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 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 from the time the last bit is transmitted at the head node of the link until the time it is received at the tail node. This is proportional to the physical distance between transmitter and receiver and is ordinarily small except in the case of a satellite 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 extensively developed 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 both on 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 \( L/C \) sec 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 was allocated to a single channel. The transmission time of a packet that is \( L \) bits long is \( Lm/C \), or \( m \) times longer 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 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.
3.2 QUEUEING MODELS—LITTLE'S THEOREM

We consider queueing systems where customers arrive at random times to obtain service. The probability distribution of the time between two successive arrivals (the interarrival time), and the probability distribution of the customers' service time are given.

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 dwell on very much), customers represent active conversations (or virtual circuits) between points in a network and service time corresponds to the duration of a conversation.

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).

We first need to clarify the meaning of the terms above. Let us denote

\[
p_n(t) = \text{Probability of } n \text{ customers waiting in queue or under service at time } t
\]

The typical situation is one whereby we are given the initial probabilities \( p_n(0) \) at time 0 and enough statistical information is provided to determine, at least in principle, the probabilities \( p_n(t) \) for all times \( t \). Then denoting

\[
\overline{N}(t) = \text{Average number in the system at time } t
\]

we have

\[
\overline{N}(t) = \sum_{n=0}^{\infty} np_n(t)
\]

Note that both \( \overline{N}(t) \) and \( p_n(t) \) depend on \( t \) as well as the initial probability distribution \( \{p_0(0), p_1(0), \ldots \} \). However, the queueing systems that we will consider typically reach equilibrium in the sense that for some \( p_n \) and \( N \) (independent of the initial distribution), we have

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

and

\[
N = \sum_{n=0}^{\infty} np_n = \lim_{t \to \infty} \overline{N}(t)
\]
We will be interested primarily in the equilibrium probabilities and the average number in the system. Note that it is possible that \( N = \infty \) and this will occur whenever the arrival rate exceeds the service capacity of the system. Individual sample functions of the number of customers in the system will be denoted by \( N(t) \). The time average of such a sample function in the interval \([0, t]\) is defined by

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

Almost every system of interest to us is ergodic in the sense that

\[
\lim_{t \to \infty} N_t = \lim_{t \to \infty} \bar{N}(t) = N
\]

holds with probability one. The equality of long term time average and ensemble average of various stochastic processes will often be accepted in this chapter on intuitive grounds since a rigorous mathematical justification requires technical arguments that are beyond the scope of this text.

Regarding average delay per customer, the situation is one whereby enough statistical information is available 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 \( k \)th customer, denoted \( \bar{T}_k \), typically converges as \( k \to \infty \) to a steady-state value

\[
T = \lim_{k \to \infty} \bar{T}_k
\]

The limit above is what we will call average delay per customer. (Again, \( T = \infty \) is possible.) For the systems of interest to us, the steady-state average delay \( T \) is also equal (with probability one) to the long-term time average of customer delay, i.e.,

\[
T = \lim_{k \to \infty} \bar{T}_k = \lim_{k \to \infty} \frac{1}{k} \sum_{i=1}^{k} T_i
\]

where \( T_i \) is the delay of the \( i \)th customer.

The average number in the system \( N \) and the average delay \( T \) 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

\[
N = \lambda T
\]

where

\[
\lambda = \text{Average customer arrival rate}
\]

and is given by

\[
\lambda = \lim_{t \to \infty} \frac{\text{Expected number of arrivals in the interval } [0, t]}{t}
\]
(We will be assuming that the limit above exists.) Phenomena reflecting Little's Theorem are familiar from everyday experience. 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.

Little's Theorem is really an accounting identity and its derivation is very simple. We will give a graphical proof, which assumes that customers are served in the order they arrive. A similar proof is possible for the case where the order of service is arbitrary (see Problems 3.31 and 3.32). For any sample system history let us denote:

$$\alpha(t) = \text{Number of arrivals in the interval } [0, t]$$
$$\beta(t) = \text{Number of departures in the interval } [0, t]$$

Assuming an empty system at time 0, the number in the system at time $t$ is

$$N(t) = \alpha(t) - \beta(t)$$

Let $t_i$ and $T_i$ be the time of arrival and the time spent in the system, respectively, by the $i^{th}$ customer. Consider any time $t$ and the shaded area in Fig. 3.1 which lies between the graphs of $\alpha(\tau)$ and $\beta(\tau)$ up to time $t$. This area can be expressed as

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

but also as

$$\sum_{i=1}^{\beta(t)} T_i + \sum_{i=\beta(t)+1}^{\alpha(t)} (t - t_i)$$

Dividing both expressions above by $t$ and equating them, we obtain

$$N_t = \lambda_t T_t \quad (3.2)$$

where

$$N_t = \frac{\int_0^t N(\tau) \, d\tau}{t} = \text{Time average of the number of customers in the system in the interval } [0, t]$$
$$\lambda_t = \frac{\alpha(t)}{t} = \text{Time average of the customer arrival rate in the interval } [0, t]$$

$$T_t = \frac{\sum_{i=1}^{\beta(t)} T_i + \sum_{i=\beta(t)+1}^{\alpha(t)} (t - t_i)}{\alpha(t)} = \text{Time average of the time a customer spends in the system in the interval } [0, t]$$
Figure 3.1 Prove of Little’s Theorem. The shaded area can be expressed both as \( \int_0^t N(r) \, dr \) and as \( \sum_{i=1}^{\alpha(t)} T_i + \sum_{i=\beta(t)+1}^{\alpha(t)} (t-t_i) \). Dividing both expressions by \( t \), equating them, and taking the limit as \( t \to \infty \) gives Little’s Theorem.

Assuming that

\[ N_t \to N, \ \lambda_t \to \lambda, \ T_t \to T \]

we obtain from Eq. (3.2) the desired formula.

Note that the expression \( T_t \) includes the total time spent in the system for all the arrivals from 1 to \( \beta(t) \), but omits the time spent beyond \( t \) for the customers still in the system at time \( t \). Assuming that \( N_t \to N < \infty \), this end effect due to customers in the system at time \( t \) will be small relative to the accumulated time in the system of customers 1 to \( \beta(t) \), and \( T_t \) for large \( t \) can be interpreted as the time average of the system time.

Strictly speaking, for the argument above to be correct, we must be assured that the time averages \( N_t, \lambda_t, T_t \) converge with probability one to the corresponding ensemble averages \( N, \lambda, T \). This is true in just about every case of interest to us, and in subsequent analysis, we will accept Little’s Theorem without further scrutiny.

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

**Example 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 $\bar{X}$ is the average transmission time, then Little’s Theorem gives the average number of packets under transmission as

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

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 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**

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

$$N = \lambda T = \alpha + \frac{P}{K}$$
Figure 3.2 The number in the system in Example 3, \( N(t) \), is deterministic and does not converge as \( t \to \infty \). However, Little’s Theorem holds if \( N \), \( \Lambda \), and \( T \) are interpreted as time averages.

One should be careful about interpreting correctly the formula in this example. Here the number in the system \( N(t) \) is a deterministic function of time. Its form is shown in Fig. 3.2 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 is correct provided \( N \) is viewed as a long-term time average of \( N(t) \), i.e.,

\[
N = \lim_{t \to \infty} \frac{\int_0^t N(\tau) \, d\tau}{t}
\]

Example 4

Consider a window flow control system (as described in subsection 2.8.1) with a window of size \( N \) for each session. Suppose that a session always has packets to send and that acknowledgements take negligible time; then, when packet \( i \) arrives at the destination, packet \( i + N \) is immediately introduced into the network. Since the number of packets in the system per session is always \( N \), 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 \( N = \Lambda T \). Thus, if congestion builds up in the network and \( T \) increases, \( \Lambda \) must decrease. Note also that if the network is congested and
capable of delivering only $\lambda$ packets per unit time for each session, then increasing
the window size $N$ for all sessions merely serves to increase the delay $T$.

**Example 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*.) Suppose that the
average customer service time is $\overline{X}$. We want to find the average 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

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

**Example 6: 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. 3.3.

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. 3.3), we have

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

(3.3)

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

$$T = R + D$$

(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
Figure 3.3  $N$ 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.

where the user's job has to wait for the jobs of all the other users; compare with Ex. 5), we have

$$R + P \leq T \leq R + NP$$

(3.5)

Combining this relation with Eq. (3.3), we obtain

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

(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, i.e.,

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

(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 relations (3.6) and (3.7), we obtain the bounds

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

(3.8)
for the maximum attainable throughput. By using \( T = N/\lambda \), we also obtain bounds for the average user delay when the system is fully loaded

\[
\max \{ NP, R + P \} \leq T \leq R + NP
\]

(3.9)

These relations are illustrated in Fig. 3.4.

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 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 interarrival 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

\[
N = \lambda T, \quad NQ = \lambda W
\]
between the basic quantities,

\[ N : \text{Average number of customers in the system} \]
\[ T : \text{Average customer time in the system} \]
\[ N_Q : \text{Average number of customers waiting in queue} \]
\[ W : \text{Average customer waiting time in queue} \]
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

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

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 $M/C/1$ system. The price paid for this generality is that the characterization of the steady-state probabilities is less convenient and simple 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 subsections 3.3.1 and 3.3.2.

### 3.3.1 Main Results

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 number 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 $t, \tau > 0$,

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

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

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

$$P \{ \tau_n \leq s \} = 1 - e^{-\lambda s}, \quad s \geq 0$$  \hspace{1cm} (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.)

(b) For every $t \geq 0$ and $\delta \geq 0$

$$P \{ A(t+\delta) - A(t) = 0 \} = 1 - \lambda \delta + o(\delta)$$  \hspace{1cm} (3.12)
$$P \{ A(t+\delta) - A(t) = 1 \} = \lambda \delta + o(\delta)$$  \hspace{1cm} (3.13)
$$P \{ A(t+\delta) - A(t) \geq 2 \} = o(\delta)$$  \hspace{1cm} (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 using Eq. (3.10) (see Prob. 3.10).

Note that if the arrivals in $n$ disjoint intervals are independent and Poisson distributed with parameters $\lambda \tau_1, \ldots, \lambda \tau_n$, then the number of arrivals in the union of the intervals is Poisson distributed with parameter $\lambda (\tau_1 + \cdots + \tau_n)$. This follows from properties of the Poisson distribution and guarantees that the requirement of Eq. (3.10) is consistent with the independence requirement in the definition of the Poisson process (see Prob. 3.10). Another fact that we will frequently use is that if two or more independent Poisson processes $A_1, \ldots, A_k$ are merged into a single process $A = A_1 + A_2 + \cdots + A_k$, then the latter process is Poisson with a rate equal to the sum of the rates of its components (see Prob. 3.10).

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

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

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

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

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

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

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

The memoryless 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 $\{N(t)|t \geq 0\}$ 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. 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 $\{N_k|k = 0, 1, \ldots\}$ is a discrete-time Markov chain. Let $P_{ij}$ denote the corresponding transition probabilities

$$P_{ij} = P\{N_{k+1} = j|N_k = i\}$$
Figure 3.5 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.

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), we have

\[ P_{00} = 1 - \lambda \delta + o(\delta) \]  
\[ P_{ii} = 1 - \lambda \delta - \mu \delta + o(\delta), \quad i \geq 1 \]  
\[ P_{i,i+1} = \lambda \delta + o(\delta), \quad i \geq 0 \]  
\[ P_{i,i-1} = \mu \delta + o(\delta), \quad i \geq 1 \]  
\[ P_{ij} = o(\delta), \quad i \text{ and } j \neq i, i+1, i-1 \]

To see how these equations are verified, note that, when at a state $i \geq 1$, the probabilities of $0$ arrivals and $0$ departures in an 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 $\delta$,

\[ P\{0 \text{ customers arrive and } 0 \text{ depart in } I_k\} = 1 - \lambda \delta - \mu \delta + o(\delta) \]  
\[ \text{Similarly, we have that} \]
\[ P\{0 \text{ customers arrive and } 1 \text{ departs in } I_k\} = \mu \delta + o(\delta) \]
\[ P\{1 \text{ customer arrives and } 0 \text{ depart in } I_k\} = \lambda \delta + o(\delta) \]

These probabilities add up to one plus $o(\delta)$. Thus, the probability of more than one arrival or departure is negligible for $\delta$ small. This means 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. 3.5 where we have omitted the terms $o(\delta)$.

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 for any \( k \geq 1, n \geq 0 \), during the time from \( \delta \) to \( k\delta \), 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, in steady state, the probability that the system is in state \( n \) and makes a transition to \( n + 1 \) at the next transition instant is the same as the probability that the system is in state \( n + 1 \) and makes a transition to \( n \), i.e.,

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

(3.20)

(These 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 to derive Eq. (3.20).) Since \( p_n \) is independent of \( \delta \), by taking the limit in Eq. (3.20) as \( \delta \to 0 \), we obtain

\[
p_{n+1} = \rho p_n, \quad n = 0, 1, \ldots
\]

where

\[
\rho = \frac{\lambda}{\mu}
\]

It follows that

\[
p_{n+1} = \rho^{n+1} p_0, \quad n = 0, 1, \ldots
\]

(3.21)

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

\[
1 = \sum_{n=0}^{\infty} p_n = \sum_{n=0}^{\infty} \rho^n p_0 = \frac{p_0}{1 - \rho}
\]

(3.22)

This equation, together with Eq. (3.21), gives finally

\[
p_n = \rho^n (1 - \rho), \quad n = 0, 1, \ldots
\]

(3.23)

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

\[
N = \lim_{t \to \infty} E\{N(t)\} = \sum_{n=0}^{\infty} n p_n = \sum_{n=0}^{\infty} n \rho^n (1 - \rho)
\]

\[
= \rho (1 - \rho) \sum_{n=0}^{\infty} n \rho^{n-1} = \rho (1 - \rho) \frac{\partial}{\partial \rho} \left( \sum_{n=0}^{\infty} \rho^n \right)
\]

\[
= \rho (1 - \rho) \frac{\partial}{\partial \rho} \left( \frac{1}{1 - \rho} \right) = \rho (1 - \rho) \frac{1}{(1 - \rho)^2}
\]

and, finally, using \( \rho = \lambda/\mu \)

\[
N = \frac{\rho}{1 - \rho} = \frac{\lambda}{\mu - \lambda}
\]

(3.24)
This equation is shown in the diagram of Fig. 3.6. As \( \rho \) increases, so does \( N \), and as \( \rho \to 1 \), we have \( N \to \infty \). The diagram 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.

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

\[
T = \frac{N}{\lambda} = \frac{\rho}{\lambda(1-\rho)}
\]  

(3.25)

Using \( \rho = \lambda/\mu \), this becomes

\[
T = \frac{1}{\mu - \lambda}
\]

(3.26)

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 (Ex. 1 in section 3.2). 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 7: 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 as \( K \) times smaller average delay per packet. This result is quite general, even applying to networks of queues. What is happening, as illustrated in Fig. 3.7, 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 8: 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 to be 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

\[ T = \frac{1}{\mu - \lambda} \]
Sec. 3.3  The $M/M/1$ Queueing System

Figure 3.7  Increasing the arrival rate and the service rate by the same factor (see Ex. 7). (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 two. The average number in the system is the same as before, but the average delay is reduced by a factor of two since customers are moving twice as fast.
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

\[
T = \frac{m}{\mu - \lambda}
\]

i.e., \( m \) 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 that must contend with many low duty cycle traffic streams are 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 average delay per packet will decrease, but the average waiting time in queue will become positive. For example in telephony each traffic stream is a voice conversation that is regular in the above sense, and 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. In some systems, the times of customer arrivals are in some sense nontypical, so that the steady-state occupancy probabilities upon arrival

\[
a_n = \lim_{t \to \infty} P\{N(t) = n \mid \text{an arrival occurred just after time } t\}\]  \hspace{1cm} (3.27)

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

\[
p_n = \lim_{t \to \infty} P\{N(t) = n\}\]  \hspace{1cm} (3.28)

It turns out, however, that for the \( M/M/1 \) system, we have

\[
p_n = a_n, \; \; n = 0, 1, \ldots\]  \hspace{1cm} (3.29)

Indeed this equality 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 interval \( \delta > 0 \), 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 very reasonable for packet transmission systems. In particular, the assumption holds if the arrival process is Poisson and interarrival times and service times are independent.

The basic reason why \( a_n = p_n \) is that the events \( \{N(t) = n\} \) and \( \{\text{an arrival occurred just after } t\} \) are independent under our hypothesis. As a result, the conditional probability in Eq. (3.27) equals the unconditional probability in Eq. (3.28). For a more formal proof, let \( A(t, t + \delta) \) be the event that an arrival occurs in the interval \( (t, t + \delta) \). Let

\[
p_n(t) = P\{N(t) = n\} \\
a_n(t) = P\{N(t) = n | \text{an arrival occurred just after time } t\}
\]

We have, using Bayes' rule,

\[
a_n(t) = \lim_{\delta \to 0} P\{N(t) = n | A(t, t + \delta)\} \\
= \lim_{\delta \to 0} \frac{P\{N(t) = n, A(t, t + \delta)\}}{P\{A(t, t + \delta)\}} \\
= \lim_{\delta \to 0} \frac{P\{A(t, t + \delta) | N(t) = n\} P\{N(t) = n\}}{P\{A(t, t + \delta)\}} \\
= \frac{P\{A(t, t + \delta) | N(t) = n\}}{P\{A(t, t + \delta)\}}
\]

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

\[
P\{A(t, t + \delta) | N(t) = n\} = P\{A(t, t + \delta)\}
\]

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 Eq. (3.29).

Thus, we have shown that the probability of an arrival finding \( n \) customers in the system equals the (unconditional) probability of \( n \) in the system. This is true at every time instant as well as in steady state regardless of the service time distribution. We can summarize this by saying that when the arrival process is Poisson, an arriving customer finds the system in a "typical" state.

As an example of what can happen if the arrival process is not Poisson, suppose that interarrival times are independent and uniformly distributed between two and four seconds, while customer service times are all equal to one second. 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 \( 1/3 \).
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 $n^{th}$ packet equals one half the interarrival time between packets $n$ and $n + 1$. A packet upon arrival finds the system empty. However, the average number in the system, as seen by an outside observer looking at the system 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, i.e., the probabilities

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

The corresponding steady-state values are denoted

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

It turns out that

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

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 one). 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 proportion of transitions from $n$ to $n + 1$ out of transitions from any $k$ to $k + 1$ equals the proportion of transitions from $n + 1$ to $n$ out of transitions from any $k + 1$ to $k$ which implies $d_n = a_n$. 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 a random time.

### 3.4 The $M/M/m$, $M/M/\infty$, and $M/M/m/m$ Systems

We consider now a number of queueing systems that are similar to $M/M/1$ in that the arrival process is Poisson, the service times are independent, exponentially distributed, and independent of the interarrival times. Because of these assumptions, these systems can be modelled with continuous- or discrete-time Markov chains.
Figure 3.8 Discrete-time Markov chain for the $M/M/m$ system.

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. 3.8.

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

$$
\lambda p_{n-1} = n \mu p_n, \quad n \leq m
$$

$$
\lambda p_{n-1} = m \mu p_n, \quad n > m
$$

From these equations, we obtain

$$
p_n = \begin{cases} 
  p_0 \frac{(m \rho)^n}{n!}, & n \leq m \\
  p_0 \frac{m^n \rho^n}{m!}, & n > m 
\end{cases} \quad (3.33)
$$

where $\rho$ is given by

$$
\rho = \frac{\lambda}{m \mu} < 1 \quad (3.34)
$$

We can calculate $p_0$ using Eq. (3.33) and the condition $\sum_{n=0}^{\infty} p_n = 1$. We obtain

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

and, finally,

$$
p_0 = \left[ \sum_{n=0}^{m-1} \frac{(m \rho)^n}{n!} + \frac{(m \rho)^m}{m!(1 - \rho)} \right]^{-1} \quad (3.35)
$$
The probability that an arrival will find all servers busy and will be forced to wait in queue is

\[
P\{\text{Queueing}\} = \sum_{n=m}^{\infty} p_n
\]

\[
= \sum_{n=m}^{\infty} \frac{p_0 m^n \rho^n}{m!} = \frac{p_0 (m \rho)^m}{m!} \sum_{n=m}^{\infty} \rho^{n-m}
\]

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 equation is known as the **Erlang C formula** and is in wide use in telephony. (Denmark's A. K. Erlang is viewed as the foremost pioneer of queueing theory.)

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 Eq. (3.33), we obtain

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

Using Eq. (3.36) 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 Eq. (3.37), we obtain the average time \(W\) a customer has to wait in queue:

\[
W = \frac{N_Q}{\lambda} = \frac{\rho P_Q}{\lambda(1 - \rho)} \tag{3.38}
\]
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 finally obtain

\[ T = \frac{1}{\mu} + W = \frac{1}{\mu} + \frac{P_Q}{m\mu - \lambda} \] (3.39)

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 obtain

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

Example 9: Using one vs. using multiple channels in statistical multiplexing

Consider a communication link serving \( m \) independent Poisson traffic streams with rate \( \lambda/m \) each. 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 Eq. (3.39)

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

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 \( P_Q \approx 0 \), \( \hat{P}_Q \approx 0 \) and

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

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 one.

### 3.4.2 M/M/∞: Infinite-Server Case

In the limiting case where $m = \infty$ in the $M/M/m$ system, we obtain from Fig. 3.8
\[
\lambda p_{n-1} = n \mu p_n, \quad n = 1, 2, \ldots
\]
so
\[
p_n = p_0 \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}, \quad n = 1, 2, \ldots
\]

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

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 by simply 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 Prob. 3.37).
Example 10: 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. As analytical substantiation, let us 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

\[
N = E\{n\} = \lambda/\mu
\]

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

Suppose the \( i \)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

\[
F = E\{f\} = (\lambda/\mu)\Gamma
\]

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 (Prob. 3.21). The result is

\[
\sigma_f = (\lambda/\mu)^{1/2} s_\gamma
\]

Therefore, we have

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

Suppose now that the average bit rate \( \Gamma \) of a session is small relative to the total \( F \), i.e., 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 Eq. (3.40) 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

This system 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—a model that is in wide use in telephony. (The last m in the M/M/m/m notation indicates the limit on the number of customers in the system.) In data networks, it can be used as a model where arrivals correspond to requests for virtual circuit connections between two nodes and the number of virtual circuits allowed is m. The average service time $1/\mu$ is then the average duration of a virtual circuit conversation.

The corresponding state transition diagram is shown in Fig. 3.9. We have

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

so

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

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. It can be shown to hold even if the service time probability distribution is arbitrary, i.e., for an M/G/m/m system (see [Ros83], p. 170).

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