EE 683 at NJIT

Computer Network Design & Analysis

LECTURE 9: Queueing theory

taught by

Dr. Constantine N. Manikopoulos
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 inter-arrival times at the first queue. Then the first queue is $M/M/1$. The second queue, however, cannot be modeled as $M/M/l$. 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.
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.

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 3.26  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

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

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

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

![Diagram](image)

**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 \( M/M/l \) 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 \textbf{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

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

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

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 $M/M/l$ 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

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

(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

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

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

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 $X_0, X_{n+1}, ...$ having transition probabilities $P_{ij}$ and stationary distribution $\{p_j | j \geq 0\}$ with $p_j > 0$ for all $j$. Suppose that the chain is in steady-state, that is,

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

(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 $X_n, X_{n-1}, ...$. This sequence is itself a Markov chain, as seen by the following calculation:
\[ P\{X_m = j \mid X_{m+1} = i, X_{m+2} = i_2, \ldots, X_{m+k} = i_k \} = \frac{P\{X_m = j, X_{m+1} = i, X_{m+2} = i_2, \ldots, X_{m+k} = i_k \}}{P\{X_{m+1} = i, X_{m+2} = i_2, \ldots, X_{m+k} = i_k \}} \]

\[ = \frac{P\{X_m = j, X_{m+1} = i\} P\{X_{m+2} = i_2, \ldots, X_{m+k} = i_k \mid X_m = j, X_{m+1} = i\}}{P\{X_{m+1} = i\} P\{X_{m+2} = i_2, \ldots, X_{m+k} = i_k \mid X_{m+1} = i\}} \]

\[ = \frac{P\{X_m = j, X_{m+1} = i\}}{P\{X_{m+1} = i\}} \]

\[ = \frac{P\{X_m = j\} P\{X_{m+1} = i \mid X_m = j\}}{P\{X_{m+1} = i\}} \]

\[ = \frac{p_j P_{ji}}{p_i} \]

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

\[ P_{ij}^* = P\{X_m = j \mid X_{m+1} = i\} = \frac{p_j P_{ji}}{p_i}, \quad i, j \geq 0 \quad (3.107) \]

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*. 

\[ \mu \]
We list some properties of the reversed chain:

1. The reversed chain is irreducible, aperiodic, and has the same distribution as the forward chain. [This property can be shown either by reasoning using the definition of the reversed chain, or by verifying $p_j = \sum_{i=0}^{\infty} p_i P_{ij}^*$ 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 reverse chains are equal. Note that in view of this equality, the transition probabilities of the reversed chain $P_{ij}^* = p_j P_{ji} / p_i$ 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 ($p_j P_{ji}$) 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

$$P_{ij}^* = \frac{p_j P_{ji}}{p_i}, \quad i, j \geq 0$$

(3.108)

form a transition probability matrix, that is,

$$\sum_{j=0}^{\infty} P_{ij}^* = 1, \quad i = 0, 1, \ldots$$

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:

\[ p_i P_{ij} = p_j P_{ji}, \quad i, j \geq 0 \]

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 discrete-time chain, and passing back to the continuous chain by taking the limit as $\delta \to 0$.

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.
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

$$q_{ij}^* = \frac{p_j q_{ji}}{p_i}, \quad i, j \geq 0 \quad (3.109)$$

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

$$q_{ij}^* = \frac{p_j q_{ji}}{p_i}, \quad i, j \geq 0 \quad (3.110)$$

satisfy for all $i \geq 0$

$$\sum_{j=0}^{\infty} q_{ij} = \sum_{j=0}^{\infty} q_{ij}^* \quad (3.111)$$

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 \underline{time} reversible if and only if its stationary distribution and transition rates satisfy the detailed balanced equations

\[ p_i q_{ij} = p_j q_{ji}, \quad i, j \geq 0 \]

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 $[0, T]$. 

23
Departures prior to $t$ in the forward process

Arrivals after $t$ in the reversed process

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 $M/M/l$ 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,

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

To prove this we first note that queue 1 is an $M/M/l$ 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,

$$P\{n \text{ at queue 1}\} = \rho_1^n (1 - \rho_1)$$  \hspace{1cm} (3.113)

$$P\{m \text{ at queue 2}\} = \rho_2^m (1 - \rho_2)$$
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. (3.13) 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.
\[ \lambda_j = r_j + \sum_{i=1}^{K} \lambda_i P_{ij}, \quad j = 1, \ldots, K \]  

(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 \( j \)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

\[ \rho_j = \frac{\lambda_j}{\mu_j}, \quad j = 1, \ldots, K \]  

(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 $P_{00} = I$, $P_{10} = 1 - \sum_{j=1}^{K} P_{ij}$ 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 $i^{th}$ row elements of the matrix $P^m$ ($P$ to the $m^{th}$ 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 (3.1 14) 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

\[
n = (n_1, n_2, \ldots, n_K)
\]

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 \( n_1, \ldots, n_K \geq 0, \)

\[
P(n) = P_1(n_1)P_2(n_2) \cdots P_K(n_K)
\]  

(3.116)

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

\[
P_j(n_j) = \rho_j^{n_j}(1 - \rho_j), \quad n_j \geq 0
\]  

(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).
not time reversible here. Nonetheless, the use of the reversed process is both analytically
covenient 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

\[
q_{nn'}^* = \frac{P(n')q_{n'n}}{P(n)}
\]  

(3.118)

satisfy, for all \( n \), the total rate equation

\[
\sum_m q_{nm} = \sum_m q_{nm}^*
\]  

(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

\[
q_{nn(j^+)} = r_j
\]  

(3.120)

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

(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 \( P(n(j^+)) = \rho_j P(n) = \lambda_j P(n)/\mu_j \) [cf. Eqs. (3.115)–(3.117)], we obtain

\[
q_{nn(j^+)}^* = \lambda_j \left(1 - \sum_i P_{ji}\right)
\]  

(3.122)

\[
q_{nn(j^-)}^* = \frac{\mu_j r_j}{\lambda_j}
\]  

(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 \( P(n(i^+, j^-)) = \rho_i P(n)/\rho_j = \lambda_i \mu_j P(n)/(\lambda_j \mu_i) \), 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

\[
\sum_m q_{nm} = \sum_m q^*_{nm} \text{.}
\]

We have for the forward system, using Eqs. (3.120), (3.121), and (3.124),

\[
\sum_m q_{nm} = \sum_{j=1}^{K} q_{nn(j^+)} + \sum_{(j,i)|n_j>0} q_{nn(i^+, j^-)} + \sum_{j|n_j>0} q_{nn(j^-)}
\]

34
Similarly, using Eqs. (3.122), (3.123), (3.125), and (3.114), we obtain for the reversed system

\[
\sum_m Q^*_m = \sum_{j=1}^K q_{nn(j+)}^{*} + \sum_{(j,i)|n_j>0} q_{nn(i+,j-)}^{*} + \sum_{j|n_j>0} q_{nn(j-)}^{*}
\]

\[
= \sum_{j=1}^K \lambda_j \left(1 - \sum_{i=1}^K P_{ji} \right) + \sum_{(j,i)|n_j>0} \frac{\mu_j \lambda_i P_{ij}}{\lambda_j} + \sum_{j|n_j>0} \frac{\mu_j r_j}{\lambda_j}
\]

\[
= \sum_{j=1}^K \lambda_j \left(1 - \sum_{i=1}^K P_{ji} \right) + \sum_{j|n_j>0} \frac{\mu_j (r_j + \sum_{i=1}^K \lambda_i P_{ij})}{\lambda_j}
\]

\[
= \sum_{j=1}^K \lambda_j \left(1 - \sum_{i=1}^K P_{ji} \right) + \sum_{j|n_j>0} \mu_j
\]
By writing Eq. (3.114) as \( r_j = \lambda_j - \sum_{i=1}^{l} \lambda_i P_{ij} \) and adding over \( j = 1, \ldots, K \), we obtain

\[
\sum_{j=1}^{K} r_j = \sum_{j=1}^{K} \lambda_j \left( 1 - \sum_{i=1}^{K} P_{ji} \right)
\]

By combining the last three equations, we see that the total rate equation \( \sum_m q_{nm} = \sum_m q_{nm}^* \) is satisfied.

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)

\[
\lambda_1 = \lambda + \lambda_2, \quad \lambda_2 = p_2 \lambda_1
\]

[These are Eqs. (3.114) specialized to this example.] By solving for \( \lambda_1 \) and \( \lambda_2 \) we obtain

\[
\begin{align*}
\lambda_1 &= \frac{\lambda}{p_1}, \\
\lambda_2 &= \frac{\lambda p_2}{p_1}
\end{align*}
\]

(3.131)

Let

\[
\begin{align*}
\rho_1 &= \frac{\lambda_1}{\mu_1}, \\
\rho_2 &= \frac{\lambda_2}{\mu_2}
\end{align*}
\]

(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 total number in the system is

\[ \frac{2d - 1}{2d} + \frac{1d - 1}{1d} = \pi_N + \lambda N = N \]

utilization factor \( \rho \), that is,

The average number of jobs in the \( \text{in} \) queue is the same as for an \( M/M/1 \) system with.
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
\[
T = \frac{\lambda_1/\mu_1}{\lambda(1 - \lambda_1/\mu_1)} + \frac{\lambda_2/\mu_2}{\lambda(1 - \lambda_2/\mu_2)} = \frac{(1 - \lambda/\mu_1)}{(1 - \lambda/\mu_1) + \lambda p_2/(\mu_2 p_1)}
\]
\[
= \frac{S_1}{1 - \lambda S_1} + \frac{S_2}{1 - \lambda S_2}
\]  
(3.133)
where
\[
S_1 = \frac{1}{\mu_1 p_1}, \quad S_2 = \frac{p_2}{\mu_2 p_1}
\]  
(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 or 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 3.31. Despite this fact, the total arrival process at each queue need not be Poisson. As an example (see Fig. 3.39, 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.
Customer exits the system with probability $1 - p$.

Customer returns to the queue with high probability $p$ (successive services of the same customer are assumed independent).

**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 $A$ 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 $f_A(t)$ the content of the delay line at time $t$. The interpretation here is that $f_A(t)$ is a function of time that specifies the customer output of the delay line in the subsequent $A$ interval $(t, t+A]$. 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,

$$P\{n(0) = n\} = \rho^n (1 - \rho)$$  \hspace{1cm} (3.135)

where $\rho = \lambda / (\mu(1-p))$ is the utilization factor. Suppose also that $f_A(0)$ is a portion of a Poisson arrival process with rate $\lambda$. The customers in $f_A(0)$ have service times that are independent, exponentially distributed with parameter $\mu$. We assume that $n(0)$ and $f_A(0)$ are independent. Then, the input to the queue over the interval $[0, A)$ will

**Figure 3.36** Heuristic explanation of Jackson’s Theorem. Consider the introduction of an arbitrarily small positive delay $A$ 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 $A$ 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 \([0, A]\) like an \(M/M/1\) queue in equilibrium. Therefore, \(n(A)\) 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 \([0, A)\), or, equivalently, of \(f_\Delta(\Delta)\) - the delay line content at time \(A\). 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 \(\dot{f}(0)\) which had the equilibrium distribution of an \(M/M/1\) queue and the statistics of a Poisson process, respectively, and \(A\) seconds later we obtained corresponding quantities \(n(A)\) and \(\dot{f}(A)\) with the same properties. Using the same reasoning, we can show that for all \(t\) which are multiples of \(A\), \(n(t)\) and \(\dot{f}(t)\) 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 \(A\), 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 \(A\) portions of the feedback arrival stream are correlated since, with probability \(p\), a departing customer from the queue appears as an arrival \(A\) 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 $j^{th}$ queue is exponentially distributed with mean $1/\mu_j(m)$, where $m$ is the number of customers in the $j^{th}$ 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

$$\rho_j(m) = \frac{\lambda_j}{\mu_j(m)}, \quad j = 1, \ldots, K, \quad m = 1, 2, \ldots \quad (3.136)$$

where $\lambda_j$ is the total arrival rate at the $j^{th}$ queue determined by Eq. (3.114). Let us also define

$$\hat{P}_j(n_j) = \begin{cases} 1, & \text{if } n_j = 0 \\ \rho_j(1)\rho_j(2)\cdots\rho_j(n_j), & \text{if } n_j > 0 \end{cases} \quad (3.137)$$
Jackson’s Theorem for State-Dependent Service Rates. We have for all states $n = (n_1, \ldots, n_K)$

$$P(n) = \frac{\hat{P}_1(n_1) \cdots \hat{P}_K(n_K)}{G}$$

(3.138)

assuming that $0 < G < \infty$, where the normalizing constant $G$ is given by

$$G = \sum_{n_1=0}^{\infty} \cdots \sum_{n_K=0}^{\infty} \hat{P}_1(n_1) \cdots \hat{P}_K(n_K)$$

(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

\[
\lambda_j(c) = r_j(c) + \sum_{i=1}^{K} \lambda_i(c) P_{ij}(c), \quad j = 1, \ldots, K, \ c = 1, 2, \ldots, C
\]  

(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** \( j^{th} \) **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 \( i^{th} \) 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

\[ \hat{\rho}_j(c, m) = \frac{\lambda_j(c)}{\mu_j(m)}, \quad j = 1, \ldots, K, \quad c = 1, 2, \ldots, C \quad (3.141) \]

\[ \hat{\rho}_j(z_j) = \begin{cases} 1, & \text{if } n_j = 0 \\ \hat{\rho}_j(c_1, 1)\hat{\rho}_j(c_2, 2) \cdots \hat{\rho}_j(c_{n_j}, n_j), & \text{if } n_j > 0 \end{cases} \quad (3.142) \]

\[ G = \sum_{(z_1, \ldots, z_K)} \prod_{j=1}^K \hat{\rho}_j(z_j) \quad (3.143) \]

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 $z = (z_1, z_2, ..., z_K)$ is given by

$$\hat{P}(z) = \frac{\hat{P}_1(z_1) \cdots \hat{P}_K(z_K)}{G} \quad (3.144)$$

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 straightforward 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

$$P(n) = \prod_{j=1}^{K} \rho_j^{n_j} (1 - \rho_j) \quad (3.145)$$

where

$$\rho_j = \frac{\sum_{c=1}^{C} \lambda_j(c)}{\mu_j} \quad (3.146)$$

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

$$\lambda_{ij} = \sum_{\{c | (i, j) \text{ lies on the path } p_c\}} x_c$$

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. (3.149, 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
Let also $\mu_j(m)$ be the service rate at the $j$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

$$\lambda_j = \sum_{i=1}^{K} \lambda_i P_{ij}, \quad j = 1, \ldots, K$$

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 $\lambda_j = a P_{1j} + \sum_{i=2}^{K} \lambda_i P_{ij}, j = 2, \ldots, K$. 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.
\[ \lambda_j = \alpha \bar{\lambda}_j, \quad j = 1, \ldots, K \]
where \( \alpha \) is a scalar and \( \bar{\lambda}_j, j = 1, \ldots, K \) is a particular solution with \( \bar{\lambda}_j > 0 \) for all \( j \). Thus the true arrival rates are given by

\[ \lambda_j(M) = \alpha(M) \bar{\lambda}_j, \quad j = 1, \ldots, K \quad (3.148) \]
where \( \alpha(M) \) is the constant of proportionality corresponding to \( M \). Note that while \( \bar{\lambda}_j \) 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

\[ P_j(m) = \begin{cases} 1, & \text{if } n_j = 0 \\ \rho_j(1)P_j(2) \cdots P_j(n_j), & \text{if } n_j > 0 \end{cases} \]

Denote

\[ \hat{P}_j(n_j) = \begin{cases} 1, & \text{if } n_j = 0 \\ \rho_j(1)\rho_j(2) \cdots \rho_j(n_j), & \text{if } n_j > 0 \end{cases} \]

\[ G(M) = \sum_{\{(n_1, \ldots, n_K)|n_1 + \cdots + n_K = M\}} \hat{P}_1(n_1) \cdots \hat{P}_K(n_K) \quad (3.151) \]

We have:
Jackson’s Theorem for Closed Networks. Under the preceding assumptions, we have for all states \( n = (n_1, \ldots, n_K) \) with \( n_1 + \cdots + n_K = M \)

\[
P(n) = \frac{\hat{P}_1(n_1) \cdots \hat{P}_K(n_K)}{G(M)}
\]  

(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

\[
\begin{align*}
n &= (n_1, \ldots, n_i, \ldots, n_j, \ldots, n_K) \\
n' &= (n_1, \ldots, n_{i-1}, n_i + 1, n_{i+1}, \ldots, n_{j-1}, n_j - 1, n_{j+1}, \ldots, n_K)
\end{align*}
\]

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

\[
q_{nn'}^* = \frac{P(n')q_{n'n}}{P(n)}
\]

(3.153)
satisfy, for all states \( n \), the total rate equation

\[
\sum_m q_{nm} = \sum_m q_{nm}^*
\]  

(3.154)

Indeed, let us assume for the purpose of the proof that the particular solutions \( \bar{\lambda}_j \) are taken to be equal to the true arrival rates \( \lambda_j(M) \), and for convenience let us denote both \( \bar{\lambda}_j \) and \( \lambda_j(M) \) as \( \lambda_j \). Then we have [cf. Eqs. (3.124) and (3.125)]

\[
q_{nn'} = \mu_j(n_j)P_{ji}, \quad q_{nn'}^* = \frac{\mu_j(n_j)\lambda_iP_{ij}}{\lambda_j}
\]

and the total rate equation (3.154) is written as

\[
\sum_{\{j,i|n_j > 0\}} \mu_j(n_j)P_{ji} = \sum_{\{j,i|n_j > 0\}} \frac{\mu_j(n_j)\lambda_iP_{ij}}{\lambda_j}
\]

We have

\[
\sum_{\{j,i|n_j > 0\}} \mu_j(n_j)P_{ji} = \sum_{i=1}^{K} \sum_{\{j|n_j > 0\}} \mu_j(n_j)P_{ji} = \sum_{\{j|n_j > 0\}} \mu_j(n_j) \sum_{i=1}^{K} P_{ji}
\]

\[
= \sum_{\{j|n_j > 0\}} \mu_j(n_j)
\]  

(3.155)
We also have
\[
\sum_{\{j|i,j\}|n_j > 0}\frac{\mu_j(n_j)\lambda_i P_{ij}}{\lambda_j} = \sum_{i=1}^{K} \sum_{\{j|n_j > 0\}} \frac{\mu_j(n_j)\lambda_i P_{ij}}{\lambda_j} = \sum_{\{j|n_j > 0\}} \frac{\mu_j(n_j)}{\lambda_j} \sum_{i=1}^{K} \lambda_i P_{ij}
\]
\[
= \sum_{\{j|n_j > 0\}} \mu_j(n_j)
\]
(3.156)

From Eqs. (3.155) and (3.156) we see that the total rate equation (3.154) holds. \textbf{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.](image-url)
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

$$\bar{\lambda}_1 = \mu_1, \quad \bar{\lambda}_2 = p_2 \mu_1$$

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

$$\rho_1 = 1, \quad \rho_2 = \frac{p_2 \mu_1}{\mu_2}$$

and the steady-state distribution of the system is given by [cf. Eqs. (3.149) to (3.151)]

$$P(M - n, n) = \frac{\rho_2^n}{G(M)}, \quad n = 0, 1, \ldots, M$$

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 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 $M/M/1$ queue and an $M/M/1$ queue, respectively [see Fig. 3.38(b)]. Let $\lambda = 1$ be the particular solution of the arrival rate equation for the system [cf. Eq. (3.147)]. We have

$$\rho_1 = \frac{R}{P}, \quad \rho_2 = 1$$

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 $A(N)$ as a function of the number of terminals compared with the upper and lower bounds derived in Example 3.7.

$$\frac{N}{R + NP} \leq X(N) \leq \min \left\{ \frac{T}{P}, \frac{N}{R + P} \right\}$$
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 $X(N)$ is the system throughput. Therefore, we have

$$X(N) = U(N)/P, \text{ or}$$

$$X(N) = \frac{G(N - 1)}{P G(N)}$$

This expression for $\lambda(N)$ is shown in Fig. 3.38(c) and is contrasted with the upper and lower bounds

$$\frac{N}{R + NP} \leq \lambda(N) \leq \min \left\{ \frac{1}{P}, \frac{N}{R + P} \right\} \quad (3.157)$$

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

$$\lambda_j(M) = \frac{N_j(M)}{T_j(M)} \quad (3.158)$$

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 $m$.]

[61]
Let us assume for simplicity that the service rate at the \( j \)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

\[
T_j(0) = N_j(0) = 0, \quad j = 1, \ldots, K
\]

(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)

\[
T_j(s) = \frac{1}{\mu_j} (1 + N_j(s-1)), \quad j = 1, \ldots, K, \quad s = 1, \ldots, M
\]

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)

\[
N_j(s) = s \frac{\bar{\lambda}_j T_j(s)}{\sum_{i=1}^K \bar{\lambda}_i T_i(s)}, \quad j = 1, \ldots, K, \quad s = 1, \ldots, M
\]

where \( \bar{\lambda}_j, j = 1, \ldots, K \) 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 \( j, \lambda_j(s) = \alpha(s) \bar{\lambda}_j \) 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 \( j^{th} \) queue is the same as the steady-state distribution of the \( j^{th} \) 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 \( j^{th} \) 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

\[
\alpha_{ij}(n) = P \{ x(t) = n \mid \text{a customer moved from queue } i \text{ to queue } j \text{ just after time } t \}
\]

(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

$$
\alpha_{ij}(n) = \frac{P\{x(t) = n, M_{ij}(t) \mid M_i(t)\}}{P\{M_{ij}(t) \mid M_i(t)\}}
$$

$$
= \frac{P\{x(t) = n \mid M_i(t)\} P\{M_{ij}(t) \mid x(t) = n, M_i(t)\}}{P\{M_{ij}(t) \mid M_i(t)\}}
$$

$$
= \frac{P(n)P_{ij}}{\sum_{\{n'=(n'_1,\ldots,n'_K)\mid n'_i > 0\}} P(n')P_{ij}}
$$

and finally, using Eqs. (3.149) to (3.152) for the steady-state probabilities. $P(n)$,

$$
\alpha_{ij}(n) = \frac{\hat{P}_1(n_1) \cdots \hat{P}_K(n_K)}{\sum_{\{(n'_1,\ldots,n'_K)\mid n'_i + \cdots + n'_K = s, n'_i > 0\}} \hat{P}_1(n'_1) \cdots \hat{P}_K(n'_K)}
$$

(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

$$
\alpha_{ij}(n) = \frac{\hat{P}_1(n_1) \cdots \hat{P}_{i-1}(n_{i-1}) \hat{P}_i(n_i - 1) \hat{P}_{i+1}(n_{i+1}) \cdots \hat{P}_K(n_K)}{\sum_{\{(n'_1,\ldots,n'_K)\mid n'_i + \cdots + n'_K = s-1\}} \hat{P}_1(n'_1) \cdots \hat{P}_K(n'_K)}
$$

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

\[ N_j(M - 1) = f_j(N_j(M)) \]

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

\[ T_j(M) = \frac{1}{\mu_j} \left( 1 + f_j(N_j(M)) \right), \quad j = 1, \ldots, K \]

\[ N_j(M) = M \frac{\lambda_j T_j(M)}{\sum_{i=1}^{K} \lambda_i T_i(M)}, \quad j = 1, \ldots, K \]

which can be solved by iterative methods to yield approximate values for \( T_j(M) \) and \( N_j(M) \).
2-10 Consider a finite M/M/1 queue capable of accommodating N packets (customers). Calculate the values of N required for the following situations:

1. \( \rho = 0.5, \ P_a = 10^{-5}, 10^{-6} \)
2. \( \rho = 0.8, \ P_a = 10^{-5}, 10^{-6} \)

Compare the results obtained.

2-11 The probability \( P_n \) that an infinite M/M/1 queue is in state \( n \) is given by \( \rho_n = (1 - \rho)\rho^n, \ \rho = \lambda/\mu. \)
a. Show that the average queue occupancy is given by

\[
E(n) = \sum n\rho_n = \rho/(1 - \rho)
\]

b. Plot \( P_n \) as a function of \( n \) for \( \rho = 0.8 \).
c. Plot \( E(n) \) versus \( \rho \) and compare with Fig. 2-17.

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

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

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

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

a. Find the blocking probability, \( P_b \) of this system.
b. Find the average number, \( E(n) \), in the system.
c. Find the normalized throughput \( \gamma/\mu \), with \( \gamma \) the average throughput, in packets/sec.
d. Find the average delay \( E(T) \) through the system, in units of \( 1/\mu \). (Alternatively, find \( E(T)/1/\mu \).)

2-22 A data concentrator has 40 terminals connected to it. Each terminal inputs packets with an average length of 680 bits. Forty bits of control information are added to each packet before transmission over an outgoing link with capacity \( C = 7200 \) bps.

\[
\begin{align*}
&\text{Twenty of the terminals input 1 packet/10 sec each, on the average.} \\
&\text{Ten of the terminals input 1 packet/5 sec each, on the average.} \\
&\text{Ten of the terminals input 1 packet/2.5 sec each, on the average.}
\end{align*}
\]

The input statistics are Poisson.

a. The data units transmitted (called frames) are exponentially distributed in length. Find (1) the average wait time on queue, not including service time and (2) the average number of packets in the concentrator, including the one in service.
b. Repeat if the packets are all of constant length.
c. Repeat if the second moment of the frame length is \( E(T^2) = 3(\mu_2)/(\mu_1)^2; \mu_4 \) is the average frame length.
2-10 Consider a finite M/M/l queue capable of accommodating N packets. Calculate the values of N required for the following situations:

1. $\rho = 0.5, \ P_0 = 10^{-3}, 10^{-6}$
2. $\rho = 0.8, \ P_0 = 10^{-3}, 10^{-6}$

Compare the results obtained.

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

a. Show that the average queue occupancy is given by

$$E(n) = \sum n p_n = \rho/(1 - \rho)$$

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

c. Plot $E(n)$ versus $\rho$ and compare with Fig. 2-17.

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

1. Ten terminals are connected to the statistical multiplexer. Each generates, on average, one 960-bit packet, assumed to be distributed exponentially, every 8 sec. A 2400-bits/sec outgoing line is used.

2. Repeat if each terminal now generates a packet every 5 sec, on the average.

3. Repeat 1. above if 16 terminals are connected.

4. Forty terminals are now connected and a 4600-bits/sec output line is used. Repeat 1. and 2. in this case. Now increase the average packet length to 1600 bits. What is the average buffer occupancy if a packet is generated every 8 sec at each terminal? What would happen if each terminal were allowed to increase its packet generation rate to 1 per 5 sec on the average? (Hint: It might now be appropriate to use a finite M/M/l model with your own choice of buffer size.)

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

2-21 A queueing system has two outgoing lines, used randomly by packets requiring service. Each transmits at a rate of $\mu$ packets/sec. When both lines are transmitting (serving) packets, packets are blocked from entering--i.e., there is no buffering in this system. Packets are exponentially distributed in length; arrivals are Poisson, with average rate $\lambda, \ P_0 = \lambda/\mu = 1$.

a. Find the blocking probability, $P_b$, of this system.

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

c. Find the normalized throughput $y/\mu$, with $y$ the average throughput in packets/sec.

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

2-22 A data concentrator has 40 terminals connected to it. Each terminal inputs packets with an average length of 680 bits. Forty bits of control information are added to each packet before transmission over an outgoing link with capacity $C = 7200$ bps.

20 of the terminals input 1 packet/10 sec each, on the average.

10 of the terminals input 1 packet/5 sec each, on the average.

Ten of the terminals input 1 packet/2.5 sec each, on the average.

The input statistics are Poisson.

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

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

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