9 Queueing Systems

9.1 Queueing Processes

A queueing system consists of "customers" arriving at random times to some facility where they receive service of some kind and then depart. We use "customer" as a generic term. It may refer, e.g., to *bona fide* customers demanding service at a counter, to ships entering a port, to batches of data flowing into a computer subsystem, to broken machines awaiting repair, and so on. Queueing systems are classified according to

1. *The input process*, the probability distribution of the pattern of arrivals of customers in time;
2. *The service distribution*, the probability distribution of the random time to serve a customer (or group of customers in the case of batch service); and
3. *The queue discipline*, the number of servers and the order of customer service.

While a variety of input processes may arise in practice, two simple and frequently occurring types are mathematically tractable and give insights into more complex cases. First is the scheduled input, where customers arrive at fixed times $T, 2T, 3T, \ldots$. The second most common model is the "completely random" arrival process, where the times of customer arrivals form a Poisson process. Understanding the axiomatic development of the Poisson process in V may help one to evaluate the validity of the Poisson assumption in any given application. Many theoretical results are available when the times of customer arrivals form a renewal process. Exponentially distributed interarrival times, then, correspond to a Poisson process of arrivals as a special case.

We will always assume that the durations of service for individual customers are independent and identically distributed nonnegative random variables and are independent of the arrival process. The situation in which all service times are the same fixed duration D, is, then, a special case.

The most common queue discipline is *first come, first served*, where customers are served in the same order in which they arrive. All of the models that we consider in this chapter are of this type.

Queueing models aid the design process by predicting system performance. For example, a queueing model might be used to evaluate the costs and benefits of adding a server to an existing system. The models enable us to calculate system performance measures in terms of more basic quantities. Some important measures of system behavior are

1. *The probability distribution of the number of customers in the system.* Not only do customers in the system often incur costs, but in many systems, physical space for waiting customers

An Introduction to Stochastic Modeling. DOI: 10.1016/B978-0-12-381416-6.00009-5

must be planned for and provided. Large numbers of waiting customers can also adversely affect the input process by turning potential new customers away. (See Section 9.4.1 on queueing with balking.)

2. *The utilization of the server*(*s*). Idle servers may incur costs without contributing to system performance.
3. *System throughput*. The long run number of customers passing through the system is a direct measure of system performance.
4. *Customer waiting time*. Long waits for service are annoying in the simplest queueing situations and directly associated with major costs in many large systems such as those describing ships waiting to unload at a port facility or patients awaiting emergency care at a hospital.

9.1.1 The Queueing Formula $L = \lambda W$

Consider a queueing system that has been operating sufficiently long to have reached an appropriate steady state, or a position of statistical equilibrium. Let

$L =$ the average number of customers in the system;

$\lambda =$ the rate of arrival of customers to the system; and

$W =$ the average time spent by a customer in the system.

The equation $L = \lambda W$ is valid under great generality for such systems and is of basic importance in the theory of queues, since it directly relates two of our most important measures of system performance, the mean queue size and the mean customer waiting time in the steady state, i.e., mean queue size and mean customer waiting time evaluated with respect to a limiting or stationary distribution for the process.

The validity of $L = \lambda W$ does not rest on the details of any particular model, but depends only upon long run mass flow balance relations. To sketch this reasoning, consider a time T sufficiently long so that statistical fluctuations have averaged out. Then, the total number of customers to have entered the system is λT, the total number to have departed is $\lambda(T - W)$, and the net number remaining in the system L must be the difference

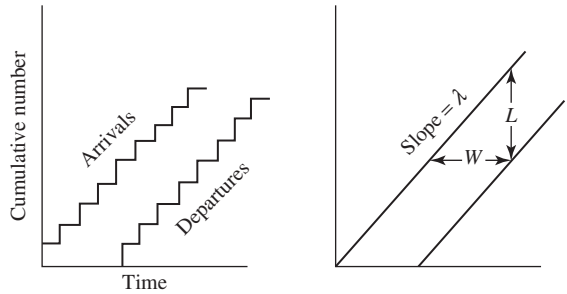
$$L = \lambda T - [\lambda(T - W)] = \lambda W.$$

Figure 9.1 depicts the relation $L = \lambda W$.

Of course, what we have done is by no means a proof, and indeed, we shall give no proof. We shall, however, provide several sample verifications of $L = \lambda W$ where L is the mean of the stationary distribution of customers in the system, W is the mean customer time in the system determined from the stationary distribution, and λ is the arrival rate in a Poisson arrival process.

Let L_0 be the average number of customers waiting in the system who are not yet being served, and let W_0 be the average waiting time in the system excluding service time. In parallel to $L = \lambda W$, we have the formula

$$L_0 = \lambda W_0 \tag{9.1}$$



(a) Random arrivals, departures (b) Smoothed values

Figure 9.1 The cumulative number of arrivals and departures in a queueing system. The smoothed values in (b) are meant to symbolize long run averages. The rate of arrivals per unit time is λ, the mean number in the system is L, and the mean time a customer spends in the system is W.

The total waiting time in the system is the sum of the waiting time before service and the service time. In terms of means, we have

$$W = W_0 + \text{mean service time}. \tag{9.2}$$

9.1.2 A Sampling of Queueing Models

In the remainder of this chapter, we will study a variety of queueing systems. A standard shorthand is used in much of the queueing literature for identifying simple queueing models. The shorthand assumes that the arrival times form a renewal process, and the format $A/B/c$ uses A to describe the interarrival distribution, B to specify the individual customer service time distribution, and c to indicate the number of servers. The common cases for the first two positions are $G = GI$ for a general or arbitrary distribution, M (memoryless) for the exponential distribution, E_k (Erlang) for the gamma distribution of order k, and D for a deterministic distribution, a schedule of arrivals or fixed service times.

Some examples discussed in the sequel are the following:

The M/M/1 queue Arrivals follow a Poisson process; service times are exponentially distributed; and there is a single server. The number $X(t)$ of customers in the system at time t forms a birth and death process. (See Section 9.2.)

The M/M/∞ queue There are Poisson arrivals and exponentially distributed service times. Any number of customers are processed simultaneously and independently. Often self-service situations may be described by this model. In the older literature, this was called the "telephone trunking problem."

The M/G/1 queue In this model, there are Poisson arrivals but arbitrarily distributed service times. The analysis proceeds with the help of an embedded Markov chain.

More elaborate variations will also be set forth. *Balking* is the refusal of new customers to enter the system if the waiting line is too long. More generally, in a

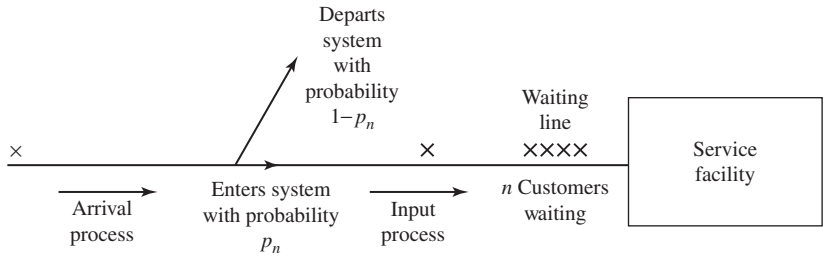


Figure 9.2 If n customers are waiting in a queueing system with balking, an arriving customer enters the system with probability p_n and does not enter with probability $1 - p_n$.

queueing system with balking, an arriving customer enters the system with a probability that depends on the size of the queue. Here it is important to distinguish between the *arrival process* and the *input process*, as shown in Figure 9.2. A special case is a queue with *overflow*, in which an arriving customer enters the queue if and only if there is at least one server free to begin service immediately.

In a *priority queue*, customers are allowed to be of different types. Both the service discipline and the service time distribution may vary with the customer type.

A *queueing network* is a collection of service facilities where the departures from some stations form the arrivals of others. The network is *closed* if the total number of customers is fixed, with these customers continuously circulating through the system. The machine repair model (see the example entitled "Repairman Models" in Chapter 6, Section 6.4) is an example of a closed queueing network. In an open queueing network, customers may arrive from, and depart to, places outside the network, as well as move from station to station. Queueing network models have found much recent application in the design of complex information processing systems.

Exercises

9.1.1 What design questions might be answered by modeling the following queueing systems?

The Customer	The Server
(a) Arriving airplanes	The runway
(b) Cars	A parking lot
(c) Broken TVs	Repairman
(d) Patients	Doctor
(e) Fires	Fire engine company

What might be reasonable assumptions concerning the arrival process, service distribution, and priority in these instances?

9.1.2 Consider a system, such as a barber shop, where the service required is essentially identical for each customer. Then, actual service times would tend to cluster near the mean service time. Argue that the exponential distribution would not be appropriate in this case. For what types of service situations might the exponential distribution be quite plausible?

9.1.3 Oil tankers arrive at an offloading facility according to a Poisson process whose rate is $\lambda = 2$ ships per day. Daily records show that there is an average of 3 ships unloading or waiting to unload at any instant in time. On average, what is the duration of time that a ship spends in port? Assume that a ship departs immediately after unloading.

Problem

9.1.1 Two dump trucks cycle between a gravel loader and a gravel unloader. Suppose that the travel times are insignificant relative to the load and unload times, which are exponentially distributed with parameters μ and λ, respectively. Model the system as a closed queueing network. Determine the long run gravel loads moved per unit time.

Hint: Refer to the example entitled "Repairman Models" in Section 6.4.

9.2 Poisson Arrivals, Exponential Service Times

The simplest and most extensively studied queueing models are those having a Poisson arrival process and exponentially distributed service times. In this case, the queue size forms a birth and death process (see Sections 6.3 and Sections 6.4 of Chapter 6), and the corresponding stationary distribution is readily found.

We let λ denote the intensity, or rate, of the Poisson arrival process and assume that the service time distribution is exponential with parameter μ. The corresponding density function is

$$g(x) = \mu e^{-\mu x} \quad \text{for } x > 0. \tag{9.3}$$

For the Poisson arrival process we have

$$\Pr\{\text{An arrival in } [t, t+h)\} = \lambda h + o(h) \tag{9.4}$$

and

$$\Pr\{\text{No arrivals in } [t, t+h)\} = 1 - \lambda h + o(h). \tag{9.5}$$

Similarly, the memoryless property of the exponential distribution as expressed by its constant hazard rate (see Chapter 1, Section 1.4.2) implies that

$$\Pr\{\text{A service is completed in } [t, t+h) | \text{Service in progress at time } t\} = \mu h + o(h), \tag{9.6}$$

and

$$\Pr\{\text{Service not completed in } [t, t+h) | \text{Service in progress at time } t\} = 1 - \mu h + o(h), \tag{9.7}$$

The service rate μ applies to a particular server. If k servers are simultaneously operating, the probability that one of them completes service in a time interval of duration h is $(k\mu)h + o(h)$, so that the system service rate is $k\mu$. The principle used here is the same as that used in deriving the infinitesimal parameters of the Yule process (Chapter 6, Section 6.1).

We let $X(t)$ denote the number of customers in the system at time t, counting the customers undergoing service as well as those awaiting service. The independence of arrivals in disjoint time intervals together with the memoryless property of the exponential service time distribution implies that $X(t)$ is a time homogeneous Markov chain, in particular, a birth and death process. (See Sections 6.3 and 6.4 of Chapter 6.)

9.2.1 The M/M/1 System

We consider first the case of a single server and let $X(t)$ denote the number of customers in the system at time t. An increase in $X(t)$ by one unit corresponds to a customer arrival, and in view of (9.4) and (9.7) and the postulated independence of service times and the arrival process, we have

$$\begin{aligned}\Pr\{X(t+h) = k+1 | X(t) = k\} &= [\lambda h + o(h)] \times [1 - \mu h + o(h)] \\ &= \lambda h + o(h) \quad \text{for } k = 0, 1, \ldots.\end{aligned}$$

Similarly, a decrease in $X(t)$ by one unit corresponds to a completion of service, whence

$$\Pr\{X(t+h) = k-1 | X(t) = k\} = \mu h + o(h) \quad \text{for } k = 1, 2, \ldots.$$

Then, $X(t)$ is a birth and death process with birth parameters

$$\lambda_k = \lambda \quad \text{for } k = 0, 1, 2, \ldots$$

and death parameters

$$\mu_k = \mu \quad \text{for } k = 1, 2, \ldots.$$

Of course, no completion of service is possible when the queue is empty. We, thus, specify $\mu_0 = 0$.

Let

$$\pi_k = \lim_{t\to\infty} \Pr\{X(t) = k\} \quad \text{for } k = 0, 1, \ldots$$

be the limiting, or equilibrium, distribution of queue length. Section 6.4 of Chapter 6 describes a straightforward procedure for determining the limiting distribution π_k from the birth and death parameters λ_k and μ_k. The technique is to first obtain intermediate quantities θ_j defined by

$$\theta_0 = 1 \quad \text{and} \quad \theta_j = \frac{\lambda_0\lambda_1\cdots\lambda_{j-1}}{\mu_1\mu_2\cdots\mu_j} \quad \text{for } j \geq 1, \tag{9.8}$$

and then

$$\pi_0 = \frac{1}{\Sigma_{J=0}^{\infty}\theta_j} \quad \text{and} \quad \pi_k = \theta_k\pi_0 = \frac{\theta_k}{\Sigma_{j=0}^{\infty}\theta_j} \quad \text{for } k \geq 1. \tag{9.9}$$

When $\Sigma_{j=0}^{\infty}\ \theta_j = \infty$, then $\lim_{t\to\infty} \Pr\{X(t) = k\} = 0$ for all k, and the queue length grows unboundedly in time.

For the $M/M/1$ queue at hand we readily compute $\theta_0 = 1$ and $\theta_j = (\lambda/\mu)^j$ for $j = 1, 2, \ldots$. Then,

$$\begin{aligned}\sum_{j=0}^{\infty}\pi_j = \sum_{j=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^j &= \frac{1}{(1-\lambda/\mu)} \quad \text{if } \lambda < \mu,\\ &= \infty \qquad\qquad \text{if } \lambda \geq \mu.\end{aligned}$$

Thus, no equilibrium distribution exists when the arrival rate λ is equal to or greater than the service rate μ. In this case, the queue length grows without bound.

When $\lambda < \mu$, a *bona fide* limiting distribution exists, given by

$$\pi_0 = \frac{1}{\Sigma\theta_j} = 1 - \frac{\lambda}{\mu} \tag{9.10}$$

and

$$\pi_k = \pi_0\theta_k = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^k \quad \text{for } k = 0, 1, \ldots. \tag{9.11}$$

The equilibrium distribution (9.11) gives us the answer to many questions involving the limiting behavior of the system. We recognize the form of (9.11) as that of a geometric distribution, and then reference to Chapter 1, Section 1.3.3 gives us the mean queue length in equilibrium to be

$$L = \frac{\lambda}{\mu - \lambda}. \tag{9.12}$$

The ratio $\rho = \lambda/\mu$ is called the *traffic intensity*,

$$\rho = \frac{\text{arrival rate}}{\text{system service rate}} = \frac{\lambda}{\mu}. \tag{9.13}$$

As the traffic intensity approaches one, the mean queue length $L = \rho/(1-\rho)$ becomes infinite. Again using (9.10), the probability of being served immediately upon arrival is

$$\pi_0 = 1 - \frac{\lambda}{\mu},$$

the probability, in the long run, of finding the server idle. The server utilization, or long run fraction of time that the server is busy, is $1 - \pi_0 = \lambda/\mu$.

We can also calculate the distribution of waiting time in the stationary case when $\lambda < \mu$. If an arriving customer finds n people in front of him, his total waiting time T, including his own service time, is the sum of the service times of himself and those ahead, all distributed exponentially with parameter μ, and since the service times are independent of the queue size, T has a gamma distribution of order $n+1$ with scale parameter μ,

$$\Pr\{T \le t | n \text{ ahead}\} = \int_0^t \frac{\mu^{n+1}\tau^n e^{-\mu\tau}}{\Gamma(n+1)} d\tau. \tag{9.14}$$

By the law of total probability, we have

$$\Pr\{T \le t\} = \sum_{n=0}^{\infty} \Pr\{T \le t | n \text{ ahead}\} \times \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right),$$

since $(\lambda/\mu)^n(1-\lambda/\mu)$ is the probability that in the stationary case a customer on arrival will find n ahead in line. Now, substituting from (9.14), we obtain

$$\begin{aligned}
\Pr\{T \le t\} &= \sum_{n=0}^{\infty} \int_0^t \frac{\mu^{n+1}\tau^n e^{-\mu\tau}}{\Gamma(n+1)} \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right) d\tau \\
&= \int_0^t \mu e^{-\mu\tau} \left(1 - \frac{\lambda}{\mu}\right) \sum_{n=0}^{\infty} \frac{\tau^n \lambda^n}{\Gamma(n+1)} d\tau \\
&= \int_0^t \left(1 - \frac{\lambda}{\mu}\right) \mu \exp\left\{-\tau\mu\left(1 - \frac{\lambda}{\mu}\right)\right\} d\tau = 1 - \exp[-t(\mu - \lambda)],
\end{aligned}$$

which is also an exponential distribution.

The mean of this exponential waiting time distribution is the reciprocal of the exponential parameter, or

$$W = \frac{1}{\mu - \lambda}. \tag{9.15}$$

Reference to (9.12) and (9.15) verifies the fundamental queueing formula $L = \lambda W$.

A queueing system alternates between durations when the servers are busy and durations when the system is empty and the servers are idle. An *idle period* begins the instant the last customer leaves, and endures until the arrival of the next customer. When the arrival process is Poisson of rate λ, then an idle period is exponentially distributed with mean

$$E[I_1] = \frac{1}{\lambda}.$$

A busy period is an uninterrupted duration in which the system is not empty. When arrivals to a queue follow a Poisson process, then the successive durations X_k from the commencement of the kth busy period to the start of the next busy period form a renewal process (see Figure 9.3). Each X_k is composed of a busy portion B_k and an idle portion I_k. Then, the renewal theorem (see "A Queueing Model" in Chapter 7, Section 7.5.3) applies to tell us that $p_0(t)$, the probability that the system is empty at time t, converges to

$$\lim_{t\to\infty} p_0(t) = \pi_0 = \frac{E[I_1]}{E[I_1] + E[B_1]}.$$

We substitute the known quantities $\pi_0 = 1 - \lambda/\mu$ and $E[I_1] = 1/\lambda$ to obtain

$$1 - \frac{\lambda}{\mu} = \frac{1/\lambda}{1/\lambda + E[B_1]},$$

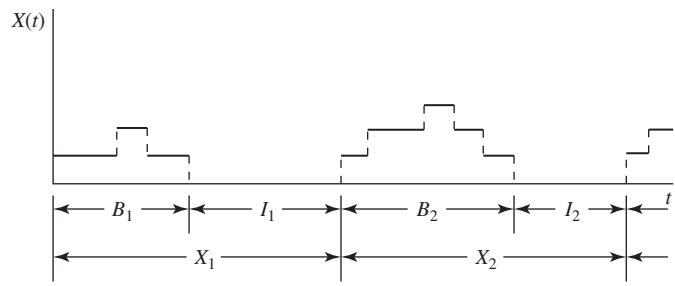


Figure 9.3 The busy periods B_k and idle periods I_k of a queueing system. When arrivals form a Poisson process, then $X_k = B_k + I_k$, $k = 1, 2, \ldots$, are independent, identically distributed non-negative random variables, and thus form a renewal process.

which gives

$$E[B_1] = \frac{1}{\mu - \lambda}$$

for the mean length of a busy period.

In Section 9.3, in studying the $M/G/1$ system, we will reverse this reasoning, calculate the mean busy period directly, and then use renewal theory to determine the server idle fraction π_0.

9.2.2 The M/M/∞ System

When an unlimited number of servers are always available, then all customers in the system at any instant are simultaneously being served. With the departure rate of a single customer being μ, the departure rate of k customers is $k\mu$, and we obtain the birth and death parameters

$$\lambda_k = \lambda \quad \text{and} \quad \mu_k = k\mu \quad \text{for } k = 0, 1, \ldots.$$

The auxiliary quantities of (9.8) are

$$\theta_k = \frac{\lambda_0 \lambda_1 \cdots \lambda_{k-1}}{\mu_1 \mu_2 \cdots \mu_k} = \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k \quad \text{for } k = 0, 1, \ldots,$$

which sum to

$$\sum_{k=0}^{\infty} \theta_k = \sum_{k=0}^{\infty} \frac{1}{\mathrm{k!}}\left(\frac{\lambda}{\mu}\right)^k = \mathrm{e}^{\lambda/\mu},$$

whence

$$\pi_0 = \frac{1}{\Sigma_{k=0}^{\infty} \theta_k} = \mathrm{e}^{-\lambda/\mu}$$

and

$$\pi_k = \theta_k \pi_0 = \frac{(\lambda/\mu)^k \mathrm{e}^{-\lambda/\mu}}{k!} \quad \text{for } k = 0, 1, \ldots, \tag{9.16}$$

a Poisson distribution with mean queue length

$$L = \frac{\lambda}{\mu}.$$

Since a customer in this system begins service immediately upon arrival, customer waiting time consists only of the exponentially distributed service time, and the mean waiting time is $W = 1/\mu$. Again, the basic queueing formula $L = \lambda W$ is verified.

The $M/G/\infty$ queue will be developed extensively in the next section.

9.2.3 The M/M/s System

When a fixed number s of servers are available and the assumption is made that a server is never idle if customers are waiting, then the appropriate birth and death parameters are

$$\lambda_k = \lambda \quad \text{for } k = 1, 2, \ldots,$$

$$\mu_k = \begin{cases} k\mu & \text{for } k = 0, 1, \ldots, s, \\ s\mu & \text{for } k > s. \end{cases}$$

If $X(t)$ is the number of customers in the system at time t, then the number undergoing service is $\min\{X(t), s\}$, and the number waiting for service is $\max\{X(t) - s, 0\}$. The system is depicted in Figure 9.4.

The auxiliary quantities are given by

$$\theta_k = \frac{\lambda_0 \lambda_1 \cdots \lambda_{k-1}}{\mu_1 \mu_2 \cdots \mu_k} = \begin{cases} \dfrac{1}{k!}\left(\dfrac{\lambda}{\mu}\right)^k & \text{for } k = 0, 1, \ldots, s, \\ \dfrac{1}{s!}\left(\dfrac{\lambda}{\mu}\right)^s \left(\dfrac{\lambda}{s\mu}\right)^{k-s} & \text{for } k \geq s, \end{cases}$$

and when $\lambda < s\mu$, then

$$\begin{aligned} \sum_{j=0}^{\infty} \theta_j &= \sum_{j=0}^{s-1} \frac{1}{j!}\left(\frac{\lambda}{\mu}\right)^j + \sum_{j=s}^{\infty} \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \left(\frac{\lambda}{s\mu}\right)^{j-s} \\ &= \sum_{j=0}^{s-1} \frac{1}{j!}\left(\frac{\lambda}{\mu}\right)^j + \frac{(\lambda/\mu)^s}{s!\,(1 - \lambda/s\mu)} \quad \text{for } \lambda < s\mu. \end{aligned} \tag{9.17}$$

The traffic intensity in an $M/M/s$ system is $\rho = \lambda/(s\mu)$. Again, as the traffic intensity approaches one, the mean queue length becomes unbounded. When $\lambda < s\mu$, then

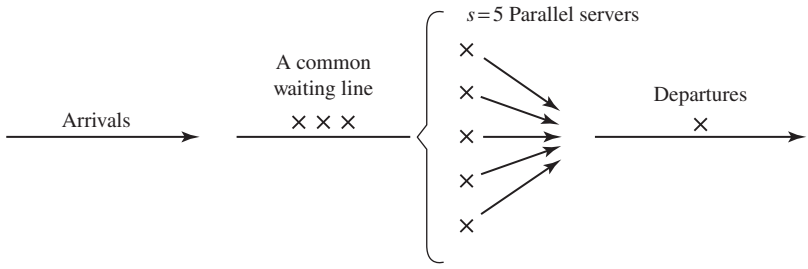


Figure 9.4 A queueing system with s servers.

from (9.10) and (9.17),

$$\pi_0 = \left\{ \sum_{j=0}^{s-1} \frac{1}{j!} \left(\frac{\lambda}{\mu} \right)^j + \frac{(\lambda/\mu)^s}{s!\,(1-\lambda/s\mu)} \right\}^{-1},$$

and

$$\pi_k = \begin{cases} \dfrac{1}{k!} \left(\dfrac{\lambda}{\mu} \right)^k \pi_0 & \text{for } k = 0, 1, \ldots, s, \\ \dfrac{1}{s!} \left(\dfrac{\lambda}{\mu} \right)^s \left(\dfrac{\lambda}{s\mu} \right)^{k-s} \pi_0 & \text{for } k \geq s. \end{cases} \tag{9.18}$$

We evaluate L_0, the mean number of customers in the system waiting for, and not undergoing, service. Then,

$$\begin{aligned} L_0 &= \sum_{j=s}^{\infty} (j-s)\pi_j = \sum_{k=0}^{\infty} k\pi_{s+k} \\ &= \pi_0 \sum_{k=0}^{\infty} k \frac{1}{s!} \left(\frac{\lambda}{\mu} \right)^s \left(\frac{\lambda}{s\mu} \right)^k \\ &= \frac{\pi_0}{s!} \left(\frac{\lambda}{\mu} \right)^s \sum_{k=0}^{\infty} k \left(\frac{\lambda}{s\mu} \right)^k \\ &= \frac{\pi_0}{s!} \left(\frac{\lambda}{\mu} \right)^s \frac{(\lambda/s\mu)}{(1-\lambda/s\mu)^2}. \end{aligned} \tag{9.19}$$

Then,

$$W_0 = \frac{L_0}{\lambda},$$

$$W = W_0 + \frac{1}{\mu},$$

and

$$L = \lambda W = \lambda \left(W_0 + \frac{1}{\mu} \right) = L_0 + \frac{\lambda}{\mu}.$$

Exercises

9.2.1 Customers arrive at a tool crib according to a Poisson process of rate $\lambda = 5$ per hour. There is a single tool crib employee, and the individual service times are

exponentially distributed with a mean service time of 10 min. In the long run, what is the probability that two or more workers are at the tool crib being served or waiting to be served?

9.2.2 On a single graph, plot the server utilization $1 - \pi_0 = \rho$ and the mean queue length $L = \rho/(1-\rho)$ for the $M/M/1$ queue as a function of the traffic intensity $\rho = \lambda/\mu$ for $0 < \rho < 1$.

9.2.3 Customers arrive at a checkout station in a market according to a Poisson process of rate $\lambda = 1$ customer per minute. The checkout station can be operated with or without a bagger. The checkout times for customers are exponentially distributed, and with a bagger the mean checkout time is 30 s, while without a bagger this mean time increases to 50 s. Compare the mean queue lengths with and without a bagger.

Problems

9.2.1 Determine explicit expressions for π_0 and L for the $M/M/s$ queue when $s = 2$. Plot $1 - \pi_0$ and L as a function of the traffic intensity $\rho = \lambda/2\mu$.

9.2.2 Determine the mean waiting time W for an $M/M/2$ system when $\lambda = 2$ and $\mu = 1.2$. Compare this with the mean waiting time in an $M/M/1$ system whose arrival rate is $\lambda = 1$ and service rate is $\mu = 1.2$. Why is there a difference when the arrival rate per server is the same in both cases?

9.2.3 Determine the stationary distribution for an $M/M/2$ system as a function of the traffic intensity $\rho = \lambda/2\mu$, and verify that $L = \lambda W$.

9.2.4 The problem is to model a queueing system having finite capacity. We assume arrivals according to a Poisson process of rate λ, with independent exponentially distributed service times having mean $1/\mu$, a single server, and a finite system capacity N. By this we mean that if an arriving customer finds that there are already N customers in the system, then that customer does not enter the system and is lost.

Let $X(t)$ be the number of customers in the system at time t. Suppose that $N = 3$ (2 waiting, 1 being served).

(a) Specify the birth and death parameters for $X(t)$.

(b) In the long run, what fraction of time is the system idle?

(c) In the long run, what fraction of customers are lost?

9.2.5 Customers arrive at a service facility according to a Poisson process having rate λ. There is a single server, whose service times are exponentially distributed with parameter μ. Let $N(t)$ be the number of people in the system at time t. Then, $N(t)$ is a birth and death process with parameters $\lambda_n = \lambda$ for $n \geq 0$ and $\mu_n = \mu$ for $n \geq 1$. Assume $\lambda < \mu$. Then, $\pi_k = (1 - \lambda/\mu)(\lambda/\mu)^k, k \geq 0$, is a stationary distribution for $N(t)$; cf. equation (9.11).

Suppose the process begins according to the stationary distribution. That is, suppose $\Pr\{N(0) = k\} = \pi_k$ for $k = 0, 1, \ldots$. Let $D(t)$ be the number of people completing service up to time t. Show that $D(t)$ has a Poisson distribution with mean λt.

Hint: Let $P_{kj}(t) = \Pr\{D(t) = j | N(0) = k\}$ and $P_j(t) = \Sigma \pi_k P_{kj}(t) = \Pr\{D(t) = j\}$. Use a first step analysis to show that $P_{0j}(t + \Delta t) = \lambda(\Delta t)P_{1j}(t) + [1 - \lambda(\Delta t)]$ $P_{0j}(t) + o(\Delta t)$, and for $k = 1, 2, \ldots,$

$$P_{kj}(t + \Delta t) = \mu(\Delta t)P_{k-1,j-1}(t) + \lambda(\Delta t)P_{k+1,j}(t) + [1 - (\lambda + \mu)(\Delta t)]P_{kj}(t) + o(t).$$

Then, use $P_j(t) = \sum_k \pi_k P_{kj}(t)$ to establish a differential equation. Use the explicit form of π_k given in the problem.

9.2.6 Customers arrive at a service facility according to a Poisson process of rate λ. There is a single server, whose service times are exponentially distributed with parameter μ. Suppose that "gridlock" occurs whenever the total number of customers in the system exceeds a capacity C. What is the smallest capacity C that will keep the probability of gridlock, under the limiting distributing of queue length, below 0.001? Express your answer in terms of the traffic intensity $\rho = \lambda/\mu$.

9.2.7 Let $X(t)$ be the number of customers in an $M/M/\infty$ queueing system at time t. Suppose that $X(0) = 0$.

(a) Derive the forward equations that are appropriate for this process by substituting the birth and death parameters into Chapter 6, equation (6.24).

(b) Show that $M(t) = E[X(t)]$ satisfies the differential equation $M'(t) = \lambda - \mu M(t)$ by multiplying the jth forward equation by j and summing.

(c) Solve for $M(t)$.

9.3 General Service Time Distributions

We continue to assume that the arrivals follow a Poisson process of rate λ. The successive customer service times $Y_1, Y_2, \ldots,$ however, are now allowed to follow an arbitrary distribution $G(y) = \Pr\{Y_k \leq y\}$ having a finite mean service time $\nu = E[Y_k]$. The long run service rate is $\mu = 1/\nu$. Deterministic service times of an equal fixed duration are an important special case.

9.3.1 The M/G/1 System

If arrivals to a queue follow a Poisson process, then the successive durations X_k from the commencement of the kth busy period to the start of the next busy period form a renewal process. (A busy period is an uninterrupted duration when the queue is not empty. See Figure 9.3.) Each X_k is composed of a busy portion B_k and an idle portion I_k. Then $p_0(t)$, the probability that the system is empty at time t, converges to

$$\lim_{t\to\infty} p_0(t) = \pi_0 = \frac{E[I_1]}{E[X_1]} = \frac{E[I_1]}{E[I_1] + E[B_1]} \tag{9.20}$$

by the renewal theorem (see "A Queueing Model" in Chapter 7, Section 7.5.3).

The idle time is the duration from the completion of a service that empties the queue to the instant of the next arrival. Because of the memoryless property that characterizes the interarrival times in a Poisson process, each idle time is exponentially distributed with mean $E[I_1] = 1/\lambda$.

The busy period is composed of the first service time Y_1, plus busy periods generated by all customers who arrive during this first service time. Let A denote this random number of new arrivals. We will evaluate the conditional mean busy period given that $A = n$ and $Y_1 = y$. First,

$$E[B_1|A = 0, Y_1 = y] = y,$$

because when no customers arrive, the busy period is composed of the first customer's service time alone. Next, consider the case in which $A = 1$, and let B' be the duration from the beginning of this customer's service to the next instant that the queue is empty. Then,

$$\begin{aligned} E[B_1|A = 1, Y_1 = y] &= y + E[B'] \\ &= y + E[B_1], \end{aligned}$$

because upon the completion of service for the initial customer, the single arrival begins a busy period B' that is statistically identical to the first, so that $E[B'] = E[B_1]$. Continuing in this manner we deduce that

$$E[B_1|A = n, Y_1 = y] = y + nE[B_1]$$

and then, using the law of total probability, that

$$\begin{aligned} E[B_1|Y_1 = y] &= \sum_{n=0}^{\infty} E[B_1|A = n, Y_1 = y] \Pr\{A = n|Y_1 = y\} \\ &= \sum_{n=0}^{\infty} \{y + nE[B_1]\} \frac{(\lambda y)^n \mathrm{e}^{-\lambda y}}{n!} \\ &= y + \lambda y E[B_1]. \end{aligned}$$

Finally,

$$\begin{aligned} E[B_1] &= \int_0^{\infty} E[B_1|Y_1 = y] \mathrm{d}G(y) \\ &= \int_0^{\infty} \{y + \lambda y E[B_1]\} \mathrm{d}G(y) \\ &= \nu\{1 + \lambda E[B_1]\}. \end{aligned} \tag{9.21}$$

Since $E[B_1]$ appears on both sides of (9.21), we may solve to obtain

$$E[B_1] = \frac{\nu}{1-\lambda\nu}, \quad \text{provided that } \lambda\nu < 1. \tag{9.22}$$

To compute the long run fraction of idle time, we use (9.22) and

$$\begin{aligned} \pi_0 &= \frac{E[I_1]}{E[I_1]+E[B_1]} \\ &= \frac{1/\lambda}{1/\lambda + \nu/(1-\lambda\nu)} \\ &= 1-\lambda\nu \quad \text{if } \lambda\nu < 1. \end{aligned} \tag{9.23}$$

Note that (9.23) agrees, as it must, with the corresponding expression (9.10) obtained for the $M/M/1$ queue where $\nu = 1/\mu$. For example, if arrivals occur at the rate of $\lambda = 2$ per hour and the mean service time is 20 min, or $\nu = \frac{1}{3}$ h, then in the long run, the server is idle $1 - 2\left(\frac{1}{3}\right) = \frac{1}{3}$ of the time.

The Embedded Markov Chain

The number $X(t)$ of customers in the system at time t is not a Markov process for a general $M/G/1$ system, because if one is to predict the future behavior of the system, one must know, in addition, the time expended in service for the customer currently in service. (It is the memoryless property of the exponential service time distribution that makes this additional information unnecessary in the $M/M/1$ case.)

Let X_n, however, denote the number of customers in the system immediately after the departure of the nth customer. Then, $\{X_n\}$ is a Markov chain. Indeed, we can write

$$\begin{aligned} X_n &= \begin{cases} X_{n-1} - 1 + A_n & \text{if } X_{n-1} > 0, \\ A_n & \text{if } X_{n-1} = 0, \end{cases} \\ &= (X_{n-1}-1)^+ + A_n, \end{aligned} \tag{9.24}$$

where A_n is the number of customers that arrive during the service of the nth customer and where $x^+ = \max\{x, 0\}$. Since the arrival process is Poisson, the number of customers A_n that arrive during the service of the nth customer is independent of earlier arrivals, and the Markov property follows instantly. We calculate

$$\begin{aligned} \alpha_k = \Pr\{A_n = k\} &= \int_0^\infty \Pr\{A_n = k | Y_n = y\} \, \mathrm{d}G(y) \\ &= \int_0^\infty \frac{(\lambda y)^k \mathrm{e}^{-\lambda y}}{k!} \mathrm{d}G(y), \end{aligned} \tag{9.25}$$

and then, for $j = 0, 1, \ldots,$

$$P_{ij} = \Pr\{X_n = j | X_{n-1} = i\} = \Pr\{A_n = j - (i-1)^+\}$$
$$= \begin{cases} \alpha_{j-i+1} & \text{for } i \geq 1, j \geq i+1, \\ \alpha_j & \text{for } i = 0. \end{cases} \tag{9.26}$$

The Mean Queue Length in Equilibrium L

The embedded Markov chain is of special interest in the $M/G/1$ queue because in this particular instance, the stationary distribution $\{\pi_j\}$ for the Markov chain $\{X_n\}$ equals the limiting distribution for the queue length process $\{X(t)\}$. That is, $\lim_{t\to\infty} \Pr\{X(t) = j\} = \lim_{n\to\infty} \Pr\{X_n = j\}$. We will use this helpful fact to evaluate the mean queue length L.

The equivalence between the stationary distribution for the Markov chain $\{X_n\}$ and that for the non-Markov process $\{X(t)\}$ is rather subtle. It is not the consequence of a general principle and should not be assumed to hold in other circumstances without careful justification. The equivalence in the case at hand is sketched in an appendix to this section.

We will calculate the expected queue length in equilibrium $L = \lim_{t\to\infty} E[X(t)]$ by calculating the corresponding quantity in the embedded Markov chain, $L = \lim_{n\to\infty} E[X_n]$. If $X = X_\infty$ is the number of customers in the system after a customer departs and X' is the number after the next departure, then in accordance with (9.24),

$$X' = X - \delta + N, \tag{9.27}$$

where N is the number of arrivals during the service period and

$$\delta = \begin{cases} 1 & \text{if } X > 0, \\ 0 & \text{if } X = 0. \end{cases}$$

In equilibrium, X has the same distribution as does X', and in particular,

$$L = E[X] = E[X'], \tag{9.28}$$

and taking expectation in (9.27) gives

$$E[X'] = E[X] - E[\delta] + E[N],$$

and, by (9.28) and (9.23), then

$$E[N] = E[\delta] = 1 - \pi_0 = \lambda\nu. \tag{9.29}$$

Squaring (9.27) gives

$$(X')^2 = X^2 + \delta^2 + N^2 - 2\delta X + 2N(X - \delta),$$

and since $\delta^2 = \delta$ and $X\delta = X$, then

$$(X')^2 = X^2 + \delta + N^2 - 2X + 2N(X - \delta). \tag{9.30}$$

Now N, the number of customers that arrive during a service period, is independent of X, and hence of δ, so that

$$E[N(X - \delta)] = E[N]E[X - \delta], \tag{9.31}$$

and because X and X' have the same distribution, then

$$E\big[(X')^2\big] = E\big[X^2\big]. \tag{9.32}$$

Taking expectations in (9.30) we deduce that

$$E\big[(X')^2\big] = E\big[X^2\big] + E[\delta] + E\big[N^2\big] - 2E[X] + 2E[N]E[X - \delta],$$

and then substituting from (9.29) and (9.32), we obtain

$$0 = \lambda\nu + E\big[N^2\big] - 2L + 2\lambda\nu\{L - \lambda\nu\},$$

or

$$L = \frac{\lambda\nu + E\big[N^2\big] - 2(\lambda\nu)^2}{2(1 - \lambda\nu)}. \tag{9.33}$$

It remains to evaluate $E\big[N^2\big]$, where N is the number of arrivals during a service time Y. Conditioned on $Y = y$, the random variable N has a Poisson distribution with a mean (and variance) equal to λy [see equation (9.26)], whence $E\big[N^2|Y = y\big] = \lambda y + (\lambda y)^2$. Using the law of total probability, then, gives

$$\begin{aligned} E\big[N^2\big] &= \int_0^\infty E\big[N^2|Y = y\big]\mathrm{d}G(y) \\ &= \lambda \int_0^\infty y\mathrm{d}G(y) + \lambda^2 \int_0^\infty y^2\mathrm{d}G(y) \\ &= \lambda\nu + \lambda^2\big(\tau^2 + \nu^2\big), \end{aligned} \tag{9.34}$$

where τ^2 is the variance of the service time distribution $G(y)$. Substituting (9.34) into (9.33) gives

$$\begin{aligned} L &= \frac{2\lambda\nu + \lambda^2\tau^2 - (\lambda\nu)^2}{2(1 - \lambda\nu)} \\ &= \rho + \frac{\lambda^2\tau^2 + \rho^2}{2(1 - \rho)}, \end{aligned} \tag{9.35}$$

where $\rho = \lambda\nu$ is the traffic intensity.

Finally, $W = L/\lambda$, which simplifies to

$$W = \nu + \frac{\lambda\left(\tau^2 + \nu^2\right)}{2(1-\rho)}. \tag{9.36}$$

The results (9.35) and (9.36) express somewhat surprising facts. They say that for a given average arrival rate λ and mean service time ν, we can decrease the expected queue size L and waiting time W by decreasing the variance of service time. Clearly, the best possible case in this respect corresponds to constant service times, for which $\tau^2 = 0$.

Appendix

We sketch a proof of the equivalence between the limiting queue size distribution and the limiting distribution for the embedded Markov chain in an $M/G/1$ model. First, beginning at $t = 0$ let η_n denote those instants when the queue size $X(t)$ increases by one (an arrival), and let ξ_n denote those instants when $X(t)$ decreases by one (a departure). Let $Y_n = X(\eta_n -)$ denote the queue length immediately prior to an arrival and let $X_n = X(\xi_n +)$ denote the queue length immediately after a departure. For any queue length i and any time t, the number of visits of Y_n to i up to time t differs from the number of visits of X_n to i by at most one unit. Therefore, in the long run the average visits per unit time of Y_n to i must equal the average visits of X_n to i, which is π_i, the stationary distribution of the Markov chain $\{X_n\}$. Thus, we need only show that the limiting distribution of $\{X(t)\}$ is the same as that of $\{Y_n\}$, which is $X(t)$ just prior to an arrival. But because the arrivals are Poisson, and arrivals in disjoint time intervals are independent, it must be that $X(t)$ is independent of an arrival that occurs at time t. It follows that $\{X(t)\}$ and $\{Y_n\}$ have the same limiting distribution, and therefore $\{X(t)\}$ and the embedded Markov chain $\{X_n\}$ have the same limiting distribution.

9.3.2 The M/G/∞ System

Complete results are available when each customer begins service immediately upon arrival independently of other customers in the system. Such situations may arise when modeling customer self-service systems. Let $W_1, W_2, \ldots$ be the successive arrival times of customers, and let $V_1, V_2, \ldots$ be the corresponding service times. In this notation, the kth customer is in the system at time t if and only if $W_k \le t$ (the customer arrived prior to t) and $W_k + V_k > t$ (the service extends beyond t).

The sequence of pairs $(W_1, V_1), (W_2, V_2), \ldots$ forms a *marked Poisson process* (see Chapter 5, Section 5.6.2), and we may use the corresponding theory to quickly obtain results in this model. Figure 9.5 illustrates the marked Poisson process. Then $X(t)$, the number of customers in the system at time t, is also the number of points (W_k, V_k) for which $W_k \le t$ and $W_k + V_k > t$. That is, it is the number of points (W_k, V_k) in the unbounded trapezoid described by

$$A_t = \{(w, v) :\ 0 \le w \le t \text{ and } v > t - w\}. \tag{9.37}$$

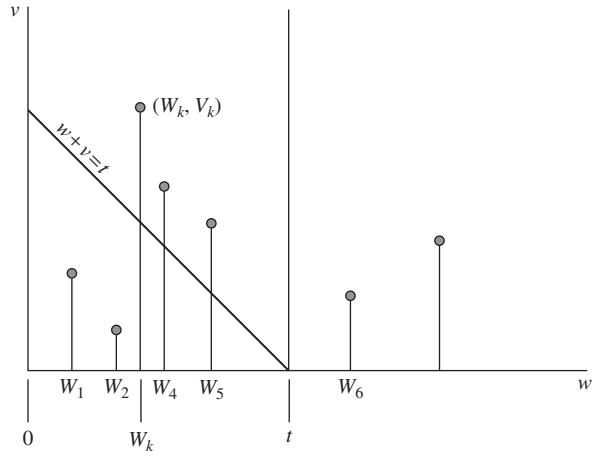


Figure 9.5 For the $M/G/\infty$ queue, the number of customers in the system at time t corresponds to the number of pairs (W_k, V_k) for which $W_k \le t$ and $W_k + V_k > t$. In the sample illustrated here, the number of customers in the system at time t is 3.

According to Chapter 5, Theorem 5.8, the number of points in A_t follows a Poisson distribution with mean

$$\begin{aligned}\mu(A_t) &= \iint_{A_i} \lambda(\mathrm{d}w)\,\mathrm{d}G(v) \\ &= \lambda \int_0^t \left\{ \int_{t-w}^{\infty} \mathrm{d}G(v) \right\} \mathrm{d}w \\ &= \lambda \int_0^t [1 - G(t-w)]\mathrm{d}w \\ &= \lambda \int_0^t [1 - G(x)]\mathrm{d}x.\end{aligned} \tag{9.38}$$

In summary,

$$\begin{aligned}p_k(t) &= \Pr\{X(t) = k\} \\ &= \frac{\mu(A_t)^k \mathrm{e}^{-\mu(A_t)}}{k!} \quad \text{for } k = 0, 1, \ldots,\end{aligned}$$

where $\mu(A_t)$ is given by (9.38). As $t \to \infty$, then

$$\lim_{t\to\infty} \mu(A_t) = \lambda \int_0^\infty [1 - G(x)]\mathrm{d}x = \lambda\nu,$$

where ν is the mean service time. Thus, we obtain the limiting distribution

$$\pi_k = \frac{(\lambda\nu)^k \mathrm{e}^{-\lambda\nu}}{k!} \quad \text{for } k = 0, 1, \ldots.$$

Exercises

9.3.1 Suppose that the service distribution in a single server queue is exponential with rate μ; i.e., $G(\nu) = 1 - \mathrm{e}^{-\mu\nu}$ for $\nu \geq 0$. Substitute the mean and variance of this distribution into (9.35) and verify that the result agrees with that derived for the $M/M/1$ system in (9.12).

9.3.2 Consider a single-server queueing system having Poisson arrivals at rate λ. Suppose that the service times have the gamma density

$$g(y) = \frac{\mu^\alpha y^{\alpha-1} \mathrm{e}^{-\mu y}}{\Gamma(\alpha)} \quad \text{for } y \geq 0,$$

where $\alpha > 0$ and $\mu > 0$ are fixed parameters. The mean service time is α/μ and the variance is α/μ^2. Determine the equilibrium mean queue length L.

9.3.3 Customers arrive at a tool crib according to a Poisson process of rate $\lambda = 5$ per hour. There is a single tool crib employee, and the individual service times are random with a mean service time of 10 min and a standard deviation of 4 min. In the long run, what is the mean number of workers at the tool crib either being served or waiting to be served?

9.3.4 Customers arrive at a checkout station in a market according to a Poisson process of rate $\lambda = 1$ customer per minute. The checkout station can be operated with or without a bagger. The checkout times for customers are random. With a bagger the mean checkout time is 30 s, while without a bagger this mean time increases to 50 s. In both cases, the standard deviation of service time is 10 s. Compare the mean queue lengths with and without a bagger.

9.3.5 Let $X(t)$ be the number of customers in an $M/G/\infty$ queueing system at time t. Suppose that $X(0) = 0$. Evaluate $M(t) = E[X(t)]$, and show that it increases monotonically to its limiting value as $t \to \infty$.

Problems

9.3.1 Let $X(t)$ be the number of customers in an $M/G/\infty$ queueing system at time t, and let $Y(t)$ be the number of customers who have entered the system and completed service by time t. Determine the joint distribution of $X(t)$ and $Y(t)$.

9.3.2 In operating a queueing system with Poisson arrivals at a rate of $\lambda = 1$ per unit time and a single server, you have a choice of server mechanisms. Method A has a mean service time of $\nu = 0.5$ and a variance in service time of $\tau^2 = 0.2$, while Method B has a mean service time of $\nu = 0.4$ and a variance of $\tau^2 = 0.9$. In terms of minimizing the waiting time of a typical customer, which method do you prefer? Would your answer change if the arrival rate were to increase significantly?

9.4 Variations and Extensions

In this section, we consider a few variations on the simple queueing models studied so far. These examples do not exhaust the possibilities but serve only to suggest the richness of the area.

Throughout we restrict ourselves to Poisson arrivals and exponentially distributed service times.

9.4.1 Systems with Balking

Suppose that a customer who arrives when there are n customers in the systems enters with probability p_n and departs with probability $q_n = 1 - p_n$. If long queues discourage customers, then p_n would be a decreasing function of n. As a special case, if there is a finite waiting room of capacity C, we might suppose that

$$p_n = \begin{cases} 1 & \text{for } n < C, \\ 0 & \text{for } n \geq C, \end{cases}$$

indicating that once the waiting room is filled, no more customers can enter the system.

Let $X(t)$ be the number of customers in the system at time t. If the arrival process is Poisson at rate λ and a customer who arrives when there are n customers in the system enters with probability p_n, then the appropriate birth parameters are

$$\lambda_n = \lambda p_n \quad \text{for } n = 0, 1, \ldots .$$

In the case of a single server, then $\mu_n = \mu$ for $n = 1, 2, \ldots$, and we may evaluate the stationary distribution π_k of queue length by the usual means.

In systems with balking, not all arriving customers enter the system, and some are lost. The *input rate* is the rate at which customers actually enter the system in the stationary state and is given by

$$\lambda_I = \lambda \sum_{n=0}^{\infty} \pi_n p_n.$$

The rate at which customers are lost is $\lambda \sum_{n=0}^{\infty} \pi_n q_n$, and the fraction of customers lost in the long run is

$$\text{fraction lost} = \sum_{n=0}^{\infty} \pi_n q_n.$$

Let us examine in detail the case of an $M/M/s$ system in which an arriving customer enters the system if and only if a server is free. Then,

$$\lambda_k = \begin{cases} \lambda & \text{for } k = 0, 1, \ldots, s-1, \\ 0 & \text{for } k = s, \end{cases}$$

and

$$\mu_k = k\mu \quad \text{for } k = 0, 1, \ldots, s.$$

To determine the limiting distribution, we have

$$\theta_k = \frac{1}{k!} \left(\frac{\lambda}{\mu} \right)^k \quad \text{for } k = 0, 1, \ldots, s,$$

and then

$$\pi_k = \frac{\dfrac{1}{k!} \left(\dfrac{\lambda}{\mu} \right)^k}{\displaystyle\sum_{j=0}^{s} \frac{1}{j!} \left(\frac{\lambda}{\mu} \right)^j} \quad \text{for } k = 0, 1, \ldots, s. \tag{9.39}$$

The long run fraction of customers lost is $\pi_s q_s = \pi_s$, since $q_s = 1$ in this case.

9.4.2 Variable Service Rates

In a similar vein, one can consider a system whose service rate depends on the number of customers in the system. For example, a second server might be added to a single-server system whenever the queue length exceeds a critical point ξ. If arrivals are Poisson and service rates are memoryless, then the appropriate birth and death parameters are

$$\lambda_k = \lambda \quad \text{for } k = 0, 1, \ldots, \quad \text{and} \quad \mu_k = \begin{cases} \mu & \text{for } k \le \xi, \\ 2\mu & \text{for } k > \xi. \end{cases}$$

More generally, let us consider Poisson arrivals $\lambda_k = \lambda$ for $k = 0, 1, \ldots$, and arbitrary service rates μ_k for $k = 1, 2, \ldots$. The stationary distribution in this case is given by

$$\pi_k = \frac{\pi_0 \lambda^k}{\mu_1 \mu_2 \cdots \mu_k} \quad \text{for } k \ge 1, \tag{9.40}$$

where

$$\pi_0 = \left\{ 1 + \sum_{k=1}^{\infty} \frac{\lambda^k}{\mu_1 \mu_2 \cdots \mu_k} \right\}^{-1}. \tag{9.41}$$

9.4.3 A System with Feedback

Consider a single-server system with Poisson arrivals and exponentially distributed service times, but suppose that some customers, upon leaving the server, return to the end of the queue for additional service. In particular, suppose that a customer leaving the server departs from the system with probability q and returns to the queue for additional service with probability $p = 1 - q$. Suppose that all such decisions are statistically independent, and that a returning customer's demands for service are statistically the same as those of a customer arriving from outside the system. Let the arrival rate be λ and the service rate be μ. The queue system is depicted in Figure 9.6.

Let $X(t)$ denote the number of customers in the system at time t. Then, $X(t)$ is a birth and death process with parameters $\lambda_n = \lambda$ for $n = 0, 1, \ldots$ and $\mu_n = q\mu$ for $n = 1, 2, \ldots$. It is easily deduced that the stationary distribution in the case that $\lambda < q\mu$ is

$$\pi_k = \left(1 - \frac{\lambda}{q\mu}\right)\left(\frac{\lambda}{q\mu}\right)^k \quad \text{for } k = 0, 1, \ldots. \tag{9.42}$$

9.4.4 A Two-Server Overflow Queue

Consider a two-server system where server i has rate μ_i for $i = 1, 2$. Arrivals to the system follow a Poisson process of rate λ. A customer arriving when the system is empty goes to the first server. A customer arriving when the first server is occupied goes to the second server. If both servers are occupied, the customer is lost. The flow is depicted in Figure 9.7.

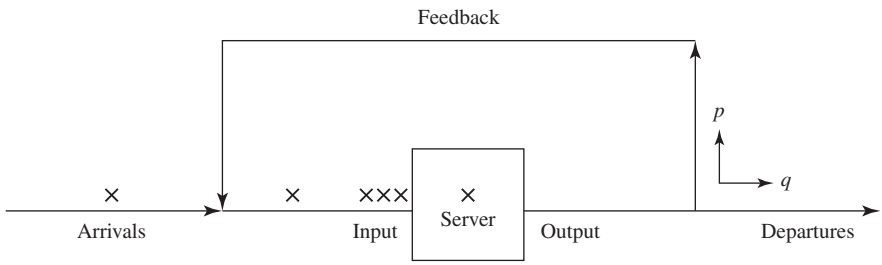


Figure 9.6 A queue with feedback.

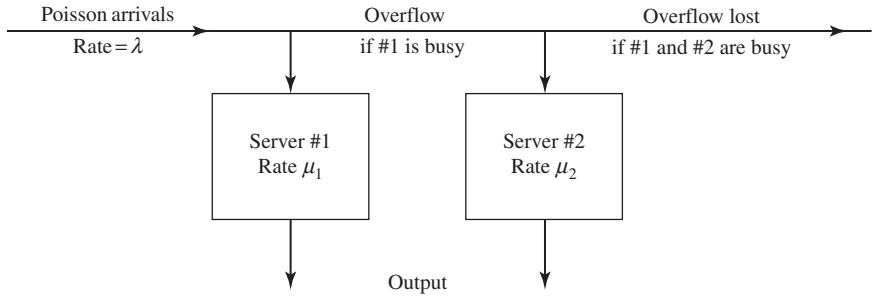


Figure 9.7 A two-server overflow model.

The system state is described by the pair $(X(t), Y(t))$, where

$$X(t) = \begin{cases} 1 & \text{if Server \#1 is busy,} \\ 0 & \text{if Server \#1 is idle.} \end{cases}$$

and

$$Y(t) = \begin{cases} 1 & \text{if Server \#2 is busy,} \\ 0 & \text{if Server \#2 is idle.} \end{cases}$$

The four states of the system are $\{(0,0), (1,0), (0,1), (1,1)\}$, and transitions among these states occur at the rate given in the following table:

From State	To State	Transition Rate	Description
$(0,0)$	$(1,0)$	λ	Arrival when system is empty
$(1,0)$	$(0,0)$	μ_1	Service completion by #1 when #2 is free
$(1,0)$	$(1,1)$	λ	Arrival when #1 is busy
$(1,1)$	$(1,0)$	μ_2	Service completion by #2 when #1 is busy
$(1,1)$	$(0,1)$	μ_1	Service completion by #1 when #2 is busy
$(0,1)$	$(1,1)$	λ	Arrival when #2 is busy and #1 is free
$(0,1)$	$(0,0)$	μ_2	Service completion by #2 when #1 is free

The process $(X(t), Y(t))$ is a finite-state, continuous-time Markov chain (see Chapter 6, Section 6.6), and the transition rates in the table furnish the infinitesimal matrix of the Markov chain:

$$\mathbf{A} = \begin{array}{c} \\ (0,0) \\ (0,1) \\ (1,0) \\ (1,1) \end{array} \begin{array}{c} \begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array} \\ \left\| \begin{array}{cccc} -\lambda & 0 & \lambda & 0 \\ \mu_2 & -(\lambda+\mu_2) & 0 & \lambda \\ \mu_1 & 0 & -(\lambda+\mu_1) & \lambda \\ 0 & \mu_1 & \mu_2 & -(\mu_1+\mu_2) \end{array} \right\| \end{array}.$$

From Chapter 6, equations (6.68) and (6.69), we find the stationary distribution $\pi = (\pi_{(0,0)}, \pi_{(0,1)}, \pi_{(1,0)}\pi_{(1,1)})$ by solving $\pi\mathbf{A} = 0$, or

$$\begin{aligned}
-\lambda\pi_{(0,0)} \quad &+\mu_2\pi_{(0,1)} \quad &+\mu_1\pi_{(1,0)} \quad &= 0,\\
-(\lambda+\mu_2)\pi_{(0,1)} \quad &&+\mu_1\pi_{(1,1)} &= 0,\\
\lambda\pi_{(0,0)} \quad &-(\lambda+\mu_1)\pi_{(1,0)} \quad &+\mu_2\pi_{(1,1)} &= 0,\\
\lambda\pi_{(0,1)} \quad &+\lambda\pi_{(1,0)} \quad &-(\mu_1+\mu_2)\pi_{(1,1)} &= 0,
\end{aligned}$$

together with

$$\pi_{(0,0)} + \pi_{(0,1)} + \pi_{(1,0)} + \pi_{(1,1)} = 1.$$

Tedious but elementary algebra yields the solution:

$$\begin{aligned}
\pi_{(0,0)} &= \frac{\mu_1\mu_2(2\lambda+\mu_1+\mu_2)}{D},\\
\pi_{(0,1)} &= \frac{\lambda^2\mu_1}{D},\\
\pi_{(1,0)} &= \frac{\lambda\mu_2(\lambda+\mu_1+\mu_2)}{D},\\
\pi_{(1,1)} &= \frac{\lambda^2(\lambda+\mu_2)}{D},
\end{aligned} \tag{9.43}$$

where

$$\begin{aligned}
D = {} & \mu_1\mu_2(2\lambda+\mu_1+\mu_2) + \lambda^2\mu_1 + \lambda\mu_2(\lambda+\mu_1+\mu_2)\\
& + \lambda^2(\lambda+\mu_2).
\end{aligned}$$

The fraction of customers that are lost, in the long run, is the same as the fraction of time that both servers are busy, $\pi_{(1,1)} = \lambda^2(\lambda+\mu_2)/D$.

9.4.5 Preemptive Priority Queues

Consider a single-server queueing process that has two classes of customers, *priority* and *nonpriority*, forming independent Poisson arrival processes of rates α and β, respectively. The customer service times are independent and exponentially distributed with parameters γ and δ, respectively. Within classes there is a first come, first served discipline, and the service of priority customers is never interrupted. If a priority customer arrives during the service of a nonpriority customer, then the latter's service is immediately stopped in favor of the priority customer. The interrupted customer's service is resumed when there are no priority customers present.

Let us introduce some convenient notation. The system arrival rate is $\lambda = \alpha + \beta$, of which the fraction $p = \alpha/\lambda$ are priority customers and $q = \beta/\lambda$ are nonpriority customers. The system mean service time is given by the appropriately weighted means

$1/\gamma$ and $1/\delta$ of the priority and nonpriority customers, respectively, or

$$\frac{1}{\mu} = p\left(\frac{1}{\gamma}\right) + q\left(\frac{1}{\delta}\right) = \frac{1}{\lambda}\left(\frac{\alpha}{\gamma} + \frac{\beta}{\delta}\right), \tag{9.44}$$

where μ is the system service rate. Finally, we introduce the traffic intensities $\rho = \lambda/\mu$ for the system, and $\sigma = \alpha/\gamma$ and $\tau = \beta/\delta$ for the priority and nonpriority customers, respectively. From (9.44) we see that $\rho = \sigma + \tau$.

The state of the system is described by the pair $(X(t), Y(t))$, where $X(t)$ is the number of priority customers in the system and $Y(t)$ is the number of nonpriority customers. Observe that the priority customers view the system as simply an $M/M/1$ queue. Accordingly, we have the limiting distribution from (9.11) to be

$$\lim_{t\to\infty} \Pr\{X(t) = m\} = (1-\sigma)\sigma^m \quad \text{for } m = 0, 1, \ldots \tag{9.45}$$

provided $\sigma = \alpha/\gamma < 1$.

Reference to (9.12) and (9.15) gives us, respectively, the mean queue length for priority customers

$$L_{\mathrm{p}} = \frac{\alpha}{\gamma - \alpha} = \frac{\sigma}{1-\sigma} \tag{9.46}$$

and the mean wait for priority customers

$$W_{\mathrm{p}} = \frac{1}{\gamma - \alpha}. \tag{9.47}$$

To obtain information about the nonpriority customers is not as easy, since these arrivals are strongly affected by the priority customers. Nevertheless, $(X(t), Y(t))$ is a discrete-state, continuous-time Markov chain, and the techniques of Chapter 6, Section 6.6 enable us to describe the limiting distribution, when it exists. The transition rates of the $(X(t), Y(t))$ Markov chain are described in the following table:

From State	To State	Transition Rate	Description
(m, n)	$(m+1, n)$	α	Arrival of priority customer
(m, n)	$(m, n+1)$	β	Arrival of nonpriority customer
$(0, n)$ $n \geq 1$	$(0, n-1)$	δ	Completion of nonpriority service
(m, n) $m \geq 1$	$(m-1, n)$	γ	Completion of priority service

Let

$$\pi_{m,n} = \lim_{t\to\infty} \Pr\{X(t) = m, Y(t) = n\}$$

be the limiting distribution of the process. Reasoning analogous to that of Chapter 6, equations (6.68) and (6.69) (where the theory was derived for a finite-state Markov chain) leads to the following equations for the stationary distribution:

$$(\alpha+\beta)\pi_{0,0} = \gamma\pi_{1,0} + \delta\pi_{0,1}, \tag{9.48}$$

$$(\alpha+\beta+\gamma)\pi_{m,0} = \gamma\pi_{m+1,0} + \alpha\pi_{m-1,0}, \quad m \geq 1, \tag{9.49}$$

$$(\alpha+\beta+\delta)\pi_{0,n} = \gamma\pi_{1,n} + \delta\pi_{0,n+1} + \beta\pi_{0,n-1}, \quad n \geq 1, \tag{9.50}$$

$$(\alpha+\beta+\gamma)\pi_{m,n} = \gamma\pi_{m+1,n} + \beta\pi_{m,n-1} + \alpha\pi_{m-1,n}, \quad m,n \geq 1. \tag{9.51}$$

The transition rates leading to equation (9.8) are shown in Figure 9.8.

In principle, these equations, augmented with the condition $\Sigma_m \Sigma_n \pi_{m,n} = 1$, may be solved for the stationary distribution, when it exists. We will content ourselves with determining the mean number L_{n} of nonpriority customers in the system in steady state, given by

$$L_{\mathrm{n}} = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} n\pi_{m,n}. \tag{9.52}$$

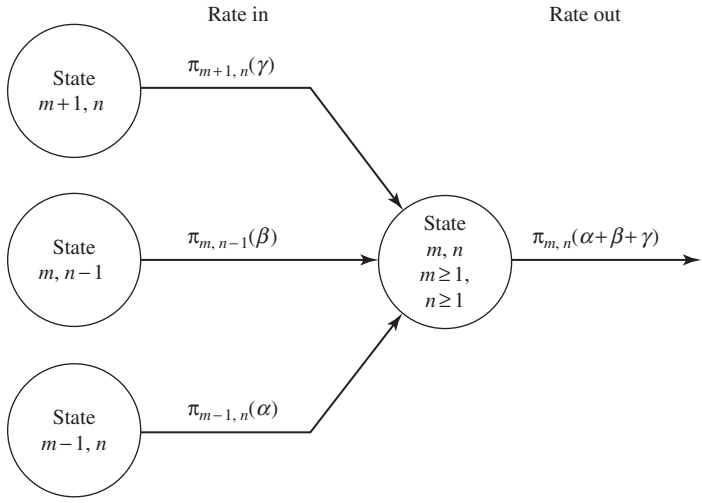


Figure 9.8 In equilibrium, the rate of flow into any state must equal the rate of flow out. Illustrated here is the state (m, n) when $m \geq 1$ and $n \geq 1$, leading to equation (9.51).

We introduce the notation

$$M_m = \sum_{n=0}^{\infty} n\pi_{m,n} = \sum_{n=1}^{\infty} n\pi_{m,n}, \tag{9.53}$$

so that

$$L_{\mathrm{n}} = M_0 + M_1 + \cdots. \tag{9.54}$$

Using (9.45), let

$$p_m = \Pr\{X(t) = m\} = \sum_{n=0}^{\infty} \pi_{m,n} = (1-\sigma)\sigma^m \tag{9.55}$$

and

$$\pi_n = \Pr\{Y(t) = n\} = \sum_{m=0}^{\infty} \pi_{m,n}. \tag{9.56}$$

We begin by summing both sides of (9.48) and (9.49) for $m = 0, 1, \ldots$ to obtain

$$(\alpha+\beta)\pi_0 + \gamma \sum_{m=1}^{\infty} \pi_{m,0} = \gamma \sum_{m=1}^{\infty} \pi_{m,0} + \delta\pi_{0,1} + \alpha\pi_0,$$

which simplifies to give

$$\beta\pi_0 = \delta\pi_{0,1}. \tag{9.57}$$

Next, we sum (9.50) and (9.51) over $m = 0, 1, \ldots$ to obtain

$$(\alpha+\beta)\pi_n + \delta\pi_{0,n} + \gamma \sum_{m=1}^{\infty} \pi_{m,n} = \gamma \sum_{m=1}^{\infty} \pi_{m,n} + \delta\pi_{0,n+1} + \beta\pi_{n-1} + \alpha\pi_n,$$

which simplifies to

$$\beta\pi_n + \delta\pi_{0,n} = \beta\pi_{n-1} + \delta\pi_{0,n+1},$$

and inductively with (9.57), we obtain

$$\beta\pi_n = \delta\pi_{0,n+1} \quad \text{for } n = 0, 1, \ldots. \tag{9.58}$$

Summing (9.58) over $n = 0, 1, \ldots$ and using $\Sigma\, \pi_n = 1$, we get

$$\beta = \delta \sum_{n=0}^{\infty} \pi_{0,n+1} = \delta \Pr\{X(t) = 0, Y(t) > 0\},$$

or

$$\Pr\{X(t) = 0, Y(t) > 0\} = \sum_{n=1}^{\infty} \pi_{0,n} = \frac{\beta}{\delta} = \tau. \tag{9.59}$$

Since (9.55) asserts that $\Pr\{X(t) = 0\} = 1 - (\alpha/\gamma) = 1 - \sigma$, we have

$$\begin{aligned} \pi_{0,0} &= \Pr\{X(t) = 0, Y(t) = 0\} = \Pr\{X(t) = 0\} - \Pr\{X(t) = 0, Y(t) > 0\} \\ &= 1 - \frac{\alpha}{\gamma} - \frac{\beta}{\delta} = 1 - \sigma - \tau \quad \text{when } \sigma + \tau < 1. \end{aligned} \tag{9.60}$$

With these preliminary results in hand, we turn to determining $M_m = \Sigma_{n=1}^{\infty} n\pi_{m,n}$. Multiplying (9.50) by n and summing, we derive

$$\begin{aligned} (\alpha + \beta + \delta)M_0 &= \gamma M_1 + \delta \sum_{n=1}^{\infty} n\pi_{0,n+1} + \beta \sum_{n=1}^{\infty} n\pi_{0,n-1} \\ &= \gamma M_1 + \delta M_0 - \delta \sum_{n=0}^{\infty} \pi_{0,n+1} + \beta M_0 + \beta \sum_{n=1}^{\infty} \pi_{0,n-1} \\ &= \gamma M_1 + \delta M_0 - \delta \left(\frac{\beta}{\delta}\right) + \beta M_0 + \beta(1 - \sigma), \end{aligned}$$

where the last line results from (9.55) and (9.59). After simplification and rearrangement, the result is

$$M_1 = \sigma M_0 + \frac{\beta}{\gamma}\sigma. \tag{9.61}$$

We next multiply (9.51) by n and sum to obtain

$$\begin{aligned} (\alpha + \beta + \gamma)M_m &= \gamma M_{m+1} + \beta \sum_{n=1}^{\infty} n\pi_{m,n-1} + \alpha M_{m-1} \\ &= \gamma M_{m+1} + \beta M_m + \beta \sum_{n=1}^{\infty} \pi_{m,n-1} + \alpha M_{m-1}. \end{aligned}$$

Again, referring to (9.55) and simplifying, we see that

$$(\alpha + \gamma)M_m = \gamma M_{m+1} + \alpha M_{m-1} + \beta(1 - \sigma)\sigma^m \quad \text{for } m = 1, 2, \ldots. \tag{9.62}$$

Equation (9.61) and (9.62) can be solved inductively to give

$$M_m = M_0\sigma^m + \frac{\beta}{\gamma} m\sigma^m \quad \text{for } m = 0, 1, \ldots,$$

which we sum to obtain

$$L_{\mathrm{n}} = \sum_{m=0}^{\infty} M_m = \frac{1}{1-\sigma}\left[M_0 + \frac{\beta}{\gamma}\frac{\sigma}{(1-\sigma)}\right]. \tag{9.63}$$

This determines L_{n} in terms of M_0. To obtain a second relation, we multiply (9.58) by n and sum to obtain

$$\begin{aligned}\beta L_{\mathrm{n}} &= \delta \sum_{n=0}^{\infty} n\pi_{0,n+1} = \delta M_0 - \delta \sum_{n=0}^{\infty} \pi_{0,n+1} \\ &= \delta M_0 - \delta\left(\frac{\beta}{\delta}\right) \quad \text{[see (9.59)]},\end{aligned}$$

or

$$M_0 = \frac{\beta}{\delta}(L_{\mathrm{n}} + 1) = \tau(L_{\mathrm{n}} + 1). \tag{9.64}$$

We substitute (9.64) into (9.63) and simplify, yielding

$$\begin{aligned} L_{\mathrm{n}} &= \frac{1}{1-\sigma}\left[\tau(L_{\mathrm{n}}+1) + \frac{\beta}{\gamma}\frac{\sigma}{1-\sigma}\right], \\ \left(1 - \frac{\tau}{1-\sigma}\right)L_{\mathrm{n}} &= \frac{1}{1-\sigma}\left[\tau + \frac{\beta}{\gamma}\frac{\sigma}{1-\sigma}\right], \end{aligned}$$

and finally,

$$L_{\mathrm{n}} = \left(\frac{\tau}{1-\sigma-\tau}\right)\left[1 + \left(\frac{\delta}{\gamma}\right)\frac{\sigma}{1-\sigma}\right]. \tag{9.65}$$

The condition that L_{n} be finite (and that a stationary distribution exist) is that

$$\rho = \sigma + \tau < 1.$$

That is, the system traffic intensity ρ must be less than one.

Since the arrival rate for nonpriority customers is β, we know that the mean waiting time for nonpriority customers is given by $W_{\mathrm{n}} = L_{\mathrm{n}}/\beta$.

Some simple numerical studies of (9.46) and (9.65) yield surprising results concerning adding priority to an existing system. Let us consider first a simple $M/M/1$ system with traffic intensity ρ whose mean queue length is given by (9.12) to be $L = \rho/(1-\rho)$. Let us propose modifying the system in such a way that a fraction $p = \frac{1}{2}$ of the customers have priority. We assume that priority is independent of service time. These assumptions lead to the values $\alpha = \beta = \frac{1}{2}\lambda$ and $\gamma = \delta = \mu$, whence

$\sigma = \tau = \rho/2$. Then, the mean queue lengths for priority and nonpriority customers are given by

$$L_{\mathrm{p}} = \frac{\sigma}{1-\sigma} = \frac{\rho/2}{1-(\rho/2)} = \frac{\rho}{2-\rho}$$

and

$$L_{\mathrm{n}} = \left(\frac{\rho/2}{1-\rho}\right)\left[1 + \frac{\rho/2}{1-(\rho/2)}\right] = \frac{\rho}{(2-\rho)(1-\rho)}.$$

The mean queue lengths L, L_{p}, and L_{n} were determined for several values of the traffic intensity ρ. The results are listed in the following table:

ρ	L	L_{p}	L_{n}
0.6	1.50	0.43	1.07
0.8	4.00	0.67	3.34
0.9	9.00	0.82	8.19
0.95	19.00	0.90	18.10

It is seen that the burden of increased queue length, as the traffic intensity increases, is carried almost exclusively by the nonpriority customers!

Exercises

9.4.1 Consider a two-server system in which an arriving customer enters the system if and only if a server is free. Suppose that customers arrive according to a Poisson process of rate $\lambda = 10$ customers per hour, and that service times are exponentially distributed with a mean service time of 6 min. In the long run, what is the rate of customers served per hour?

9.4.2 Customers arrive at a checkout station in a small grocery store according to a Poisson process of rate $\lambda = 1$ customer per minute. The checkout station can be operated with or without a bagger. The checkout times for customers are exponentially distributed, and with a bagger the mean checkout time is 30 s, while without a bagger this mean time increases to 50 s. Suppose the store's policy is to have the bagger help whenever there are two or more customers in the checkout line. In the long run, what fraction of time is the bagger helping the cashier?

9.4.3 Consider a two-server system in which an arriving customer enters the system if and only if a server is free. Suppose that customers arrive according to a Poisson process of rate $\lambda = 10$ customers per hour, and that service times are exponentially distributed. The servers have different experience in the job, and the newer server has a mean service time of 6 min, while the older has a mean service time of 4 min. In the long run, what is the rate of customers served per hour? Be explicit about any additional assumptions that you make.

9.4.4 Suppose that incoming calls to an office follow a Poisson process of rate $\lambda = 6$ per hour. If the line is in use at the time of an incoming call, the secretary has a HOLD button that will enable a single additional caller to wait. Suppose that the lengths of conversations are exponentially distributed with a mean length of 5 min, that incoming calls while a caller is on hold are lost, and that outgoing calls can be ignored. Apply the results of Section 9.4.1 to determine the fraction of calls that are lost.

Problems

9.4.1 Consider the two-server overflow queue of Section 9.4.4 and suppose the arrival rate is $\lambda = 10$ per hour. The two servers have rates 6 and 4 per hour. Recommend which server should be placed first. That is, choose between

$$\begin{matrix} \mu_1 = 6, \\ \mu_2 = 4 \end{matrix} \quad \text{and} \quad \begin{matrix} \mu_1 = 4, \\ \mu_2 = 6, \end{matrix}$$

and justify your answer. Be explicit about your criterion.

9.4.2 Consider the preemptive priority queue of Section 9.4.5 and suppose that the arrival rate is $\lambda = 4$ per hour. Two classes of customers can be identified, having mean service times of 12 min and 8 min, and it is proposed to give one of these classes priority over the other. Recommend which class should have priority. Be explicit about your criterion and justify your answer. Assume that the two classes appear in equal proportions and that all service times are exponentially distributed.

9.4.3 *Balking* refers to the refusal of an arriving customer to enter the queue. *Reneging* refers to the departure of a customer in the queue before obtaining service. Consider an $M/M/1$ system with reneging such that the probability that a specified single customer in line will depart prior to service in a short time interval $(t, t + \Delta t]$ is $r_n(\Delta t) + o(\Delta t)$ when n is the number of customers in the system. (Note that $r_0 = r_1 = 0$.) Assume Poisson arrivals at rate λ and exponential service times with parameter μ, and determine the stationary distribution when it exists.

9.4.4 A small grocery store has a single checkout counter with a full-time cashier. Customers arrive at the checkout according to a Poisson process of rate λ per hour. When there is only a single customer at the counter, the cashier works alone at a mean service rate of α per hour. Whenever there is more than one customer at the checkout, however, a "bagger" is added, increasing the service rate to β per hour. Assume that service times are exponentially distributed and determine the stationary distribution of the queue length.

9.4.5 A ticket office has two agents answering incoming phone calls. In addition, a third caller can be put on HOLD until one of the agents becomes available. If all three phone lines (both agent lines plus the hold line) are busy, a potential caller gets a busy signal, and is assumed lost. Suppose that the calls and attempted

calls occur according to a Poisson process of rate λ, and that the length of a telephone conversation is exponentially distributed with parameter μ. Determine the stationary distribution for the process.

9.5 Open Acyclic Queueing Networks

Queueing networks, composed of groups of service stations, with the departures of some stations forming the arrivals of others, arise in computer and information processing systems, manufacturing job shops, service industries such as hospitals and airport terminals, and in many other contexts. A remarkable result often enables the steady-state behavior of these complex systems to be analyzed component by component.

9.5.1 The Basic Theorem

The result alluded to in the preceding paragraph asserts that the departures from a queue with Poisson arrivals and exponentially distributed service times in statistical equilibrium also form a Poisson process. We give the precise statement as Theorem 9.1. The proof is contained in an appendix at the end of this section. See also Problem 9.2.5.

Theorem 9.1. *Let* $\{X(t), t \geq 0\}$ *be a birth and death process with constant birth parameters* $\lambda_n = \lambda$ *for* $n = 0, 1, \ldots$, *and arbitrary death parameters* μ_n *for* $n = 1, 2, \ldots$. *Suppose there exists a stationary distribution* $\pi_k \geq 0$ *where* $\sum_k \pi_k = 1$ *and that* $\Pr\{X(0) = k\} = \pi_k$ *for* $k = 0, 1, \ldots$. *Let* $D(t)$ *denote the number of deaths in* $(0, t]$. *Then*

$$\Pr\{X(t) = k, D(t) = j\} = \Pr\{X(t) = k\} \Pr\{D(t) = j\}$$
$$= \pi_k \frac{(\lambda t)^j e^{-\lambda t}}{j!} \quad \text{for } k, j \geq 0.$$

Remark The stipulated conditions are satisfied, e.g., when $X(t)$ is the number of customers in an $M/M/s$ queueing system that is in steady state wherein $\Pr\{X(0) = j\} = \pi_j$, the stationary distribution of the process. In this case, a stationary distribution exists provided that $\lambda < s\mu$, where μ is the individual service rate.

To see the major importance of this theorem, suppose that $X(t)$ represents the number of customers in some queueing system at time t. The theorem asserts that the departures form a Poisson process of rate λ. Furthermore, the number $D(t)$ of departures up to time t is independent of the number $X(t)$ of customers remaining in the system at time t.

We caution the reader that the foregoing analysis applies only if the processes are in statistical equilibrium where the stationary distribution $\pi_k = \Pr\{X(t) = k\}$ applies. In contrast, under the condition that $X(0) = 0$, then neither will the departures form a Poisson process, nor will $D(t)$ be independent of $X(t)$.

9.5.2 Two Queues in Tandem

Let us use Theorem 9.1 to analyze a simple queueing network composed of two single-server queues connected in series as shown in Figure 9.9.

Let $X_k(t)$ be the number of customers in the kth queue at time t. We assume steady state. Beginning with the first server, the stationary distribution (9.11) for a single server queue applies, and

$$\Pr\{X_1(t) = n\} = \left(1 - \frac{\lambda}{\mu_1}\right)\left(\frac{\lambda}{\mu 1}\right)^n \quad \text{for } n = 0, 1, \ldots.$$

Theorem 9.1 asserts that the departure process from the first server, denoted by $D_1(t)$, is a Poisson process of rate λ that is statistically independent of the first queue length $X_1(t)$. These departures form the arrivals to the second server, and therefore the second system has Poisson arrivals and is thus an $M/M/1$ queue as well. Thus, again using (9.11),

$$\Pr\{X_2(t) = m\} = \left(1 - \frac{\lambda}{\mu_2}\right)\left(\frac{\lambda}{\mu_2}\right)^m \quad \text{for } m = 0, 1, \ldots.$$

Furthermore, because the departures $D_1(t)$ from the first server are independent of $X_1(t)$, it must be that $X_2(t)$ is independent of $X_1(t)$. We, thus, obtain the joint distribution

$$\begin{aligned}\Pr\{X_1(t) = n \quad \text{and} \quad X_2(t) = m\} &= \Pr\{X_1(t) = n\}\Pr\{X_2(t) = m\} \\ &= \left(1 - \frac{\lambda}{\mu_1}\right)\left(\frac{\lambda}{\mu_1}\right)^n \left(1 - \frac{\lambda}{\mu_2}\right)\left(\frac{\lambda}{\mu_2}\right)^m \\ &\qquad \text{for } n, m = 0, 1, \ldots.\end{aligned}$$

We again caution the reader that the foregoing analysis applies only when the network is in its limiting distribution. In contrast, if both queues are empty at time $t = 0$,

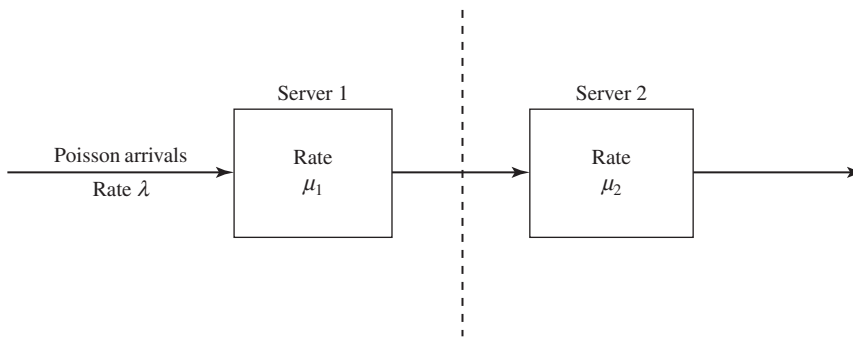


Figure 9.9 Two queues in series in which the departures from the first form the arrivals for the second.

then neither will the departures $D_1(t)$ form a Poisson process nor will $D_1(t)$ and $X_1(t)$ be independent.

9.5.3 Open Acyclic Networks

The preceding analysis of two queues in series applies to more general systems. An *open* queueing network (see Figure 9.10) has customers arriving from and departing to the outside world. (The repairman model in Chapter 6, Section 6.4 is a prototypical *closed* queueing network.) Consider an open network having K service stations, and let $X_k(t)$ be the number of customers in queue k at time t. Suppose

1. The arrivals from outside the system to distinct servers form independent Poisson processes.
2. The departures from distinct servers independently travel instantly to other servers, or leave the system, with fixed probabilities.
3. The service times for the various servers are *memoryless* in the sense that

$$\begin{aligned}&\Pr\{\text{Server \#}k \text{ completes a service in } (t, t+\Delta t] | X_k(t) = n\} \\ &\quad = \mu_{kn}(\Delta t) + o(\Delta t) \quad \text{for } n = 1, 2, \ldots,\end{aligned} \tag{9.66}$$

   and does not otherwise depend on the past.
4. The system is in statistical equilibrium (steady state).
5. The network is *acyclic* in that a customer can visit any particular server at most once. (The case where a customer can visit a server more than once is more subtle, and is treated in the next section.)

Then,

(a) $X_1(t), X_2(t), \ldots, X_K(t)$ are independent processes, where

$$\begin{aligned}&\Pr\{X_1(t) = n_1, X_2(t) = n_2, \ldots, X_K(t) = n_K\} \\ &\quad = \Pr\{X_1(t) = n_1\} \Pr\{X_2(t) = n_2\} \cdots \Pr\{X_K(t) = n_K\}.\end{aligned} \tag{9.67}$$

(b) The departure process $D_k(t)$ associated with the kth server is a Poisson process, and $D_k(t)$ and $X_k(t)$ are independent.
(c) The arrivals to the kth station form a Poisson process of rate λ_k.
(d) The departure rate at the kth server equals the rate of arrivals to that server.

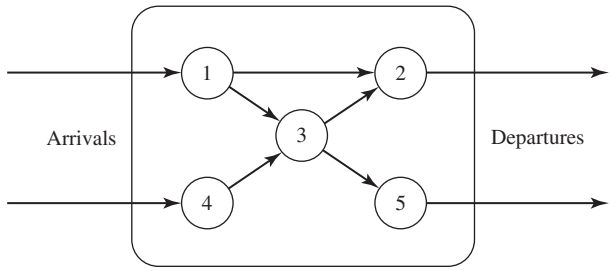


Figure 9.10 An open queueing network.

Let us add some notation so as to be able to express these results more explicitly. Let

λ_{0k} = rate of arrivals to station k from outside the system,
λ_k = rate of total arrivals to station k,
P_{kj} = probability that a customer leaving station k next visits station j.

Then, the arrivals to station k come from outside the system or from some other station j. The departure rate from j equals the arrival rate to j, the fraction P_{jk} of which go to station k, whence

$$\lambda_k = \lambda_{0k} + \sum_j \lambda_j P_{jk}. \tag{9.68}$$

Since the network is acyclic, (9.68) may be solved recursively, beginning with stations having only outside arrivals. The simple example that follows will make the procedure clear.

The arrivals to station k form a Poisson process of rate λ_k. Let

$$\psi_k(n) = \pi_{k0} \times \frac{\lambda_k^n}{\mu_{k1}\mu_{k2}\cdots\mu_{kn}} \quad \text{for } n = 1, 2, \ldots, \tag{9.69}$$

where

$$\psi_k(0) = \pi_{k0} = \left\{1 + \sum_{n=1}^{\infty}\left(\frac{\lambda_k^n}{\mu_{k1}\mu_{k2}\cdots\mu_{kn}}\right)\right\}^{-1}. \tag{9.70}$$

Referring to (9.40) and (9.41) we see that (9.69) and (9.70) give the stationary distribution for a queue having Poisson arrivals at rate λ_k and memoryless service times at rates μ_{kn} for $n = 1, 2, \ldots$. Accordingly, we may now express (9.67) explicitly as

$$\begin{aligned}\Pr\{X_1(t) = n_1, X_2(t) = n_2, \ldots, X_K = n_k\} \\ = \psi_1(n_1)\psi_2(n_2)\cdots\psi_K(n_K).\end{aligned} \tag{9.71}$$

Example Consider the three-station network as shown in Figure 9.11.

The first step in analyzing the example is to determine the arrival rates at the various stations. In equilibrium, the arrival rate at a station must equal its departure rate, as asserted in (d). Accordingly, departures from state 1 occur at rate $\lambda_1 = 4$, and since these departures independently travel to stations 2 and 3 with respective probabilities $P_{12} = \frac{1}{3}$ and $P_{13} = \frac{2}{3}$, we determine the arrival rate $\lambda = \left(\frac{1}{3}\right)4$. At station 3 the arrivals include both those from station 1 and those from station 2. Thus, $\lambda_3 = \left(\frac{2}{3}\right)4 + \left(\frac{1}{3}\right)4 = 4$.

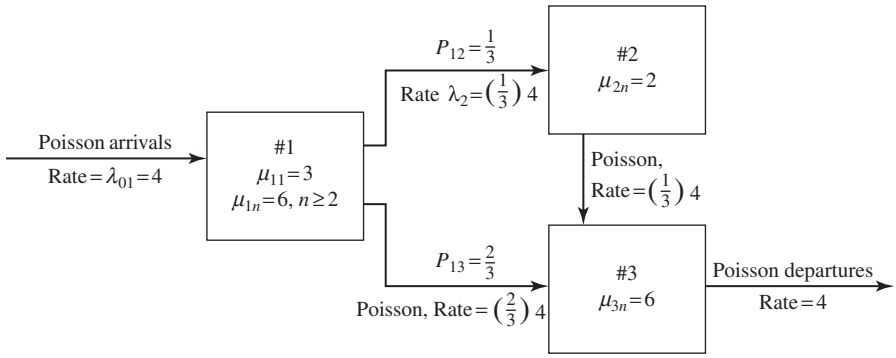


Figure 9.11 A three-station open acyclic network. Two servers, each of rate 3, at the first station give rise to the station rates $\mu_{11} = 3$ and $\mu_{1n} = 6$ for $n \geq 2$. Stations 2 and 3 each have a single server of rate 2 and 6, respectively.

Having determined the arrival rates at each station, we turn to determining the equilibrium probabilities. Station 1 is an $M/M/2$ system with $\lambda = 4$ and $\mu = 3$. From (9.18), or (9.69) and (9.70), we obtain

$$\Pr\{X_1(t) = 0\} = \pi_0 = \left\{1 + \left(\frac{4}{3}\right) + \frac{(4/3)^2}{2(1/3)}\right\}^{-1} = 0.2$$

and

$$\Pr\{X_1(t) = n\} = \begin{cases} \left(\frac{4}{3}\right)(0.2) & \text{for } n = 1, \\ (0.4)\left(\frac{2}{3}\right)^n & \text{for } n \geq 2. \end{cases}$$

Station 2 is an $M/M/1$ system with $\lambda = \frac{4}{3}$ and $\mu = 2$. From (9.11) we obtain

$$\Pr\{X_2(t) = n\} = \left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^n \quad \text{for } n = 0, 1, \ldots.$$

Similarly station 3 is an $M/M/1$ system with $\lambda = 4$ and $\mu = 6$, so that (9.11) yields

$$\Pr\{X_3(t) = n\} = \left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^n \quad \text{for } n = 0, 1, \ldots.$$

Finally, according to Property (a), the queue lengths $X_1(t)$, $X_2(t)$, and $X_3(t)$ are independent, so that

$$\begin{aligned} &\Pr\{X_1(t) = n_1, X_2(t) = n_2, X_3(t) = n_3\} \\ &\quad = \Pr\{X_1(t) = n_1\}\Pr\{X_2(t) = n_2\}\Pr\{X_3(t) = n_3\}. \end{aligned}$$

9.5.4 Appendix: Time Reversibility

Let $\{X(t), -\infty < t < +\infty\}$ be an arbitrary countable-state Markov chain having a stationary distribution $\pi_j = \Pr\{X(t) = j\}$ for all states j and all times t. Note that the time index set is the whole real line. We view the process as having begun indefinitely far in the past, so that it now is evolving in a stationary manner. Let $Y(t) = X(-t)$ be the same process, but with time reversed. The stationary process $\{X(t)\}$ is said to be *time reversible* if $\{X(t)\}$ and $\{Y(t)\}$ have the same probability laws. Clearly, $\Pr\{X(0) = j\} = \Pr\{Y(0) = j\} = \pi_j$, and it is not difficult to show that both processes are Markov. Hence, in order to show that they share the same probability laws it suffices to show that they have the same transition probabilities. Let

$$P_{ij}(t) = \Pr\{X(t) = j | X(0) = i\},$$
$$Q_{ij}(t) = \Pr\{Y(t) = j | Y(0) = i\}.$$

The process $\{X(t)\}$ is reversible if

$$P_{ij}(t) = Q_{ij}(t) \tag{9.72}$$

for all states i, j and all times t. We evaluate $Q_{ij}(t)$ as follows:

$$\begin{aligned} Q_{ij}(t) &= \Pr\{Y(t) = j | Y(0) = i\} \\ &= \Pr\{X(-t) = j | X(0) = i\} \\ &= \Pr\{X(0) = j | X(t) = i\} \quad \text{(by stationarity)} \\ &= \frac{\Pr\{X(0) = j, X(t) = i\}}{\Pr\{X(t) = i\}} \\ &= \frac{\pi_j P_{ji}(t)}{\pi_i}. \end{aligned}$$

In conjunction with (9.72) we see that the process $\{X(t)\}$ is reversible if

$$P_{ij}(t) = Q_{ij}(t) = \frac{\pi_j P_{ji}(t)}{\pi_i},$$

or

$$\pi_i P_{ij}(t) = \pi_j P_{ji}(t), \tag{9.73}$$

for all states i, j and all times t.

As a last step, we determine a criterion for reversibility in terms of the infinitesimal parameters

$$a_{ij} = \lim_{t \downarrow 0} \frac{1}{t} \Pr\{X(t) = j | X(0) = i\}, \quad i \neq j.$$

It is immediate that (9.73) holds when $i = j$. When $i \neq j$,

$$P_{ij}(t) = a_{ij}t + o(t), \tag{9.74}$$

which substituted into (9.73) gives

$$\pi_i[a_{ij}t + o(t)] = \pi_j[a_{ji}t + o(t)],$$

and after dividing by t and letting t vanish, we obtain the criterion

$$\pi_i a_{ij} = \pi_j a_{ji} \quad \text{for all } i \neq j. \tag{9.75}$$

When the transition probabilities are determined by the infinitesimal parameters, we deduce that the process $\{X(t)\}$ is time reversible whenever (9.75) holds.

All birth and death processes satisfying Chapter 6, (6.21) and having stationary distributions are time reversible! Because birth and death processes have

$$a_{i,i+1} = \lambda_i,$$
$$a_{i,i-1} = \mu_i,$$

and

$$a_{i,j} = 0 \quad \text{if } |i-j| > 1,$$

in verifying (9.75) it suffices to check that

$$\pi_i a_{i,i+1} = \pi_{i+1} a_{i+1,i},$$

or

$$\pi_i \lambda_i = \pi_{i+1}\mu_{i+1} \quad \text{for } i = 0, 1, \ldots. \tag{9.76}$$

But [see Chapter 6, equations (6.36) and (6.73)],

$$\pi_i = \pi_0 \left(\frac{\lambda_0 \lambda_1 \cdots \lambda_{i-1}}{\mu_1 \mu_2 \cdots \mu_i} \right) \quad \text{for } i = 1, 2, \ldots,$$

whence (9.76) becomes

$$\pi_0 \left(\frac{\lambda_0 \lambda_1 \cdots \lambda_{i-1}}{\mu_1 \mu_2 \cdots \mu_i} \right) \lambda_i = \pi_0 \left(\frac{\lambda_0 \lambda_1 \cdots \lambda_i}{\mu_1 \mu_2 \cdots \mu_{i+1}} \right) \mu_{i+1},$$

which is immediately seen to be true.

9.5.5 Proof of Theorem 9.1

Let us consider a birth and death process $\{X(t)\}$ having the constant birth rate $\lambda_k = \lambda$ for $k = 0, 1, \ldots$ and arbitrary death parameters $\mu_k > 0$ for $k = 1, 2, \ldots$. This process corresponds to a memoryless server queue having Poisson arrivals. A typical evolution is illustrated in Figure 9.12. The arrival process for $\{X(t)\}$ is a Poisson process of rate λ. The reversed time process $Y(t) = X(-t)$ has the same probabilistic laws as does $\{X(t)\}$, so the arrival process for $\{Y(t)\}$ also must be a Poisson process of rate λ. But the arrival process for $\{Y(t)\}$ is the departure process for $\{X(t)\}$ (see Figure 9.12). Thus, it must be that these departure instants also form a Poisson process of rate λ. In particular, if $D(t)$ counts the departures in the $X(\cdot)$ process over the duration $(0, t]$, then

$$\Pr\{D(t) = j\} = \frac{(\lambda t)^j e^{-\lambda t}}{j!} \quad \text{for } j = 0, 1, \ldots. \tag{9.77}$$

Moreover, looking at the reversed process $Y(-t) = X(t)$, the "future" arrivals for $Y(-t)$ in the Y duration $[-t, 0)$ are independent of $Y(-t) = X(t)$. (See Figure 9.12.) These future arrivals for $Y(-t)$ are the departures for $X(\cdot)$ in the interval $(0, t]$. Therefore, these departures and $X(t) = Y(-t)$ must be independent. Since $\Pr\{X(t) = k\} = \pi_k$, by the assumption of stationarity, the independence of $D(t)$ and $X(t)$ and (9.77) give

$$\begin{aligned}\Pr\{X(t) = k, D(t) = j\} &= \Pr\{X(t) = k\}\Pr\{D(t) = j\} \\ &= \frac{\pi_k e^{-\lambda t}(\lambda t)^j}{j!},\end{aligned}$$

and the proof of Theorem 9.1 is complete.

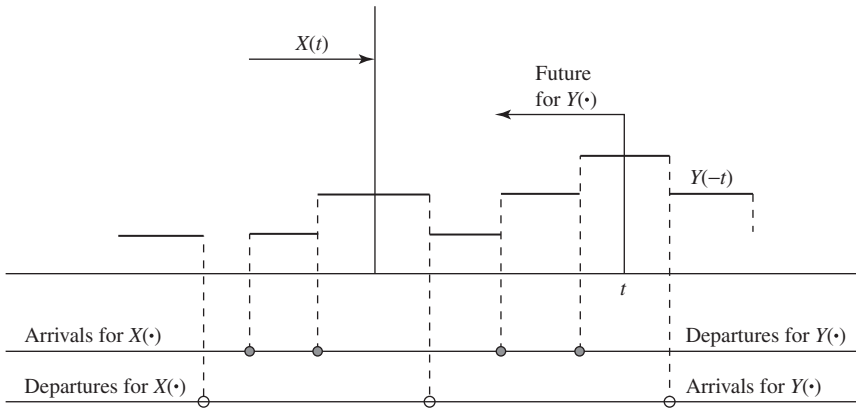
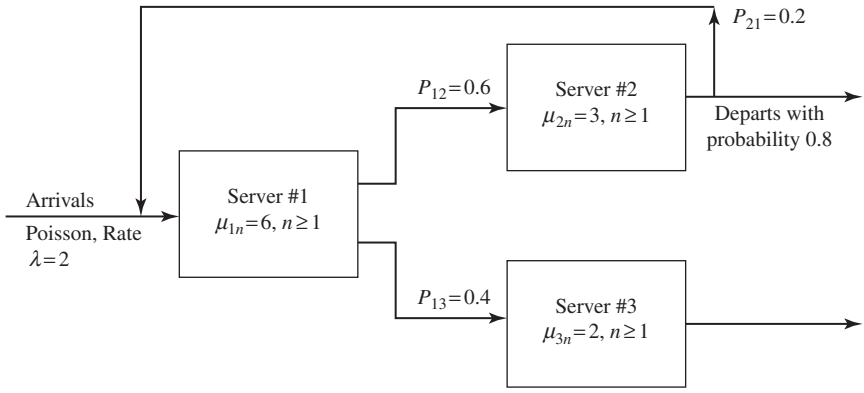


Figure 9.12 A typical evolution of a queueing process. The instants of arrivals and departures have been isolated on two time axes below the graph.

Exercises

9.5.1 Consider the three-server network pictured here:



In the long run, what fraction of time is server #2 idle while, simultaneously, server #3 is busy? Assume that all service times are exponentially distributed.

9.5.2 Refer to the network of Exercise 9.5.1. Suppose that server #2 and server #3 share a common customer waiting area. If it is desired that the total number of customers being served and waiting to be served not exceed the waiting area capacity more than 5% of the time in the long run, how large should this area be?

Problem

9.5.1 Suppose three service stations are arranged in tandem so that the departures from one form the arrivals for the next. The arrivals to the first station are a Poisson process of rate $\lambda = 10$ per hour. Each station has a single server, and the three service rates are $\mu_1 = 12$ per hour, $\mu_2 = 20$ per hour, and $\mu_3 = 15$ per hour. In-process storage is being planned for station 3. What capacity C_3 must be provided if in the long run, the probability of exceeding C_3 is to be less than or equal to 1%? That is, what is the smallest number $C_3 = c$ for which $\lim_{t\to\infty} \Pr\{X_3(t) > c\} \leq 0.01$?

9.6 General Open Networks

The preceding section covered certain memoryless queueing networks in which a customer could visit any particular server at most once. With this assumption, the departures from any service station formed a Poisson process that was independent of the number of customers at that station in steady state. As a consequence, the numbers

$X_1(t), X_2(t), \ldots, X_K(t)$ of customers at the K stations were independent random variables, and the product form solution expressed in (9.67) prevailed.

The situation where a customer can visit a server more than once is more subtle. On the one hand, many flows in the network are no longer Poisson. On the other hand, rather surprisingly, the product form solution of (9.67) remains valid.

Example To begin our explanation, let us first reexamine the simple feedback model of Section 9.4.3. The flow is depicted in Figure 9.13. The arrival process is Poisson, but the input to the server is not. (The distinction between the arrival and input processes is made in Figures 9.2 and 9.6.) The output process, as shown in Figure 9.13, is not Poisson, nor is it independent of the number of customers in the system. Recall that each customer in the output is fed back with probability p and departs with probability $q = 1 - p$. In view of this non-Poisson behavior, it is remarkable that the distribution of the number of customers in the system is the same as that in a Poisson $M/M/1$ system whose input rate is λ/q and whose service rate is μ, as verified in (9.42).

Example Let us verify the product form solution in a slightly more complex two-server network, depicted in Figure 9.14.

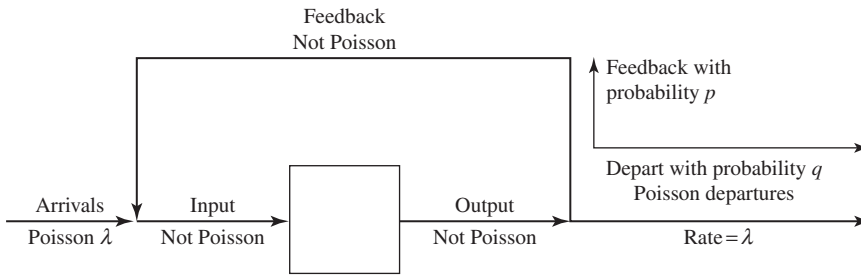


Figure 9.13 A single server with feedback.

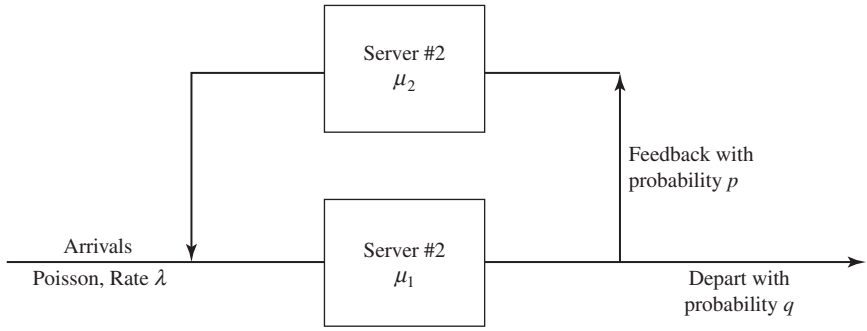


Figure 9.14 A two-server feedback system. For example, server #2 in this system might be an inspector returning a fraction p of the output for rework.

If we let $X_j(t)$ denote the number of customers at station i at time t, for $i = 1, 2$, then $\mathbf{X}(t) = [X_1(t), X_2(t)]$ is a Markov chain whose transition rates are given in the following table:

From State	To State	Transition Rate	Description
(m, n)	$(m+1, n)$	λ	Arrival of new customer
(m, n) $n \geq 1$	$(m+1, n-1)$	μ_2	Input of feedback customer
(m, n) $m \geq 1$	$(m-1, n)$	$q\mu_1$	Departure of customer
(m, n) $m \geq 1$	$(m-1, n+1)$	$p\mu_1$	Feedback to server #2

Let $\pi_{m,n} = \lim_{t\to\infty} \Pr\{X_1(t) = m, X_2(t) = n\}$ be the stationary distribution of the process. Reasoning analogous to that of (6.68) and (6.69) of Chapter 6 (where the theory was developed for finite-state Markov chains) leads to the following equations for the stationary distribution:

$$\lambda\pi_{0,0} = q\mu_1\pi_{1,0}, \tag{9.78}$$

$$(\lambda + \mu_2)\pi_{0,n} = p\mu_1\pi_{1,n-1} + q\mu_1\pi_{1,n}, \quad n \geq 1, \tag{9.79}$$

$$(\lambda + \mu_1)\pi_{m,0} = \lambda\pi_{m-1,0} + q\mu_1\pi_{m+1,0} + \mu_2\pi_{m-1,1}, \quad m \geq 1, \tag{9.80}$$

$$\begin{aligned}(\lambda + \mu_1 + \mu_2)\pi_{m,n} = {} & \lambda\pi_{m-1,n} + p\mu_1\pi_{m+1,n-1} + q\mu_1\pi_{m+1,n} \\ & + \mu_2\pi_{m-1,n+1}, m, \quad n \geq 1.\end{aligned} \tag{9.81}$$

The mass balance interpretation as explained following (6.69) in Chapter 6 may help motivate (9.78) through (9.81). For example, the left side in (9.78) measures the total rate of flow out of state $(0, 0)$ and is jointly proportional to $\pi_{0,0}$, the long run fraction of time the process is in state $(0, 0)$, and λ, the (conditional) transition rate out of $(0, 0)$. Similarly, the right side of (9.78) measures the total rate of flow into state $(0, 0)$.

Using the product form solution in the acyclic case, we will "guess" a solution and then verify that our guess indeed satisfies (9.78) through (9.81). First we need to determine the input rate, call it λ_1, to server #1. In equilibrium, the output rate must equal the input rate, and of this output, the fraction p is returned to join the new arrivals after visiting server #2. We have

$$\text{Input Rate} = \text{New Arrivals} + \text{Feedback}.$$

which translates into

$$\lambda_1 = \lambda + p\lambda_1,$$

or

$$\lambda_1 = \frac{\lambda}{1-p} = \frac{\lambda}{q}. \tag{9.82}$$

The input rate to server #2 is

$$\lambda_2 = p\lambda_1 = \frac{p\lambda}{q}. \tag{9.83}$$

The solution that we guess is to treat server #1 and server #2 as independent $M/M/1$ systems having input rates λ_1 and λ_2, respectively (even though we know from our earlier discussion that the input to server #2, while of rate λ_2, is not Poisson). That is, we attempt a solution of the form

$$\begin{aligned}\pi_{m,n} &= \left(1-\frac{\lambda_1}{\mu_1}\right)\left(\frac{\lambda_1}{\mu_1}\right)^m \left(1-\frac{\lambda_2}{\mu_2}\right)\left(\frac{\lambda_2}{\mu_2}\right)^n \\ &= \left(1-\frac{\lambda}{q\mu_1}\right)\left(\frac{\lambda}{q\mu_1}\right)^m \left(1-\frac{p\lambda}{q\mu_2}\right)\left(\frac{p\lambda}{q\mu_2}\right)^n \quad \text{for } m,n \geq 1.\end{aligned}$$

It is immediate that

$$\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\pi_{m,n} = 1,$$

provided that $\lambda_1 = (\lambda/q) < \mu_1$ and $\lambda_2 = p\lambda/q < \mu_2$.

We turn to verifying (9.78) through (9.81). Let $\theta_{m,n} = (\lambda/q\mu_1)^m \times (p\lambda/q\mu_2)^n$. It suffices to verify that $\theta_{m,n}$ satisfies (9.78) through (9.81), since $\pi_{m,n}$ and $\theta_{m,n}$ differ only by the constant multiple $\pi_{0,0} = (1-\lambda_1/\mu_1) \times (1-\lambda_2/\mu_2)$. Thus, we proceed to substitute $\theta_{m,n}$ into (9.78) through (9.81) and verify that equality is obtained.
We verify (9.78):

$$\lambda = q\mu_1\left(\frac{\lambda}{q\mu_1}\right) = \lambda.$$

We verify (9.79):

$$(\lambda+\mu_2)\left(\frac{p\lambda}{q\mu_2}\right)^n = p\mu_1\left(\frac{\lambda}{q\mu_1}\right)\left(\frac{p\lambda}{q\mu_2}\right)^{n-1} + q\mu_1\left(\frac{\lambda}{q\mu_1}\right)\left(\frac{p\lambda}{q\mu_2}\right)^n,$$

or after dividing by $(p\lambda/q\mu_2)^n$ and simplifying,

$$\lambda+\mu_2 = \left(\frac{p\lambda}{q}\right)\left(\frac{q\mu_2}{p\lambda}\right) + \lambda = \lambda+\mu_2.$$

We verify (9.80):

$$(\lambda+\mu_1)\left(\frac{\lambda}{q\mu_1}\right)^m = \lambda\left(\frac{\lambda}{q\mu_1}\right)^{m-1} + q\mu_1\left(\frac{\lambda}{q\mu_1}\right)^{m+1} + \mu_2\left(\frac{\lambda}{q\mu_1}\right)^{m-1}\left(\frac{p\lambda}{q\mu_2}\right),$$

which, after dividing by $(\lambda/q\mu_1)^m$, becomes

$$(\lambda+\mu_1) = \lambda\left(\frac{q\mu_1}{\lambda}\right) + q\mu_1\left(\frac{\lambda}{q\mu_1}\right) + \mu_2\left(\frac{q\mu_1}{\lambda}\right)\left(\frac{p\lambda}{q\mu_2}\right),$$

or

$$\lambda+\mu_1 = q\mu_1+\lambda+p\mu_1 = \lambda+\mu_1.$$

The final verification, that $\theta_{m,n}$ satisfies (9.81), is left to the reader as Exercise 9.6.1.

9.6.1 The General Open Network

Consider an open queueing network having K service stations, and let $X_k(t)$ denote the number of customers at station k at time t. We assume that

1. The arrivals from outside the network to distinct servers form independent Poisson processes, where the outside arrivals to station k occur at rate λ_{0k}.
2. The departures from distinct servers independently travel instantly to other servers, or leave the system, with fixed probabilities, where the probability that a departure from station j travels to station k is P_{jk}.
3. The service times are *memoryless,* or *Markov*, in the sense that

$$\begin{aligned}\Pr\{&\text{Server \#}k \text{ completes a service in } (t, t+\Delta t] | X_k(t) = n\} \\ &= \mu_{kn}(\Delta t) + o(\Delta t) \quad \text{for } n = 1, 2, \ldots,\end{aligned} \tag{9.84}$$

   and does not otherwise depend on the past.
4. The system is in statistical equilibrium (stationary).
5. The system is completely open in that all customers in the system eventually leave.

Let λ_k be the rate of input at station k. The input at station k is composed of customers entering from outside the system, at rate λ_{0k}, plus customers traveling from (possibly) other stations. The input to station k from station j occurs at rate $\lambda_j P_{jk}$, whence, as in (9.68),

$$\lambda_k = \lambda_{0k} + \sum_{j=1}^{K} \lambda_j P_{jk} \quad \text{for } k = 1, \ldots, K. \tag{9.85}$$

Condition 5 above, that all entering customers eventually leave, ensures that (9.85) has a unique solution.

With $\lambda_1, \ldots, \lambda_K$ given by (9.86), the main result is the product form solution

$$\begin{aligned}\Pr\{X_1(t) = n_1, X_2(t) = n_2, \ldots, X_k(t) = n_K\} \\ = \psi_1(n_1)\psi_2(n_2)\cdots\psi_K(n_K),\end{aligned} \tag{9.86}$$

where

$$\psi_k(n) = \frac{\pi_{k0}\lambda_k^n}{\mu_{k1}\mu_{k2}\cdots\mu_{kn}} \quad \text{for } n = 1, 2, \ldots, \tag{9.87}$$

and

$$\psi_k(0) = \pi_{k0} = \left\{1 + \sum_{n=1}^{\infty} \frac{\lambda_k^n}{\mu_{k1}\mu_{k2}\cdots\mu_{kn}}\right\} - 1. \tag{9.88}$$

Example The example of Figure 9.13 (see also Section 9.4.3) corresponds to $K = 1$ (a single service station) for which $P_{11} = p < 1$. The external arrivals are at rate $\lambda_{01} = \lambda$, and (9.86) becomes

$$\lambda_1 = \lambda_{01} + \lambda_1 P_{11}, \quad \text{or} \quad \lambda_1 = \lambda + \lambda_1 p,$$

which gives $\lambda_1 = \lambda/(1-p) = \lambda/q$. Since the example concerns a single server, then $\mu_{1n} = \mu$ for all n, and (9.88) becomes

$$\psi_1(n) = \pi_{10}\left(\frac{\lambda_1}{\mu}\right)^n = \pi_{10}\left(\frac{\lambda}{q\mu}\right)^n,$$

where

$$\pi_{10} = \left(1 - \frac{\lambda}{q\mu}\right),$$

in agreement with (9.42).

Example Consider next the two-server example depicted in Figure 9.14. The data given there furnish the following information:

$$\begin{aligned}&\lambda_{01} = \lambda, \quad \lambda_{02} = 0,\\ &P_{11} = 0, \quad P_{12} = p,\\ &P_{21} = 1, \quad P_{22} = 0,\end{aligned}$$

which substituted into (9.86) gives

$$\begin{aligned}\lambda_1 &= \lambda + \lambda_2(1),\\ \lambda_2 &= 0 + \lambda_1(p),\end{aligned}$$

which readily yields

$$\lambda_1 = \frac{\lambda}{q} \quad \text{and} \quad \lambda_2 = \frac{p\lambda}{q},$$

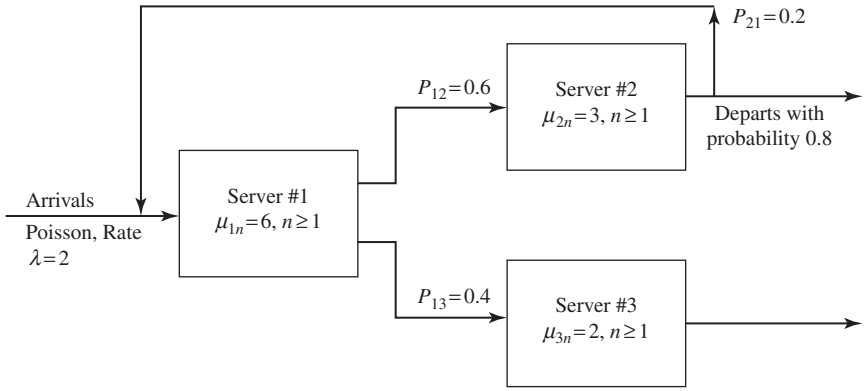
in agreement with (9.82) and (9.83). It is readily seen that the product solution of (9.86) through (9.88) is identical with (9.84), which was directly verified as the solution in this example.

Exercise

9.6.1 In the case $m \geq 1, n \geq 1$, verify that $\theta_{m,n}$ as given following (9.84) satisfies the equation for the stationary distribution (9.81).

Problem

9.6.1 Consider the three-server network pictured here:



In the long run, what fraction of the time is server #2 idle while, simultaneously, server #3 is busy? Assume that the system satisfies assumptions (1) through (5) of a general open network.