that the number in the system is Poisson distributed with mean $\lambda\overline{X}$.

**3.48** *An Improved Bound for the $G/G/1$ Queue.*

**(a)** Let $r$ be a nonnegative random variable and let $x$ be a nonnegative scalar. Show that

$$\frac{\overline{\left(\max\{0, r - x\}\right)^2}}{\overline{\left(\max\{0, r - x\}\right)}^2} \geq \frac{\overline{r^2}}{(\overline{r})^2}$$

where overbar denotes expected value. *Hint*: Prove that the left-hand expression is monotonically nondecreasing as a function of $x$.

**(b)** Using the notation of Section 3.5.4, show that

$$\sigma_I^2 \geq (1 - \rho)^2 \sigma_a^2$$

and that

$$W \leq \frac{\lambda(\sigma_a^2 + \sigma_b^2)}{2(1 - \rho)} - \frac{\lambda(1 - \rho)\sigma_a^2}{2}$$

*Hint*: Use part (a) with $r$ being the customer interarrival time and $x$ equal to the time in the system [cf. Eq. (3.93)].

**3.49** *Last-Come First-Serve $M/G/1$ System.* Consider an $M/G/1$ system with the difference that upon arrival at the queue, a customer goes immediately into service, replacing the customer who is in service at the time (if any) on a preemptive-resume basis. When a customer completes service, the customer most recently preempted resumes service. Show that:

**(a)** The expected length of a busy period, denoted $E\{B\}$, is the same as in the ordinary $M/G/1$ queue.

**(b)** Show that the expected time in the system of a customer is equal to $E\{B\}$. *Hint*: Argue that a customer who starts a busy period stays in the system for the entire duration of the busy period.

**(c)** Let $C$ be the average time in the system of a customer requiring one unit of service time. Argue that the average time in the system of a customer requiring $X$ units of service time is $XC$. *Hint*: Argue that a customer requiring two units of service time is "equivalent" to two customers with one unit service time each, and with the second customer arriving at the time that the first departs.

**(d)** Show that

$$C = \frac{E\{B\}}{E\{X\}} = \frac{1}{1 - \rho}$$

**3.50** *Truncation of Queues.* This problem illustrates one way to use simple queues to obtain results about more complicated queues.

**(a)** Consider a continuous-time Markov chain with state space $S$, stationary distribution $\{p_j\}$, and transition rates $q_{ij}$. Suppose that we have a truncated version of this chain, that is, a new chain with space $\overline{S}$, which is a subset of $S$ and has the same transition rates $q_{ij}$ between states $i$ and $j$ of $\overline{S}$. Assume that for all $j \in \overline{S}$, we have

$$p_j \sum_{i \notin \overline{S}} q_{ji} = \sum_{i \notin \overline{S}} p_i q_{ij}$$

Show that if the truncated chain is irreducible, then its stationary distribution $\{\overline{p}_j\}$ satisfies $\overline{p}_j = p_j / \sum_{i \in \overline{S}} p_i$ for all $j \in \overline{S}$. (Note that $\overline{p}_j$ is the conditional probability for the state of the original chain to be $j$ conditioned on the fact that it lies within $\overline{S}$.)

**(b)** Show that the condition of part (a) on the stationary distribution $\{p_j\}$ and the transition rates $\{q_{ij}\}$ is satisfied if the original chain is time reversible, and that in this case, the truncated chain is also time reversible.

**(c)** Consider two queues with independent Poisson arrivals and independent exponentially distributed service times. The arrival and service rates are denoted $\lambda_i, \mu_i$, for $i = 1, 2$, respectively. The two queues share a waiting room with finite capacity $B$ (including customers in service). Arriving customers that find the waiting room full are lost. Use part (b) to show that the system is reversible and that for $m + n \leq B$, the steady-state probabilities are

$$P\{m \text{ in queue } 1,\ n \text{ in queue } 2\} = \frac{\rho_1^m \rho_2^n}{G}$$

where $\rho_i = \lambda_i / \mu_i$, $i = 1, 2$, and $G$ is a normalizing constant.

**3.51** *Decomposition/Aggregation of Reversible Chains.* Consider a time reversible continuous-time Markov chain in equilibrium, with state space $S$, transition rates $q_{ij}$, and stationary probabilities $p_j$. Let $S = \cup_{k=1}^{K} S_k$ be a partition of $S$ in mutually disjoint sets, and denote for all $k$ and $j \in S_k$:

$u_k$ = Probability of the state being in $S_k$ (*i.e.*, $u_k = \sum_{j \in S_k} p_j$)

$\pi_j$ = Probability of the state being equal to $j$ conditioned on the fact that the state belongs to $S_k$ (*i.e.*, $\pi_j = P\{X_n = j \mid X_n \in S_k\} = p_j / u_k$)

Assume that all states in $S_k$ communicate with all other states in $S_k$.

**(a)** Show that $\{\pi_j \mid j \in S_k\}$ is the stationary distribution of the truncated chain with state space $S_k$ (cf. Problem 3.50).

**(b)** Show that $\{u_k \mid k = 1, \ldots, K\}$ is the stationary distribution of the so-called *aggregate chain*, which is the Markov chain with states $k = 1, \ldots, K$ and transition rates

$$\tilde{q}_{km} = \sum_{j \in S_k,\ i \in S_m} \pi_j q_{ji}, \qquad k, m = 1, \ldots, K$$

Show also that the aggregate chain is reversible. (Note that the aggregate chain corresponds to a fictitious process; the actual process, corresponding to transitions between sets of states, need not be Markov.)

**(c)** Outline a divide-and-conquer solution method that first solves for the distributions of the truncated chains and then solves for the distribution of the aggregate chain. Apply this method to Examples 3.12 and 3.13.

(d) Suppose that the truncated chains are reversible but the original chain is not. Show that the results of parts (a) and (b) hold except that the aggregate chain need not be reversible.

**3.52** *An Extension of Burke's Theorem.* Consider an $M/M/1$ system in steady state where customers are served in the order that they arrive. Show that given that a customer departs at time $t$, the arrival time of that customer is independent of the departure process prior to $t$. *Hint*: Consider a customer arriving at time $t_1$ and departing at time $t_2$. In reversed system terms, the arrival process is independent Poisson, so the arrival process to the left of $t_2$ is independent of the times spent in the system of customers that arrived at or to the right of $t_2$.

**3.53** Consider the model of two queues in tandem of Section 3.7 and assume that customers are served at each queue in the order they arrive.

(a) Show that the times (including service) spent by a customer in queue 1 and in queue 2 are mutually independent, and independent of the departure process from queue 2 prior to the customer's departure from the system. *Hint:* By Burke's Theorem, the time spent by a customer in queue 1 is independent of the sequence of arrival times at queue 2 prior to the customer's arrival at queue 2. These arrival times (together with the corresponding independent service times) determine the time the customer spends at queue 2 as well as the departure process from queue 2 prior to the customer's departure from the system.

(b) Argue by example that the times a customer spends waiting *before entering service* at the two queues are *not* independent.

**3.54** Use reversibility to characterize the departure process of the $M/M/1/m$ queue.

**3.55** Consider the feedback model of a CPU and I/O device of Example 3.19 with the difference that the CPU consists of $m$ identical parallel processors. The service time of a job at each parallel processor is exponentially distributed with mean $1/\mu_1$. Derive the stationary distribution of the system.

**3.56** Consider the discrete-time approximation to the $M/M/1$ queue of Fig. 3.6. Let $X_n$ be the state of the system at time $n\Delta$ and let $D_n$ be a random variable taking on the value 1 if a departure occurs between $n\Delta$ and $(n+1)\Delta$, and the value 0 if no departure occurs. Assume that the system is in steady-state at time $n\Delta$. Answer the following without using reversibility.

(a) Find $P\{X_n = i, D_n = j\}$ for $i \geq 0$, $j = 0, 1$.

(b) Find $P\{D_n = 1\}$.

(c) Find $P\{X_n = i, D_n = 1\}$ for $i \geq 0$.

(d) Find $P\{X_{n+1} = i, D_n = 1\}$ and show that $X_{n+1}$ is statistically independent of $D_n$. *Hint*: Use part (c); also show that $P\{X_{n+1} = i\} = P\{X_{n+1} = i \mid D_n = 1\}$ for all $i \geq 0$ is sufficient to show independence.

(e) Find $P\{X_{n+k} = i, D_{n+1} = j \mid D_n\}$ and show that the pair of variables $(X_{n+1}, D_{n+1})$ is statistically independent of $D_n$.

(f) For each $k > 1$, find $P\{X_{n+k} = i, D_{n+k} = j \mid D_{n+k-1}, D_{n+k-2}, \ldots, D_n\}$ and show that the pair $(X_{n+k}, D_{n+k})$ is statistically independent of $(D_{n+k-1}, D_{n+k-2}, \ldots, D_n)$. *Hint*: Use induction on $k$.

(g) Deduce a discrete-time analog to Burke's Theorem.

**3.57** Consider the network in Fig. 3.39. There are four sessions: ACE, ADE, BCEF, and BDEF sending Poisson traffic at rates 100, 200, 500, and 600 packets/min, respectively. Packet lengths are exponentially distributed with mean 1000 bits. All transmission lines have capac-

ity 50 kbits/sec, and there is a propagation delay of 2 msec on each line. Using the Kleinrock independence approximation, find the average number of packets in the system, the average delay per packet (regardless of session), and the average delay per packet of each session.

**3.58** *Jackson Networks with a Limit on the Total Number of Customers.* Consider an open Jackson network as described in the beginning of Section 3.8, with the difference that all customers who arrive when there are a total of $M$ customers in the network are blocked from entering and are lost for the system. Derive the stationary distribution. *Hint*: Convert the system into a closed network with $M$ customers by introducing an additional queue $K+1$ with service rate equal to $\sum_{j=1}^{K} r_j$. A customer exiting queue $i \in \{1, \ldots, K\}$ enters queue $K+1$ with probability $1 - \sum_j P_{ij}$, and a customer exiting queue $K+1$ enters queue $i \in \{1, \ldots, K\}$ with probability $r_i / \sum_{j=1}^{K} r_j$.

**3.59** Justify the Arrival Theorem for closed networks by inserting a very fast $M/M/1$ queue between every pair of queues. Argue that conditioning on a customer moving from one queue to another is essentially equivalent to conditioning on a single customer being in the fast $M/M/1$ queue that lies between the two queues.

**3.60** Consider a closed Jackson network where the service time at each queue is independent of the number of customers at the queue. Suppose that for a given number of customers, the utilization factor of one of the queues, say queue 1, is strictly larger than the utilization factors of the other queues. Show that as the number of customers increases, the proportion of time that a customer spends in queue 1 approaches unity.

**3.61** Consider a model of a computer CPU connected to $m$ I/O devices as shown in Fig. 3.40. Jobs enter the system according to a Poisson process with rate $\lambda$, use the CPU and with

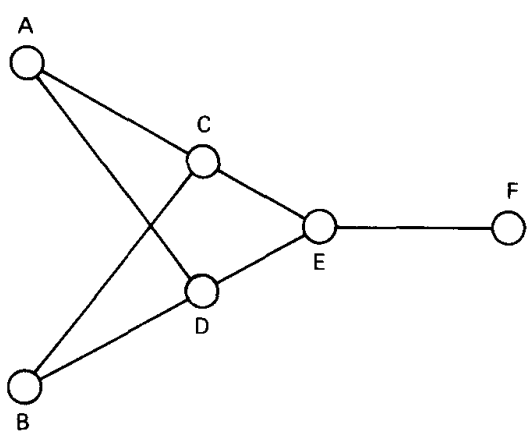



**Figure 3.39** Network of transmission lines for Problem 3.57.

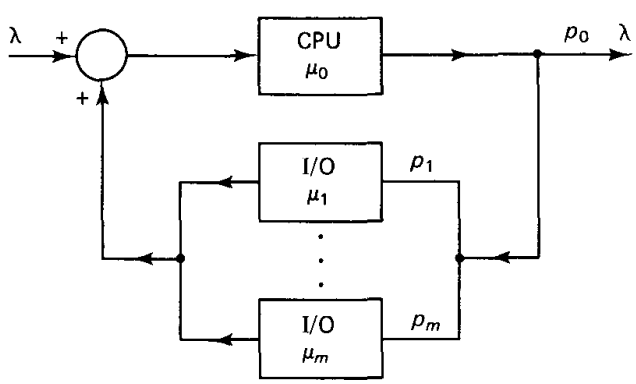



**Figure 3.40** Model of a computer CPU connected to $m$ I/O devices for Problem 3.61.

probability $p_i$, $i = 1, \ldots, m$, are routed to the $i^{\text{th}}$ I/O device, while with probability $p_0$ they exit the system. The service time of a job at the CPU (or the $i^{\text{th}}$ I/O device) is exponentially distributed with mean $1/\mu_0$ (or $1/\mu_i$, respectively). We assume that all job service times at all queues are independent (including the times of successive visits to the CPU and I/O devices of the same job). Find the occupancy distribution of the system and construct an "equivalent" system with $m+1$ queues in tandem that has the same occupancy distribution.

**3.62** Consider a closed version of the queueing system of Problem 3.61, shown in Fig. 3.41. There are $M$ jobs in the system at all times. A job uses the CPU and with probability $p_i$, $i = 1, \ldots, m$, is routed to the $i^{\text{th}}$ I/O device. The service time of a job at the CPU (or the $i^{\text{th}}$ I/O device) is exponentially distributed with mean $1/\mu_0$ (or $1/\mu_i$, respectively). We assume that all job service times at all queues are independent (including the times of successive visits to the CPU and I/O devices of the same job). Find the arrival rate of jobs at the CPU and the occupancy distribution of the system.

**3.63** *Bounds on the Throughput of a Closed Queueing Network.* Packets enter the network of transmission lines shown in Fig. 3.42 at point $A$ and exit at point $B$. A packet is first transmitted on one of the lines $L_1, \ldots, L_K$, where it requires on the average a transmission time $\overline{X}$, and is then transmitted in line $L_{K+1}$, where it requires on the average a transmission time $\overline{Y}$. To effect flow control, a maximum of $N \geq K$ packets are admitted into the system.

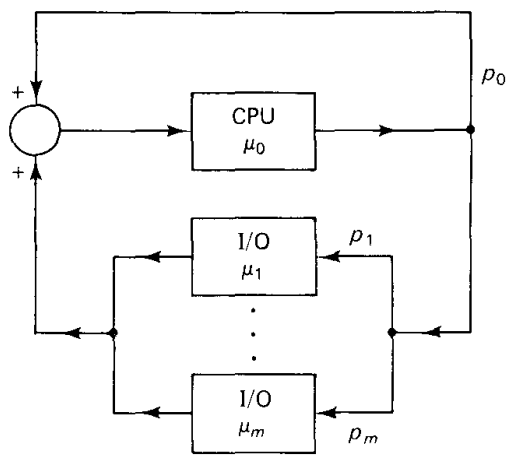



**Figure 3.41** Closed queueing system for Problem 3.62.

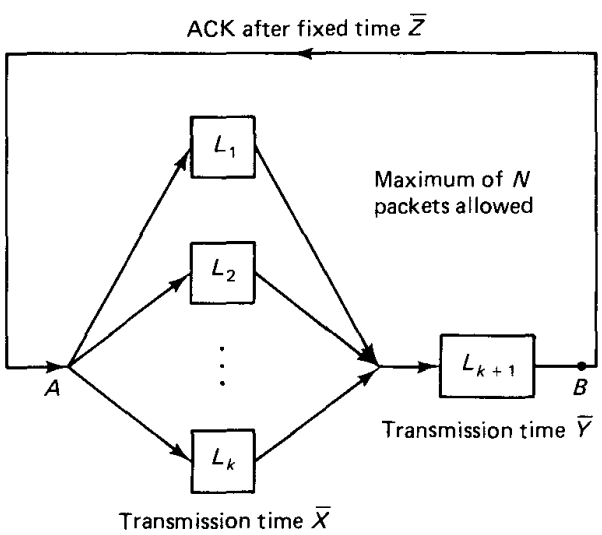



**Figure 3.42** Closed queueing network for Problem 3.63.

Each time a packet exits the system at point $B$, an acknowledgment is sent back and reaches point $A$ after a fixed time $\overline{Z}$. At that time, a new packet is allowed to enter the system. Use Little's Theorem to find upper and lower bounds for the system throughput under two circumstances:

**(a)** The method of routing a packet to one of the lines $L_1, \ldots, L_K$ is unspecified.

**(b)** The routing method is such that whenever one of the lines $L_1, \ldots, L_K$ is idle, there is no packet waiting at any of the other lines.

**3.64** Consider the closed queueing network in Fig. 3.43. There are three customers who are doomed forever to cycle between queue 1 and queue 2. The service times at the queues are independent and exponentially distributed with mean $\mu_1$ and $\mu_2$. Assume that $\mu_2 < \mu_1$.

**(a)** The system can be represented by a four-state Markov chain. Find the transition rates of the chain.

**(b)** Find the steady-state probabilities of the states.

**(c)** Find the customer arrival rate at queue 1.

**(d)** Find the rate at which a customer cycles through the system.

**(e)** Show that the Markov chain is reversible. What does a departure from queue 1 in the forward process correspond to in the reversed process? Can the transitions of a single customer in the forward proces be associated with transitions of a single customer in the reverse process?

**3.65** Consider the closed queueing network of Section 3.8.2 and assume that the service rate $\mu_j(m)$ at the $j^{\text{th}}$ queue is independent of the number of customers $m$ in the queue [$\mu_j(m) = \mu_j$ for all $m$]. Show that the utilization factor $U_j(M) = \lambda_j(M)/\mu_j$ of the $j^{\text{th}}$ queue is given by

$$U_j(M) = \rho_j \frac{G(M-1)}{G(M)}$$

where $\rho_j = \overline{\lambda}_j/\mu_j$ (compare with Examples 3.21 and 3.22).

**3.66** *$M/M/1$ System with Multiple Classes of Customers.* Consider an $M/M/1$-like system with first-come first-serve service and multiple classes of customers denoted $c = 1, 2, \ldots, C$. Let $\lambda_i$ and $\mu_i$ be the arrival and service rate of class $i$.

**(a)** Model this system by a Markov chain and show that unless $\mu_1 = \mu_2 = \cdots = \mu_C$, its steady-state distribution does not have a product form. *Hint:* Consider a state $z = (c_1, c_2, \ldots, c_n)$ such that $\mu_{c_1} \neq \mu_{c_n}$. Write the global balance equations for state $z$.

**(b)** Suppose instead that the service discipline is last-come first-serve (as defined in Problem 3.49). Model the system by a Markov chain and show that the steady-state distribution has the product form

$$P(z) = P(c_1, c_2, \ldots, c_n) = \frac{\rho_{c_1}\rho_{c_2}\cdots\rho_{c_n}}{G}$$

where $\rho_c = \lambda_c/\mu_c$ and $G$ is a normalizing constant.

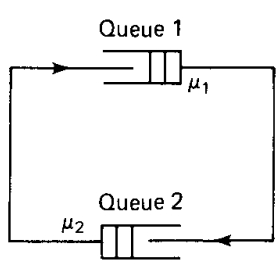



**Figure 3.43** Closed queueing network for Problem 3.64.

# APPENDIX A: REVIEW OF MARKOV CHAIN THEORY

The purpose of this appendix is to provide a brief summary of the results we need from discrete- and continuous-time Markov chain theory. We refer the reader to books on stochastic processes for detailed accounts.

## 3A.1 Discrete-Time Markov Chains

Consider a discrete-time stochastic process $\{X_n \mid n = 0, 1, 2, \ldots\}$ that takes values from the set of nonnegative integers, so the states that the process can be in are $i = 0, 1 \ldots$. The process is said to be a *Markov chain* if whenever it is in state $i$, there is a fixed probability $P_{ij}$ that it will next be in state $j$ regardless of the process history prior to arriving at $i$. That is, for all $n > 0$, $i_{n-1}, \ldots, i_0, i, j$,

$$\begin{aligned} P_{ij} &= P\{X_{n+1} = j \mid X_n = i,\ X_{n-1} = i_{n-1}, \ldots, X_0 = i_0\} \\ &= P\{X_{n+1} = j \mid X_n = i\} \end{aligned}$$

We refer to $P_{ij}$ as the *transition probabilities*. They must satisfy

$$P_{ij} \geq 0\,, \quad \sum_{j=0}^{\infty} P_{ij} = 1\,, \qquad i = 0, 1, \ldots$$

The corresponding transition probability matrix is denoted

$$P = \begin{bmatrix} P_{00} & P_{01} & P_{02} & \cdots \\ P_{10} & P_{11} & P_{12} & \cdots \\ \cdot & \cdot & \cdot & \cdot \\ P_{i0} & P_{i1} & P_{i2} & \cdots \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

We will concentrate on the case where the number of states is infinite. There are analogous notions and results for the case where the number of states is finite.

Consider the $n$-step transition probabilities

$$P_{ij}^n = P\{X_{n+m} = j \mid X_m = i\}\,, \qquad n \geq 0, i, j \geq 0\ .$$

The *Chapman–Kolmogorov equations* provide a method for calculating $P_{ij}^n$. They are given by

$$P_{ij}^{n+m} = \sum_{k=0}^{\infty} P_{ik}^n P_{kj}^m\,, \qquad n, m \geq 0, i, j \geq 0$$

From these equations, we see that $P_{ij}^n$ are the elements of the matrix $P^n$ (the transition probability matrix $P$ raised to the $n^{\text{th}}$ power).

We say that two states $i$ and $j$ *communicate* if for some $n$ and $n'$, we have $P_{ij}^n > 0$ and $P_{ji}^{n'} > 0$. If all states communicate, we say that the Markov chain is *irreducible*.

We say that a state $i$ of a Markov chain is *periodic* if there exists some integer $m \geq 1$ such that $P_{ii}^m > 0$ and some integer $d > 1$ such that $P_{ii}^n > 0$ only if $n$ is a multiple of $d$. A Markov chain is said to be *aperiodic* if none of its states is periodic. A probability distribution $\{p_j \mid j \geq 0\}$ is said to be a *stationary distribution* for the Markov chain if

$$p_j = \sum_{i=0}^{\infty} p_i P_{ij} \qquad j = 0, 1, \ldots \tag{3A.1}$$

We will restrict attention to irreducible and aperiodic Markov chains, since this is the only type we will encounter. For such a chain, it can be shown that the limit

$$p_j = \lim_{n \to \infty} P\{X_n = j \mid X_0 = i\}, \qquad i = 0, 1, \ldots$$

exists and is independent of the starting state $X_0 = i$. Furthermore, we have (with probability 1)

$$p_j = \lim_{k \to \infty} \frac{\text{Number of visits to state } j \text{ up to time } k}{k}$$

which leads to the interpretation that $p_j$ is the proportion of time or the frequency with which the process visits $j$, a time average interpretation. Note again that this frequency does not depend on the starting state. The following result will be of primary interest:

**Theorem.** In an irreducible, aperiodic Markov chain, there are two possibilities for the scalars $p_j = \lim_{n \to \infty} P\{X_n = j \mid X_0 = i\}$:

1. $p_j = 0$ for all $j \geq 0$, in which case the chain has no stationary distribution.
2. $p_j > 0$ for all $j \geq 0$, in which case $\{p_j \mid j \geq 0\}$ is the unique stationary distribution of the chain [*i.e.*, it is the only probability distribution satisfying Eq. (3A.1)].

A typical example of case 1 is a queueing system where the arrival rate exceeds the service rate, and the number of customers in the system increases to $\infty$, so the steady-state probability $p_j$ of having any finite number of customers $j$ in the system is zero. Note that case 1 never arises when the number of states is finite. In particular, for every irreducible and aperiodic Markov chain with states $j = 0, 1, \ldots, m$, there exists a unique probability distribution $\{p_j \mid j = 0, 1, \ldots, m\}$ satisfying $p_j = \sum_{i=0}^{m} p_i P_{ij}$ and $p_j > 0$ for all $j$.

In case 2, there arises the issue of characterizing the stationary distribution $\{p_j \mid j \geq 0\}$. For queueing systems, the following equations are often useful. Multiplying the equation $\sum_{i=0}^{\infty} P_{ji} = 1$ by $p_j$ and using Eq. (3A.1), we have

$$p_j \sum_{i=0}^{\infty} P_{ji} = \sum_{i=0}^{\infty} p_i P_{ij}, \qquad j = 0, 1, \ldots \tag{3A.2}$$

These equations are known as the *global balance equations*. Note that $p_i P_{ij}$ may be viewed as the long-term frequency of transitions from $i$ to $j$. Thus the global balance equations state that at equilibrium, the frequency of transitions out of $j$ [left side of Eq. (3A.2)] equals the frequency of transitions into $j$ [right side of Eq. (3A.2)].

A typical approach for finding the stationary distribution of an irreducible, aperiodic Markov chain is to try to solve the global balance equations. If a distribution satisfying these equations is found, then by the preceding theorem, it must be the stationary distribution.

The global balance equations can be generalized to apply to an entire set of states. Consider a subset of states $S$. By adding Eq. (3A.2) over all $j \in S$, we obtain

$$\sum_{j \in S} p_j \sum_{i \notin S} P_{ji} = \sum_{i \notin S} p_i \sum_{j \in S} P_{ij} \tag{3A.3}$$

An intuitive explanation of these equations is based on the fact that when the Markov chain is irreducible, the state (with probability 1) will return to the set $S$ infinitely many times. Therefore, for each transition out of $S$ there must be (with probability 1) a reverse transition into $S$ at some later time. As a result, the frequency of transitions out of $S$ equals the frequency of transitions into $S$. This is precisely the meaning of the global balance equations (3A.3).

## 3A.2 Detailed Balance Equations

As an application of the global balance equations, consider a Markov chain typical of queueing systems and, more generally, *birth–death* systems where two successive states can only differ by unity, that is, $P_{i,j} = 0$ if $|i - j| > 1$, cf. Fig. 3A.1. We assume that $P_{i,i+1} > 0$ and $P_{i+1,i} > 0$ for all $i$. This is a necessary and sufficient condition for the chain to be irreducible. Consider the sets of states

$$S = \{0, 1, \ldots, n\}$$

Application of Eq. (3A.3) yields

$$p_n P_{n,n+1} = p_{n+1} P_{n+1,n}\,, \qquad n = 0, 1, \ldots \tag{3A.4}$$

(*i.e.*, in steady-state), the frequency of transitions from $n$ to $n+1$ equals the frequency of transitions from $n+1$ to $n$. These equations can be very useful in computing the stationary distribution $\{p_j \mid j \geq 0\}$ (see Sections 3.3 and 3.4).

Equation (3A.4) is a special case of the equations

$$p_j P_{ji} = p_i P_{ij}\,, \qquad i, j \geq 0 \tag{3A.5}$$

known as the *detailed balance equations*. These equations imply the global balance equations (3A.2) but need not hold in any given Markov chain. However, in many important special cases, they do hold and greatly simplify the calculation of the stationary

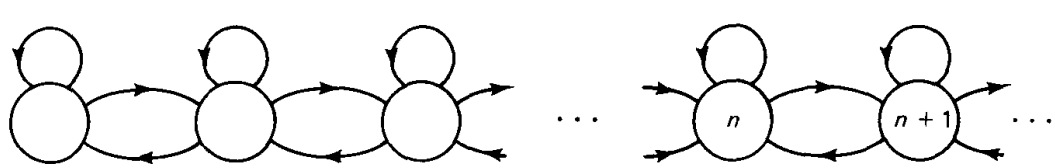



**Figure 3A.1** Transition probability diagram for a birth–death process.

distribution. A common approach is to hypothesize the validity of the detailed balance equations and to try to solve them for the steady-state probabilities $p_j$, $j \geq 0$. There are two possibilities; either the system (3A.5) together with $\sum_j p_j = 1$ is inconsistent or else a distribution $\{p_j \mid j \geq 0\}$ satisfying Eq. (3A.5) will be found. In the latter case, this distribution will also satisfy the global balance equations (3A.2), so by the theorem given earlier, it is the unique stationary distribution.

### 3A.3 Partial Balance Equations

Some Markov chains have the property that their stationary distribution $\{p_j \mid j \geq 0\}$ satisfies a set of equations which is intermediate between the global and the detailed balance equations. For every node $j$, consider a partition $S_j^1, \ldots, S_j^k$ of the complementary set of nodes $\{i \mid i \geq 0, i \neq j\}$ and the equations

$$p_j \sum_{i \in S_j^m} P_{ji} = \sum_{i \in S_j^m} p_i P_{ij}\,, \qquad m = 1, 2, \ldots, k \tag{3A.6}$$

Equations of the form above are known as a set of *partial balance equations*. If a distribution $\{p_j \mid j \geq 0\}$ solves a set of partial balance equations, it will also solve the global balance equations, so it will be the unique stationary distribution of the chain. A technique that often proves useful is to guess the right set of partial balance equations satisfied by the stationary distribution and then proceed to solve them.

### 3A.4 Continuous-Time Markov Chains

A continuous-time Markov chain is a process $\{X(t) \mid t \geq 0\}$ taking values from the set of states $i = 0, 1, \ldots$ that has the property that each time it enters state $i$:

1. The time it spends in state $i$ is exponentially distributed with parameter $\nu_i$. We may view $\nu_i$ as the rate (in transitions/sec) at which the process makes a transition when at state $i$.
2. When the process leaves state $i$, it will enter state $j$ with probability $P_{ij}$, where $\sum_j P_{ij} = 1$.

   We may view

$$q_{ij} = \nu_i P_{ij}$$

as the rate (in transitions/sec) at which the process makes a transition to $j$ when at state $i$. Consequently, we call $q_{ij}$ the *transition rate* from $i$ to $j$.

We will be interested in chains for which the discrete-time Markov chain with transition probabilities $P_{ij}$ (called the *embedded chain*) is irreducible. We also require a technical condition, namely that the number of transitions in any finite length of time is finite with probability 1; chains with this property are called *regular*. (Nonregular chains almost never arise in queueing systems of interest. For an example, see [Ros83], p. 142.)

Under the preceding conditions, it can be shown that the limit

$$p_j = \lim_{t \to \infty} P\{X(t) = j \mid X(0) = i\} \tag{3A.7}$$

exists and is independent of the initial state $i$. We refer to $p_j$ as the steady-state occupancy probability of state $j$. It can be shown that if $T_j(t)$ is the time spent in state $j$ up to time $t$, then, regardless of the initial state, we have with probability 1,

$$p_j = \lim_{t \to \infty} \frac{T_j(t)}{t} \tag{3A.8}$$

that is, $p_j$ can be viewed as the long-term proportion of time the process spends in state $j$. It can be shown also that either the occupancy probabilities are all zero or else they are all positive and they sum to unity. Queueing systems where the arrival rate is larger than the service rate provide examples where all occupancy probabilities are zero.

The *global balance equations* for a continuous-time Markov chain take the form

$$p_j \sum_{i=0}^{\infty} q_{ji} = \sum_{i=0}^{\infty} p_i q_{ij}, \qquad j = 0, 1, \ldots \tag{3A.9}$$

It can be shown that if a probability distribution $\{p_j \mid j \geq 0\}$ satisfies these equations, then each $p_j$ is the steady-state occupancy probability of state $j$.

To interpret the global balance equations, we note that since $p_i$ is the proportion of time the process spends in state $i$, it follows that $p_i q_{ij}$ is the frequency of transitions from $i$ to $j$ (average number of transitions from $i$ to $j$ per unit time). It is seen therefore that the global balance equations (3A.9) express the natural fact that the frequency of transitions out of state $j$ (the left-side term $p_j \sum_{i=1}^{\infty} q_{ji}$) is equal to the frequency of transitions into state $j$ (the right-side term $\sum_{i=0}^{\infty} p_i q_{ij}$).

The continuous-time analog of the detailed balance equations for discrete-time chains is

$$p_j q_{ji} = p_i q_{ij}, \qquad i, j = 0, 1, \ldots$$

These equations hold in birth–death systems where $q_{ij} = 0$ for $|i - j| > 1$, but need not hold in other types of Markov chains. They express the fact that the frequencies of transitions from $i$ to $j$ and from $j$ to $i$ are equal. One can also write a set of partial balance equations and attempt to solve them for the distribution $\{p_j \mid j \geq 0\}$. If a solution can be found, it provides the stationary distribution of the continuous chain.

To understand the relationship between the global balance equations (3A.9) for continuous-time chains and the global balance equations (3A.2) for discrete-time chains, consider any $\delta > 0$, and the discrete-time Markov chain $\{X_n \mid n \geq 0\}$, where

$$X_n = X(n\delta), \qquad n = 0, 1, \ldots$$

The stationary distribution of $\{X_n\}$ is clearly $\{p_j \mid j \geq 0\}$, the occupancy distribution of the continuous chain [cf. Eq. (3A.7)]. The transition probabilities of $\{X_n \mid n \geq 0\}$ can be derived by using the properties of the exponential distribution and a derivation which is very similar to the one used in Section 3.3.1 for the Markov chain of the $M/M/1$ queue. We obtain

$$\overline{P}_{ij} = \delta q_{ij} + o(\delta), \qquad i \neq j$$

$$\overline{P}_{jj} = 1 - \delta \sum_{\substack{i=0 \\ i \neq j}}^{\infty} q_{ji} + o(\delta)$$

Using these expressions and Eq. (3A.1), which is equivalent to the global balance equations for the discrete chain, we obtain

$$p_j = \sum_{i=0}^{\infty} p_i \overline{P}_{ij} = p_j\big(1 - \delta \sum_{\substack{i=0\\ i\neq j}}^{\infty} q_{ji} + o(\delta)\big) + \sum_{\substack{i=0\\ i\neq j}} p_i\big(\delta q_{ij} + o(\delta)\big)$$

and dividing by $\delta$ and letting $\delta \to 0$, we obtain the global balance equations (3A.9) for the continuous chain.

## 3A.5 Drift and Stability

Suppose that we are given an irreducible, aperiodic, discrete-time Markov chain. In many situations one is particularly interested in whether the chain has a stationary distribution $\{p_j\}$. In this case, $p_j > 0$ for all $j$ and the chain is "stable" in the sense that all states are visited infinitely often with probability 1. The notion of *drift*, defined as

$$D_i = E\{X_{n+1} - X_n \mid X_n = i\} = \sum_{k=-i}^{\infty} kP_{i(i+k)}, \qquad i = 0, 1, \ldots$$

is particularly useful in this respect. Roughly speaking, the sign of $D_i$ indicates whether, starting at $i$, the state tends to increase ($D_i > 0$) or decrease ($D_i < 0$). Intuitively, the chain will be stable if the drift is negative for all large enough states. This is established in the following lemma:

**Stability Lemma [Pak69].** Suppose that $D_i < \infty$ for all $i$, and that for some scalar $\delta > 0$ and integer $\bar{i} \geq 0$ we have

$$D_i \leq -\delta, \qquad \text{for all } i > \bar{i}$$

Then the Markov chain has a stationary distribution.

*Proof:* Let $\beta = \max_{i \leq \bar{i}} D_i$. We have for each state $i$

$$\begin{aligned}
E\{X_n \mid X_0 = i\} - i &= \sum_{k=1}^{n} E\{X_k - X_{k-1} \mid X_0 = i\} \\
&= \sum_{k=1}^{n} \sum_{j=0}^{\infty} E\{X_k - X_{k-1} \mid X_{k-1} = j\} P\{X_{k-1} = j \mid X_0 = i\} \\
&\leq \sum_{k=1}^{n} \Bigg[ \beta \sum_{j=0}^{\bar{i}} P\{X_{k-1} = j \mid X_0 = i\} \\
&\qquad -\delta \left( 1 - \sum_{j=0}^{\bar{i}} P\{X_{k-1} = j \mid X_0 = i\} \right) \Bigg]
\end{aligned}$$

$$= (\beta + \delta) \sum_{k=1}^{n} \sum_{j=0}^{\bar{i}} P\{X_{k-1} = j \mid X_0 = i\} - n\delta$$

from which we obtain

$$0 \leq E\{X_n \mid X_0 = i\} \leq n \left[ (\beta + \delta) \sum_{j=0}^{\bar{i}} \left( n^{-1} \sum_{k=1}^{n} P\{X_{k-1} = j \mid X_0 = i\} \right) - \delta \right] + i$$

Dividing by $n$ and taking the limit as $n \to \infty$, we obtain

$$0 \leq (\beta + \delta) \sum_{j=0}^{\bar{i}} p_j - \delta$$

which implies that $p_j > 0$ for some $j \in \{0, \ldots, \bar{i}\}$. Since the chain is assumed irreducible and aperiodic, it follows from the theorem of Section 3A.1 that there exists a stationary distribution. **Q.E.D.**

We also give without proof a converse to the preceding lemma:

**Instability Lemma [Kap79].** Suppose that there exist integers $\bar{i} \geq 0$ and $k$ such that

$$D_i > 0, \qquad \text{for all } i > \bar{i}$$

and

$$P_{ij} = 0, \qquad \text{for all } i \text{ and } j \text{ such that } 0 \leq j \leq i - k$$

Then the Markov chain does not have a stationary distribution; that is, $p_j = 0$ for all $j$.

## APPENDIX B: SUMMARY OF RESULTS

### Notation

$p_n$ = Steady-state probability of having $n$ customers in the system

$\lambda$ = Arrival rate (inverse of average interarrival time)

$\mu$ = Service rate (inverse of average service time)

$N$ = Average number of customers in the system

$N_Q$ = Average number of customers waiting in queue

$T$ = Average customer time in the system

$W$ = Average customer waiting time in queue (does not include service time)

$\overline{X}$ = Average service time

$\overline{X^2}$ = Second moment of service time

### Little's Theorem

$$N = \lambda T$$

$$N_Q = \lambda W$$

### Poisson distribution with parameter *m*

$$p_n = \frac{e^{-m} m^n}{n!}, \quad n = 0, 1, \ldots$$

$$\text{Mean} = \text{Variance} = m$$

### Exponential distribution with parameter λ

$$P\{\tau \le s\} = 1 - e^{-\lambda s}, \qquad s \ge 0$$

$$\text{Density: } p(\tau) = \lambda e^{-\lambda \tau}$$

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

$$\text{Variance} = \frac{1}{\lambda^2}$$

### Summary of *M*/*M*/1 system results

1. Utilization factor (proportion of time the server is busy)

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

2. Probability of $n$ customers in the system

$$p_n = \rho^n (1 - \rho), \qquad n = 0, 1, \ldots$$

3. Average number of customers in the system

$$N = \frac{\rho}{1 - \rho}$$

4. Average customer time in the system

$$T = \frac{\rho}{\lambda(1 - \rho)} = \frac{1}{\mu - \lambda}$$

5. Average number of customers in queue

$$N_Q = \frac{\rho^2}{1 - \rho}$$

6. Average waiting time in queue of a customer

$$W = \frac{\rho}{\mu - \lambda}$$

### Summary of *M*/*M*/*m* system results

1. Ratio of arrival rate to maximal system service rate

$$\rho = \frac{\lambda}{m\mu}$$

2. Probability of $n$ customers in the system

$$p_0 = \left[\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!(1-\rho)}\right]^{-1}, \qquad n = 0$$

$$p_n = \begin{cases} p_0 \dfrac{(m\rho)^n}{n!}, & n \le m \\ p_0 \dfrac{m^m \rho^n}{m!}, & n > m \end{cases}$$

3. Probability that an arriving customer has to wait in queue ($m$ customers or more in the system)

$$P_Q = \frac{p_0 (m\rho)^m}{m!(1-\rho)} \qquad \text{(Erlang C Formula)}$$

4. Average waiting time in queue of a customer

$$W = \frac{\rho P_Q}{\lambda(1-\rho)}$$

5. Average number of customers in queue

$$N_Q = \frac{\rho P_Q}{1-\rho}$$

6. Average customer time in the system

$$T = \frac{1}{\mu} + W$$

7. Average number of customers in the system

$$N = m\rho + \frac{\rho P_Q}{1-\rho}$$

### Summary of $M/M/m/m$ system results

1. Probability of $m$ customers in the system

$$p_0 = \left[\sum_{n=0}^{m} \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}\right]^{-1}$$

$$p_n = p_0 \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}, \qquad n = 1, 2, \ldots, m$$

2. Probability that an arriving customer is lost

$$p_m = \frac{(\lambda/\mu)^m / m!}{\sum_{n=0}^{m} (\lambda/\mu)^n / n!} \qquad \text{(Erlang B Formula)}$$

### Summary of $M/G/1$ system results

1. Utilization factor

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

2. Mean residual service time

$$R = \frac{\lambda \overline{X^2}}{2}$$

3. Pollaczek–Khinchin formula

$$W = \frac{R}{1-\rho} = \frac{\lambda \overline{X^2}}{2(1-\rho)}$$

$$T = \frac{1}{\mu} + W$$

$$N_Q = \frac{\lambda^2 \overline{X^2}}{2(1-\rho)}$$

$$N = \rho + \frac{\lambda^2 \overline{X^2}}{2(1-\rho)}$$

4. Pollaczek–Khinchin formula for $M/G/1$ queue with vacations

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho)} + \frac{\overline{V^2}}{2\overline{V}}$$

$$T = \frac{1}{\mu} + W$$

where $\overline{V}$ and $\overline{V^2}$ are the first two moments of the vacation interval.

### Summary of reservation/polling results

1. Average waiting time ($m$-user system, unlimited service)

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho)} + \frac{(m-\rho)\overline{V}}{2(1-\rho)} + \frac{\sigma_V^2}{2\overline{V}} \qquad \text{(exhaustive)}$$

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho)} + \frac{(m+\rho)\overline{V}}{2(1-\rho)} + \frac{\sigma_V^2}{2\overline{V}} \qquad \text{(partially gated)}$$

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho)} + \frac{(m+2-\rho)\overline{V}}{2(1-\rho)} + \frac{\sigma_V^2}{2\overline{V}} \qquad \text{(gated)}$$

where $\rho = \lambda/\mu$, and $\overline{V}$ and $\sigma_V^2$ are the mean and variance of the reservation intervals, respectively, averaged over all users

$$\overline{V} = \frac{1}{m} \sum_{\ell=0}^{m-1} \overline{V_\ell}$$

$$\sigma_V^2 = \frac{1}{m} \sum_{\ell=0}^{m-1} \left( \overline{V_\ell^2} - \overline{V_\ell}^2 \right)$$

2. Average waiting time ($m$-user system, limited service)

$$W = \frac{\lambda\overline{X^2}}{2(1-\rho-\lambda\overline{V})} + \frac{(m+\rho)\overline{V}}{2(1-\rho-\lambda\overline{V})} + \frac{\sigma_V^2(1-\rho)}{2\overline{V}(1-\rho-\lambda\overline{V})} \qquad \text{(partially gated)}$$

$$W = \frac{\lambda\overline{X^2}}{2(1-\rho-\lambda\overline{V})} + \frac{(m+2-\rho-2\lambda\overline{V})\overline{V}}{2(1-\rho-\lambda\overline{V})} + \frac{\sigma_V^2(1-\rho)}{2\overline{V}(1-\rho-\lambda\overline{V})} \qquad \text{(gated)}$$

3. Average time in the system

$$T = \frac{1}{\mu} + W$$

## Summary of priority queueing results

1. *Nonpreemptive priority.* Average waiting time in queue for class $k$ customers

$$W_k = \frac{\sum_{i=1}^{n} \lambda_i \overline{X_i^2}}{2(1-\rho_1-\cdots-\rho_{k-1})(1-\rho_1-\cdots-\rho_k)}$$

2. *Nonpreemptive priority.* Average time in the system for class $k$ customers

$$T_k = \frac{1}{\mu_k} + W_k$$

3. *Preemptive resume priority.* Average time in the system for class $k$ customers

$$T_k = \frac{(1/\mu_k)(1-\rho_1-\cdots-\rho_k) + R_k}{(1-\rho_1-\cdots-\rho_{k-1})(1-\rho_1-\cdots-\rho_k)}$$

where

$$R_k = \frac{\sum_{i=1}^{k} \lambda_i \overline{X_i^2}}{2}$$

## Heavy traffic approximation for the $G/G/1$ system

Average waiting time in queue satisfies

$$W \le \frac{\lambda(\sigma_a^2 + \sigma_b^2)}{2(1-\rho)}$$

where

$\sigma_a^2$ = Variance of the interarrival times

$\sigma_b^2$ = Variance of the service times

$\lambda$ = Average interarrival time

$\rho$ = Utilization factor $\lambda/\mu$, where $1/\mu$ is the average service time