Queueing Theory

8.1 Introduction

In this chapter we will study a class of models in which customers arrive in some random manner at a service facility. Upon arrival they are made to wait in queue until it is their turn to be served. Once served they are generally assumed to leave the system. For such models we will be interested in determining, among other things, such quantities as the average number of customers in the system (or in the queue) and the average time a customer spends in the system (or spends waiting in the queue).

In Section 8.2 we derive a series of basic queueing identities that are of great use in analyzing queueing models. We also introduce three different sets of limiting probabilities that correspond to what an arrival sees, what a departure sees, and what an outside observer would see.

In Section 8.3 we deal with queueing systems in which all of the defining probability distributions are assumed to be exponential. For instance, the simplest such model is to assume that customers arrive in accordance with a Poisson process (and thus the interarrival times are exponentially distributed) and are served one at a time by a single server who takes an exponentially distributed length of time for each service. These exponential queueing models are special examples of continuous-time Markov chains and so can be analyzed as in Chapter 6. However, at the cost of a (very) slight amount of repetition we shall not assume that you are familiar with the material of Chapter 6, but rather we shall redevelop any

Introduction to Probability Models, ISBN: 9780123756862

needed material. Specifically we shall derive anew (by a heuristic argument) the formula for the limiting probabilities.

In Section 8.4 we consider models in which customers move randomly among a network of servers. The model of Section 8.4.1 is an open system in which customers are allowed to enter and depart the system, whereas the one studied in Section 8.4.2 is closed in the sense that the set of customers in the system is constant over time.

In Section 8.5 we study the model $M/G/1$, which while assuming Poisson arrivals, allows the service distribution to be arbitrary. To analyze this model we first introduce in Section 8.5.1 the concept of work, and then use this concept in Section 8.5.2 to help analyze this system. In Section 8.5.3 we derive the average amount of time that a server remains busy between idle periods.

In Section 8.6 we consider some variations of the model $M/G/1$. In particular in Section 8.6.1 we suppose that bus loads of customers arrive according to a Poisson process and that each bus contains a random number of customers. In Section 8.6.2 we suppose that there are two different classes of customers—with type 1 customers receiving service priority over type 2.

In Section 8.6.3 we present an $M/G/1$ optimization example. We suppose that the server goes on break whenever she becomes idle, and then determine, under certain cost assumptions, the optimal time for her to return to service.

In Section 8.7 we consider a model with exponential service times but where the interarrival times between customers is allowed to have an arbitrary distribution. We analyze this model by use of an appropriately defined Markov chain. We also derive the mean length of a busy period and of an idle period for this model.

In Section 8.8 we consider a single-server system whose arrival process results from return visits of a finite number of possible sources. Assuming a general service distribution, we show how a Markov chain can be used to analyze this system.

In the final section of the chapter we talk about multiserver systems. We start with loss systems, in which arrivals finding all servers busy are assumed to depart and as such are lost to the system. This leads to the famous result known as Erlang's loss formula, which presents a simple formula for the number of busy servers in such a model when the arrival process in Poisson and the service distribution is general. We then discuss multiserver systems in which queues are allowed. However, except in the case where exponential service times are assumed, there are very few explicit formulas for these models. We end by presenting an approximation for the average time a customer waits in queue in a k-server model that assumes Poisson arrivals but allows for a general service distribution.

8.2 Preliminaries

In this section we will derive certain identities that are valid in the great majority of queueing models.

8.2.1 Cost Equations

Some fundamental quantities of interest for queueing models are

L,	the average number of customers in the system;
L_Q,	the average number of customers waiting in queue;
W,	the average amount of time a customer spends in the system;
W_Q,	the average amount of time a customer spends waiting in queue.

A large number of interesting and useful relationships between the preceding and other quantities of interest can be obtained by making use of the following idea: Imagine that entering customers are forced to pay money (according to some rule) to the system. We would then have the following basic cost identity:

$$\begin{aligned}&\text{average rate at which the system earns}\\ &\quad = \lambda_a \times \text{ average amount an entering customer pays}\end{aligned} \tag{8.1}$$

where λ_a is defined to be average arrival rate of entering customers. That is, if $N(t)$ denotes the number of customer arrivals by time t, then

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

We now present a heuristic proof of Equation (8.1).

Heuristic Proof of Equation (8.1). Let T be a fixed large number. In two different ways, we will compute the average amount of money the system has earned by time T. On one hand, this quantity approximately can be obtained by multiplying the average rate at which the system earns by the length of time T. On the other hand, we can approximately compute it by multiplying the average amount paid by an entering customer by the average number of customers entering by time T (this latter factor is approximately $\lambda_a T$). Hence, both sides of Equation (8.1) when multiplied by T are approximately equal to the average amount earned by T. The result then follows by letting $T \to \infty$.*

By choosing appropriate cost rules, many useful formulas can be obtained as special cases of Equation (8.1). For instance, by supposing that each customer pays $1 per unit time while in the system, Equation (8.1) yields the so-called Little's formula,

$$L = \lambda_a W \tag{8.2}$$

This follows since, under this cost rule, the rate at which the system earns is just the number in the system, and the amount a customer pays is just equal to its time in the system.

* This can be made into a rigorous proof provided we assume that the queueing process is regenerative in the sense of Section 7.5. Most models, including all the ones in this chapter, satisfy this condition.

Similarly if we suppose that each customer pays \$1 per unit time while in queue, then Equation (8.1) yields

$$L_Q = \lambda_a W_Q \tag{8.3}$$

By supposing the cost rule that each customer pays \$1 per unit time while in service we obtain from Equation (8.1) that the

$$\text{average number of customers in service} = \lambda_a E[S] \tag{8.4}$$

where $E[S]$ is defined as the average amount of time a customer spends in service.

It should be emphasized that Equations (8.1) through (8.4) are valid for almost all queueing models regardless of the arrival process, the number of servers, or queue discipline. ■

8.2.2 Steady-State Probabilities

Let $X(t)$ denote the number of customers in the system at time t and define $P_n, n \geqslant 0$, by

$$P_n = \lim_{t\to\infty} P\{X(t) = n\}$$

where we assume the preceding limit exists. In other words, P_n is the limiting or long-run probability that there will be exactly n customers in the system. It is sometimes referred to as the *steady-state probability* of exactly n customers in the system. It also usually turns out that P_n equals the (long-run) proportion of time that the system contains exactly n customers. For example, if $P_0 = 0.3$, then in the long run, the system will be empty of customers for 30 percent of the time. Similarly, $P_1 = 0.2$ would imply that for 20 percent of the time the system would contain exactly one customer.*

Two other sets of limiting probabilities are $\{a_n, n \geqslant 0\}$ and $\{d_n, n \geqslant 0\}$, where

$$\begin{aligned} a_n &= \text{proportion of customers that find } n \\ &\quad\ \text{in the system when they arrive, and} \\ d_n &= \text{proportion of customers leaving behind } n \\ &\quad\ \text{in the system when they depart} \end{aligned}$$

That is, P_n is the proportion of time during which there are n in the system; a_n is the proportion of arrivals that find n; and d_n is the proportion of departures that leave behind n. That these quantities need not always be equal is illustrated by the following example.

* A sufficient condition for the validity of the dual interpretation of P_n is that the queueing process be regenerative.

Example 8.1 Consider a queueing model in which all customers have service times equal to 1, and where the times between successive customers are always greater than 1 (for instance, the interarrival times could be uniformly distributed over (1, 2)). Hence, as every arrival finds the system empty and every departure leaves it empty, we have

$$a_0 = d_0 = 1$$

However,

$$P_0 \neq 1$$

as the system is not always empty of customers. ■

It was, however, no accident that a_n equaled d_n in the previous example. That arrivals and departures always see the same number of customers is always true as is shown in the next proposition.

Proposition 8.1 In any system in which customers arrive and depart one at a time

$$\text{the rate at which arrivals find } n = \text{the rate at which departures leave } n$$

and

$$a_n = d_n$$

Proof. An arrival will see n in the system whenever the number in the system goes from n to $n + 1$; similarly, a departure will leave behind n whenever the number in the system goes from $n + 1$ to n. Now in any interval of time T the number of transitions from n to $n + 1$ must equal to within 1 the number from $n + 1$ to n. (Between any two transitions from n to $n + 1$, there must be one from $n + 1$ to n, and conversely.) Hence, the rate of transitions from n to $n + 1$ equals the rate from $n + 1$ to n; or, equivalently, the rate at which arrivals find n equals the rate at which departures leave n. Now a_n, the proportion of arrivals finding n, can be expressed as

$$a_n = \frac{\text{the rate at which arrivals find } n}{\text{overall arrival rate}}$$

Similarly,

$$d_n = \frac{\text{the rate at which departures leave } n}{\text{overall departure rate}}$$

Thus, if the overall arrival rate is equal to the overall departure rate, then the preceding shows that $a_n = d_n$. On the other hand, if the overall arrival rate exceeds the overall departure rate, then the queue size will go to infinity, implying that $a_n = d_n = 0$. ■

Hence, on the average, arrivals and departures always see the same number of customers. However, as Example 8.1 illustrates, they do not, in general, see time averages. One important exception where they do is in the case of Poisson arrivals.

Proposition 8.2 Poisson arrivals always see time averages. In particular, for Poisson arrivals,

$$P_n = a_n$$

To understand why Poisson arrivals always see time averages, consider an arbitrary Poisson arrival. If we knew that it arrived at time t, then the conditional distribution of what it sees upon arrival is the same as the unconditional distribution of the system state at time t. For knowing that an arrival occurs at time t gives us no information about what occurred prior to t. (Since the Poisson process has independent increments, knowing that an event occurred at time t does not affect the distribution of what occurred prior to t.) Hence, an arrival would just see the system according to the limiting probabilities.

Contrast the foregoing with the situation of Example 8.1 where knowing that an arrival occurred at time t tells us a great deal about the past; in particular it tells us that there have been no arrivals in $(t-1, t)$. Thus, in this case, we cannot conclude that the distribution of what an arrival at time t observes is the same as the distribution of the system state at time t.

For a second argument as to why Poisson arrivals see time averages, note that the total time the system is in state n by time T is (roughly) P_nT. Hence, as Poisson arrivals always arrive at rate λ no matter what the system state, it follows that the number of arrivals in $[0, T]$ that find the system in state n is (roughly) λP_nT. In the long run, therefore, the rate at which arrivals find the system in state n is λP_n and, as λ is the overall arrival rate, it follows that $\lambda P_n/\lambda = P_n$ is the proportion of arrivals that find the system in state n.

The result that Poisson arrivals see time averages is called the *PASTA* principle.

8.3 Exponential Models

8.3.1 A Single-Server Exponential Queueing System

Suppose that customers arrive at a single-server service station in accordance with a Poisson process having rate λ. That is, the times between successive arrivals are independent exponential random variables having mean $1/\lambda$. Each customer, upon arrival, goes directly into service if the server is free and, if not, the customer joins the queue. When the server finishes serving a customer, the customer leaves the system, and the next customer in line, if there is any, enters service.

The successive service times are assumed to be independent exponential random variables having mean $1/\mu$.

The preceding is called the $M/M/1$ queue. The two Ms refer to the fact that both the interarrival and the service distributions are exponential (and thus memoryless, or Markovian), and the 1 to the fact that there is a single server. To analyze it, we shall begin by determining the limiting probabilities P_n, for $n = 0, 1, \ldots.$ To do so, think along the following lines. Suppose that we have an infinite number of rooms numbered $0, 1, 2, \ldots,$ and suppose that we instruct an individual to enter room n whenever there are n customers in the system. That is, he would be in room 2 whenever there are two customers in the system; and if another were to arrive, then he would leave room 2 and enter room 3. Similarly, if a service would take place he would leave room 2 and enter room 1 (as there would now be only one customer in the system).

Now suppose that in the long run our individual is seen to have entered room 1 at the rate of ten times an hour. Then at what rate must he have left room 1? Clearly, at this same rate of ten times an hour. For the total number of times that he enters room 1 must be equal to (or one greater than) the total number of times he leaves room 1. This sort of argument thus yields the general principle that will enable us to determine the state probabilities. Namely, for each $n \geqslant 0$, *the rate at which the process enters state n equals the rate at which it leaves state n.* Let us now determine these rates. Consider first state 0. When in state 0 the process can leave only by an arrival as clearly there cannot be a departure when the system is empty. Since the arrival rate is λ and the proportion of time that the process is in state 0 is P_0, it follows that the rate at which the process leaves state 0 is λP_0. On the other hand, state 0 can only be reached from state 1 via a departure. That is, if there is a single customer in the system and he completes service, then the system becomes empty. Since the service rate is μ and the proportion of time that the system has exactly one customer is P_1, it follows that the rate at which the process enters state 0 is μP_1.

Hence, from our rate-equality principle we get our first equation,

$$\lambda P_0 = \mu P_1$$

Now consider state 1. The process can leave this state either by an arrival (which occurs at rate λ) or a departure (which occurs at rate μ). Hence, when in state 1, the process will leave this state at a rate of $\lambda + \mu$.* Since the proportion of time the process is in state 1 is P_1, the rate at which the process leaves state 1 is $(\lambda + \mu)P_1$. On the other hand, state 1 can be entered either from state 0 via an arrival or from state 2 via a departure. Hence, the rate at which the process enters state 1 is $\lambda P_0 + \mu P_2$. Because the reasoning for other states is similar, we

* If one event occurs at a rate λ and another occurs at rate μ, then the total rate at which either event occurs is $\lambda + \mu$. Suppose one man earns \$2 per hour and another earns \$3 per hour; then together they clearly earn \$5 per hour.

obtain the following set of equations:

State	*Rate at which the process leaves = rate at which it enters*
0	$\lambda P_0 = \mu P_1$
$n, n \geqslant 1$	$(\lambda+\mu)P_n = \lambda P_{n-1} + \mu P_{n+1}$ (8.5)

Equations (8.5), which balance the rate at which the process enters each state with the rate at which it leaves that state are known as *balance equations*.

In order to solve Equations (8.5), we rewrite them to obtain

$$\begin{aligned} P_1 &= \frac{\lambda}{\mu}P_0, \\ P_{n+1} &= \frac{\lambda}{\mu}P_n + \left(P_n - \frac{\lambda}{\mu}P_{n-1}\right), \qquad n \geqslant 1 \end{aligned}$$

Solving in terms of P_0 yields

$$\begin{aligned} P_0 &= P_0, \\ P_1 &= \frac{\lambda}{\mu}P_0, \\ P_2 &= \frac{\lambda}{\mu}P_1 + \left(P_1 - \frac{\lambda}{\mu}P_0\right) = \frac{\lambda}{\mu}P_1 = \left(\frac{\lambda}{\mu}\right)^2 P_0, \\ P_3 &= \frac{\lambda}{\mu}P_2 + \left(P_2 - \frac{\lambda}{\mu}P_1\right) = \frac{\lambda}{\mu}P_2 = \left(\frac{\lambda}{\mu}\right)^3 P_0, \\ P_4 &= \frac{\lambda}{\mu}P_3 + \left(P_3 - \frac{\lambda}{\mu}P_2\right) = \frac{\lambda}{\mu}P_3 = \left(\frac{\lambda}{\mu}\right)^4 P_0, \\ P_{n+1} &= \frac{\lambda}{\mu}P_n + \left(P_n - \frac{\lambda}{\mu}P_{n-1}\right) = \frac{\lambda}{\mu}P_n = \left(\frac{\lambda}{\mu}\right)^{n+1} P_0 \end{aligned}$$

To determine P_0 we use the fact that the P_n must sum to 1, and thus

$$1 = \sum_{n=0}^{\infty} P_n = \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n P_0 = \frac{P_0}{1-\lambda/\mu}$$

or

$$\begin{aligned} P_0 &= 1 - \frac{\lambda}{\mu}, \\ P_n &= \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right), \qquad n \geqslant 1 \end{aligned} \tag{8.6}$$

Notice that for the preceding equations to make sense, it is necessary for λ/μ to be less than 1. For otherwise $\sum_{n=0}^{\infty}(\lambda/\mu)^n$ would be infinite and all the P_n would be 0. Hence, we shall assume that $\lambda/\mu < 1$. Note that it is quite intuitive that there would be no limiting probabilities if $\lambda > \mu$. For suppose that $\lambda > \mu$. Since customers arrive at a Poisson rate λ, it follows that the expected total number of arrivals by time t is λt. On the other hand, what is the expected number of customers served by time t? If there were always customers present, then the number of customers served would be a Poisson process having rate μ since the time between successive services would be independent exponentials having mean $1/\mu$. Hence, the expected number of customers served by time t is no greater than μt; and, therefore, the expected number in the system at time t is at least

$$\lambda t - \mu t = (\lambda - \mu)t$$

Now, if $\lambda > \mu$, then the preceding number goes to infinity as t becomes large. That is, $\lambda/\mu > 1$, the queue size increases without limit and there will be no limiting probabilities. Note also that the condition $\lambda/\mu < 1$ is equivalent to the condition that the mean service time be less than the mean time between successive arrivals. This is the general condition that must be satisfied for limited probabilities to exist in most single-server queueing systems.

Remarks

(i) In solving the balance equations for the $M/M/1$ queue, we obtained as an intermediate step the set of equations

$$\lambda P_n = \mu P_{n+1}, \qquad n \geqslant 0$$

These equations could have been directly argued from the general queueing result (shown in Proposition 8.1) that the rate at which arrivals find n in the system—namely λP_n—is equal to the rate at which departures leave behind n—namely, μP_{n+1}.

(ii) We can also prove that $P_n = (\lambda/\mu)^n(1 - \lambda/\mu)$ by using a queueing cost identity. Suppose that, for a fixed $n > 0$, whenever there are at least n customers in the system the nth oldest customer (with age measured from when the customer arrived) pays 1 per unit time. Letting X be the steady state number of customers in the system, because the system earns 1 per unit time whenever X is at least n, it follows that

$$\text{average rate at which the system earns} = P\{X \geqslant n\}$$

Also, because a customer who finds fewer than $n - 1$ in the system when it arrives will pay 0, while an arrival who finds at least $n - 1$ in the system will pay 1 per unit time for an exponentially distributed time with rate μ,

$$\text{average amount a customer pays} = \frac{1}{\mu}P\{X \geqslant n - 1\}$$

Therefore, the queueing cost identity yields

$$P\{X \geqslant n\} = (\lambda/\mu)P\{X \geqslant n-1\}, \qquad n > 0$$

Iterating this gives

$$\begin{aligned} P\{X \geqslant n\} &= (\lambda/\mu)P\{X \geqslant n-1\} \\ &= (\lambda/\mu)^2 P\{X \geqslant n-2\} \\ &= \cdots \\ &= (\lambda/\mu)^n P\{X \geqslant 0\} \\ &= (\lambda/\mu)^n \end{aligned}$$

Therefore,

$$P\{X = n\} = P\{X \geqslant n\} - P\{X \geqslant n+1\} = (\lambda/\mu)^n(1-\lambda/\mu) \qquad \blacksquare$$

Now let us attempt to express the quantities L, L_Q, W, and W_Q in terms of the limiting probabilities P_n. Since P_n is the long-run probability that the system contains exactly n customers, the average number of customers in the system clearly is given by

$$\begin{aligned} L &= \sum_{n=0}^{\infty} nP_n \\ &= \sum_{n=0}^{\infty} n\left(\frac{\lambda}{\mu}\right)^n \left(1-\frac{\lambda}{\mu}\right) \\ &= \frac{\lambda}{\mu-\lambda} \end{aligned} \tag{8.7}$$

where the last equation followed upon application of the algebraic identity

$$\sum_{n=0}^{\infty} nx^n = \frac{x}{(1-x)^2}$$

The quantities W, W_Q, and L_Q now can be obtained with the help of Equations (8.2) and (8.3). That is, since $\lambda_a = \lambda$, we have from Equation (8.7) that

$$\begin{aligned} W &= \frac{L}{\lambda} \\ &= \frac{1}{\mu-\lambda}, \end{aligned}$$

$$\begin{aligned} W_Q &= W - E[S] \\ &= W - \frac{1}{\mu} \\ &= \frac{\lambda}{\mu(\mu - \lambda)}, \\ L_Q &= \lambda W_Q \\ &= \frac{\lambda^2}{\mu(\mu - \lambda)} \end{aligned} \tag{8.8}$$

Example 8.2 Suppose that customers arrive at a Poisson rate of one per every 12 minutes, and that the service time is exponential at a rate of one service per 8 minutes. What are L and W?

Solution: Since $\lambda = \frac{1}{12}$, $\mu = \frac{1}{8}$, we have

$$L = 2, \qquad W = 24$$

Hence, the average number of customers in the system is 2, and the average time a customer spends in the system is 24 minutes.

Now suppose that the arrival rate increases 20 percent to $\lambda = \frac{1}{10}$. What is the corresponding change in L and W? Again using Equations (8.8), we get

$$L = 4, \qquad W = 40$$

Hence, an increase of 20 percent in the arrival rate *doubled* the average number of customers in the system.

To understand this better, write Equations (8.8) as

$$\begin{aligned} L &= \frac{\lambda/\mu}{1 - \lambda/\mu}, \\ W &= \frac{1/\mu}{1 - \lambda/\mu} \end{aligned}$$

From these equations we can see that when λ/μ is near 1, a slight increase in λ/μ will lead to a large increase in L and W. ■

A Technical Remark We have used the fact that if one event occurs at an exponential rate λ, and another independent event at an exponential rate μ, then together they occur at an exponential rate $\lambda + \mu$. To check this formally, let T_1 be the time at which the first event occurs, and T_2 the time at which the second event occurs. Then

$$\begin{aligned} P\{T_1 \leqslant t\} &= 1 - e^{-\lambda t}, \\ P\{T_2 \leqslant t\} &= 1 - e^{-\mu t} \end{aligned}$$

Now if we are interested in the time until either T_1 or T_2 occurs, then we are interested in $T = \min(T_1, T_2)$. Now,

$$\begin{aligned}P\{T \leqslant t\} &= 1 - P\{T > t\}\\ &= 1 - P\{\min(T_1, T_2) > t\}\end{aligned}$$

However, $\min(T_1, T_2) > t$ if and only if both T_1 and T_2 are greater than t; hence,

$$\begin{aligned}P\{T \leqslant t\} &= 1 - P\{T_1 > t, T_2 > t\}\\ &= 1 - P\{T_1 > t\}P\{T_2 > t\}\\ &= 1 - e^{-\lambda t}e^{-\mu t}\\ &= 1 - e^{-(\lambda+\mu)t}\end{aligned}$$

Thus, T has an exponential distribution with rate $\lambda + \mu$, and we are justified in adding the rates. ■

Given that an $M/M/1$ steady-state customer—that is, a customer who arrives after the system has been in operation a long time—spends a total of t time units in the system, let us determine the conditional distribution of N, the number of others that were present when that customer arrived. That is, letting W^* be the amount of time a customer spends in the system, we will find $P\{N = n|W^* = t\}$. Now,

$$\begin{aligned}P\{N = n|W^* = t\} &= \frac{f_{N,W^*}(n,t)}{f^*_W(t)}\\ &= \frac{P\{N = n\}f_{W^*|N}(t|n)}{f^*_W(t)}\end{aligned}$$

where $f_{W^*|N}(t|n)$ is the conditional density of W^* given that $N = n$, and $f_{W^*}(t)$ is the unconditional density of W^*. Now, given that $N = n$, the time that the customer spends in the system is distributed as the sum of $n + 1$ independent exponential random variables with a common rate μ, implying that the conditional distribution of W^* given that $N = n$ is the gamma distribution with parameters $n + 1$ and μ. Therefore, with $C = 1/f_{W^*}(t)$,

$$\begin{aligned}P\{N = n|W^* = t\} &= CP\{N = n\}\mu e^{-\mu t}\frac{(\mu t)^n}{n!}\\ &= C(\lambda/\mu)^n(1 - \lambda/\mu)\mu e^{-\mu t}\frac{(\mu t)^n}{n!} \qquad \text{(by PASTA)}\\ &= K\frac{(\lambda t)^n}{n!}\end{aligned}$$

where $K = C(1 - \lambda/\mu)\mu e^{-\mu t}$ does not depend on n. Summing over n yields

$$1 = \sum_{n=0}^{\infty} P\{N = n|T = t\} = K \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} = Ke^{\lambda t}$$

Thus, $K = e^{-\lambda t}$, showing that

$$P\{N = n| W^* = t\} = e^{-\lambda t} \frac{(\lambda t)^n}{n!}$$

Therefore, the conditional distribution of the number seen by an arrival who spends a total of t time units in the system is the Poisson distribution with mean λt.

In addition, as a by-product of our analysis, we have

$$\begin{aligned} f_{W^*}(t) &= 1/C \\ &= \frac{1}{K}(1 - \lambda/\mu)\mu e^{-\mu t} \\ &= (\mu - \lambda)e^{-(\mu-\lambda)t} \end{aligned}$$

In other words, W^*, the amount of time a customer spends in the system, is an exponential random variable with rate $\mu - \lambda$. (As a check, we note that $E[W^*] = 1/(\mu - \lambda)$, which checks with Equation (8.8) since $W = E[W^*]$.)

Remark Another argument as to why W^* is exponential with rate $\mu - \lambda$ is as follows. If we let N denote the number of customers in the system as seen by an arrival, then this arrival will spend $N + 1$ service times in the system before departing. Now,

$$P\{N + 1 = j\} = P\{N = j - 1\} = (\lambda/\mu)^{j-1}(1 - \lambda/\mu), \qquad j \geqslant 1$$

In words, the number of services that have to be completed before the arrival departs is a geometric random variable with parameter $1 - \lambda/\mu$. Therefore, after each service completion our customer will be the one departing with probability $1 - \lambda/\mu$. Thus, no matter how long the customer has already spent in the system, the probability he will depart in the next h time units is $\mu h + o(h)$, the probability that a service ends in that time, multiplied by $1 - \lambda/\mu$. That is, the customer will depart in the next h time units with probability $(\mu - \lambda)h + o(h)$, which says that the hazard rate function of W^* is the constant $\mu - \lambda$. But only the exponential has a constant hazard rate, and so we can conclude that W^* is exponential with rate $\mu - \lambda$.

Our next example illustrates the inspection paradox.

Example 8.3 For an $M/M/1$ queue in steady state, what is the probability that the next arrival finds n in the system?

Solution: Although it might initially seem, by the PASTA principle, that this probability should just be $(\lambda/\mu)^n(1-\lambda/\mu)$, we must be careful. Because if t is the current time, then the time from t until the next arrival is exponentially distributed with rate λ, and is independent of the time from t since the last arrival, which (in the limit, as t goes to infinity) is also exponential with rate λ. Thus, although the times between successive arrivals of a Poisson process are exponential with rate λ, the time between the previous arrival before t and the first arrival after t is distributed as the sum of two independent exponentials. (This is an illustration of the inspection paradox, which results because the length of an interarrival interval that contains a specified time tends to be longer than an ordinary interarrival interval—see Section 7.7.)

Let N_a denote the number found by the next arrival, and let X be the number currently in the system. Conditioning on X yields

$$\begin{aligned}
P\{N_a = n\} &= \sum_{k=0}^{\infty} P\{N_a = n | X = k\} P\{X = k\} \\
&= \sum_{k=0}^{\infty} P\{N_a = n | X = k\} (\lambda/\mu)^k (1 - \lambda/\mu) \\
&= \sum_{k=n}^{\infty} P\{N_a = n | X = k\} (\lambda/\mu)^k (1 - \lambda/\mu) \\
&= \sum_{i=0}^{\infty} P\{N_a = n | X = n + i\} (\lambda/\mu)^{n+i} (1 - \lambda/\mu)
\end{aligned}$$

Now, for $n > 0$, given there are currently $n + i$ in the system, the next arrival will find n if we have i services before an arrival and then an arrival before the next service completion. By the lack of memory property of exponential interarrival random variables, this gives

$$P\{N_a = n | X = n + i\} = \left(\frac{\mu}{\lambda + \mu}\right)^i \frac{\lambda}{\lambda + \mu}, \qquad n > 0$$

Consequently, for $n > 0$,

$$\begin{aligned}
P\{N_a = n\} &= \sum_{i=0}^{\infty} \left(\frac{\mu}{\lambda + \mu}\right)^i \frac{\lambda}{\lambda + \mu} \left(\frac{\lambda}{\mu}\right)^{n+i} (1 - \lambda/\mu) \\
&= (\lambda/\mu)^n (1 - \lambda/\mu) \frac{\lambda}{\lambda + \mu} \sum_{i=0}^{\infty} \left(\frac{\lambda}{\lambda + \mu}\right)^i \\
&= (\lambda/\mu)^{n+1} (1 - \lambda/\mu)
\end{aligned}$$

On the other hand, the probability that the next arrival will find the system empty, when there are currently i in the system, is the probability that there are i services before the next arrival. Therefore, $P\{N_a = 0|X = i\} = (\frac{\mu}{\lambda+\mu})^i$, giving

$$\begin{aligned}
P\{N_a = 0\} &= \sum_{i=0}^{\infty} \left(\frac{\mu}{\lambda + \mu}\right)^i \left(\frac{\lambda}{\mu}\right)^i (1 - \lambda/\mu) \\
&= (1 - \lambda/\mu) \sum_{i=0}^{\infty} \left(\frac{\lambda}{\lambda + \mu}\right)^i \\
&= (1 + \lambda/\mu)(1 - \lambda/\mu)
\end{aligned}$$

As a check, note that

$$\begin{aligned}
\sum_{n=0}^{\infty} P\{N_a = n\} &= (1 - \lambda/\mu)\left[1 + \lambda/\mu + \sum_{n=1}^{\infty} (\lambda/\mu)^{n+1}\right] \\
&= (1 - \lambda/\mu) \sum_{i=0}^{\infty} (\lambda/\mu)^i \\
&= 1
\end{aligned}$$

Note that $P\{N_a = 0\}$ is larger than $P_0 = 1 - \lambda/\mu$, showing that the next arrival is more likely to find an empty system than is an average arrival, and thus illustrating the inspection paradox that when the next customer arrives the elapsed time since the previous arrival is distributed as the sum of two independent exponentials with rate λ. Also, we might expect because of the inspection paradox that $E[N_a]$ is less than L, the average number of customers seen by an arrival. That this is indeed the case is seen from

$$E[N_a] = \sum_{n=1}^{\infty} n(\lambda/\mu)^{n+1}(1 - \lambda/\mu) = \frac{\lambda}{\mu} L < L \qquad \blacksquare$$

8.3.2 A Single-Server Exponential Queueing System Having Finite Capacity

In the previous model, we assumed that there was no limit on the number of customers that could be in the system at the same time. However, in reality there is always a finite system capacity N, in the sense that there can be no more than N customers in the system at any time. By this, we mean that if an arriving customer finds that there are already N customers present, then he does not enter the system.

As before, we let $P_n,\ 0 \leqslant n \leqslant N$, denote the limiting probability that there are n customers in the system. The rate-equality principle yields the following set of balance equations:

State	*Rate at which the process leaves = rate at which it enters*
0	$\lambda P_0 = \mu P_1$
$1 \leqslant n \leqslant N - 1$	$(\lambda + \mu) P_n = \lambda P_{n-1} + \mu P_{n+1}$
N	$\mu P_N = \lambda P_{N-1}$

The argument for state 0 is exactly as before. Namely, when in state 0, the process will leave only via an arrival (which occurs at rate λ) and hence the rate at which the process leaves state 0 is λP_0. On the other hand, the process can enter state 0 only from state 1 via a departure; hence, the rate at which the process enters state 0 is μP_1. The equation for state n, where $1 \leqslant n < N$, is the same as before. The equation for state N is different because now state N can only be left via a departure since an arriving customer will not enter the system when it is in state N; also, state N can now only be entered from state $N - 1$ (as there is no longer a state $N + 1$) via an arrival.

We could now either solve the balance equations exactly as we did for the infinite capacity model, or we could save a few lines by directly using the result that the rate at which departures leave behind $n - 1$ is equal to the rate at which arrivals find $n - 1$. Invoking this result yields

$$\mu P_n = \lambda P_{n-1}, \qquad n = 1, \ldots, N$$

giving

$$P_n = \frac{\lambda}{\mu} P_{n-1} = \left(\frac{\lambda}{\mu}\right)^2 P_{n-2} = \cdots = \left(\frac{\lambda}{\mu}\right)^n P_0, \qquad n = 1, \ldots, N$$

By using the fact that $\sum_{n=0}^{N} P_n = 1$ we obtain

$$\begin{aligned} 1 &= P_0 \sum_{n=0}^{N} \left(\frac{\lambda}{\mu}\right)^n \\ &= P_0 \left[\frac{1 - (\lambda/\mu)^{N+1}}{1 - \lambda/\mu}\right] \end{aligned}$$

or

$$P_0 = \frac{(1 - \lambda/\mu)}{1 - (\lambda/\mu)^{N+1}}$$

and hence from the preceding we obtain

$$P_n = \frac{(\lambda/\mu)^n(1-\lambda/\mu)}{1-(\lambda/\mu)^{N+1}}, \qquad n = 0, 1, \ldots, N$$

Note that in this case, there is no need to impose the condition that $\lambda/\mu < 1$. The queue size is, by definition, bounded so there is no possibility of its increasing indefinitely.

As before, L may be expressed in terms of P_n to yield

$$\begin{aligned} L &= \sum_{n=0}^{N} nP_n \\ &= \frac{(1-\lambda/\mu)}{1-(\lambda/\mu)^{N+1}} \sum_{n=0}^{N} n\left(\frac{\lambda}{\mu}\right)^n \end{aligned}$$

which after some algebra yields

$$L = \frac{\lambda[1 + N(\lambda/\mu)^{N+1} - (N+1)(\lambda/\mu)^N]}{(\mu-\lambda)(1-(\lambda/\mu)^{N+1})}$$

In deriving W, the expected amount of time a customer spends in the system, we must be a little careful about what we mean by a customer. Specifically, are we including those "customers" who arrive to find the system full and thus do not spend any time in the system? Or, do we just want the expected time spent in the system by a customer who actually entered the system? The two questions lead, of course, to different answers. In the first case, we have $\lambda_a = \lambda$; whereas in the second case, since the fraction of arrivals that actually enter the system is $1 - P_N$, it follows that $\lambda_a = \lambda(1 - P_N)$. Once it is clear what we mean by a customer, W can be obtained from

$$W = \frac{L}{\lambda_a}$$

Example 8.4 Suppose that it costs $c\mu$ dollars per hour to provide service at a rate μ. Suppose also that we incur a gross profit of A dollars for each customer served. If the system has a capacity N, what service rate μ maximizes our total profit?

Solution: To solve this, suppose that we use rate μ. Let us determine the amount of money coming in per hour and subtract from this the amount going out each hour. This will give us our profit per hour, and we can choose μ so as to maximize this.

Now, potential customers arrive at a rate λ. However, a certain proportion of them do not join the system—namely, those who arrive when there are N customers already in the system. Hence, since P_N is the proportion of time that the system is full, it follows that entering customers arrive at a rate of $\lambda(1 - P_N)$. Since each customer pays \$$A$, it follows that money comes in at an hourly rate of $\lambda(1 - P_N)A$ and since it goes out at an hourly rate of $c\mu$, it follows that our total profit per hour is given by

$$\begin{aligned}
\text{profit per hour} &= \lambda(1 - P_N)A - c\mu \\
&= \lambda A\left[1 - \frac{(\lambda/\mu)^N(1 - \lambda/\mu)}{1 - (\lambda/\mu)^{N+1}}\right] - c\mu \\
&= \frac{\lambda A[1 - (\lambda/\mu)^N]}{1 - (\lambda/\mu)^{N+1}} - c\mu
\end{aligned}$$

For instance if $N = 2, \lambda = 1, A = 10, c = 1$, then

$$\begin{aligned}
\text{profit per hour} &= \frac{10[1 - (1/\mu)^2]}{1 - (1/\mu)^3} - \mu \\
&= \frac{10(\mu^3 - \mu)}{\mu^3 - 1} - \mu
\end{aligned}$$

In order to maximize profit we differentiate to obtain

$$\frac{d}{d\mu}[\text{profit per hour}] = 10\frac{(2\mu^3 - 3\mu^2 + 1)}{(\mu^3 - 1)^2} - 1$$

The value of μ that maximizes our profit now can be obtained by equating to zero and solving numerically. ■

We say that a queueing system alternates between idle periods when there are no customers in the system and busy periods in which there is at least one customer in the system. We will end this section by determining the expected value and variance of the number of lost customers in a busy period, where a customer is said to be lost if it arrives when the system is at capacity.

To determine the preceding quantities, let L_n denote the number of lost customers in a busy period of a finite capacity $M/M/1$ queue in which an arrival finding n others does not join the system. To derive an expression for $E[L_n]$ and $\text{Var}(L_n)$, suppose a busy period has just begun and condition on whether the next event is an arrival or a departure. Now, with

$$I = \begin{cases} 0, & \text{if service completion occurs before next arrival} \\ 1, & \text{if arrival before service completion} \end{cases}$$

note that if $I = 0$ then the busy period will end before the next arrival and so there will be no lost customers in that busy period. As a result

$$E[L_n|I = 0] = \text{Var}(L_n|I = 0) = 0$$

Now suppose that the next arrival appears before the end of the first service time, and so $I = 1$. Then if $n = 1$ that arrival will be lost and it will be as if the busy period were just beginning anew at that point, yielding that the conditional number of lost customers has the same distribution as does $1 + L_1$. On the other hand, if $n > 1$ then at the moment of the arrival there will be two customers in the system, the one in service and the "second customer" who has just arrived. Because the distribution of the number of lost customers in a busy period does not depend on the order in which customers are served, let us suppose that the "second customer" is put aside and does not receive any service until it is the only remaining customer. Then it is easy to see that the number of lost customers until that "second customer" begins service has the same distribution as the number of lost customers in a busy period when the system capacity is $n - 1$. Moreover, the additional number of lost customers in the busy period starting when service begins on the "second customer" has the distribution of the number of lost customers in a busy period when the system capacity is n. Consequently, given $I = 1$, L_n has the distribution of the sum of two independent random variables: one of which is distributed as L_{n-1} and represents the number of lost customers before there is again only a single customer in the system, and the other which is distributed as L_n and represents the additional number of lost customers from the moment when there is again a single customer until the busy period ends. Hence,

$$E[L_n|I = 1] = \begin{cases} 1 + E[L_1], & \text{if } n = 1 \\ E[L_{n-1}] + E[L_n], & \text{if } n > 1 \end{cases}$$

and

$$\text{Var}(L_n|I = 1) = \begin{cases} \text{Var}(L_1), & \text{if } n = 1 \\ \text{Var}(L_{n-1}) + \text{Var}(L_n), & \text{if } n > 1 \end{cases}$$

Letting

$$m_n = E[L_n] \qquad \text{and} \qquad v_n = \text{Var}(L_n)$$

then, with $m_0 = 1, v_0 = 0$, the preceding equations can be rewritten as

$$E[L_n|I] = I(m_{n-1} + m_n), \tag{8.9}$$

$$\text{Var}(L_n|I) = I(v_{n-1} + v_n) \tag{8.10}$$

Using that $P(I = 1) = P(\text{arrival before service}) = \frac{\lambda}{\lambda+\mu} = 1 - P(I = 0)$, we obtain upon taking expectations of both sides of Equation (8.9) that

$$m_n = \frac{\lambda}{\lambda + \mu}[m_n + m_{n-1}]$$

or

$$m_n = \frac{\lambda}{\mu} m_{n-1}$$

Starting with $m_1 = \lambda/\mu$, this yields the result

$$m_n = (\lambda/\mu)^n$$

To determine v_n, we use the conditional variance formula. Using Equations (8.9) and (8.10) it gives

$$\begin{aligned}
v_n &= (v_n + v_{n-1})E[I] + (m_n + m_{n-1})^2 \text{Var}(I) \\
&= \frac{\lambda}{\lambda + \mu}(v_n + v_{n-1}) + [(\lambda/\mu)^n + (\lambda/\mu)^{n-1}]^2 \frac{\lambda}{\lambda + \mu}\frac{\mu}{\lambda + \mu} \\
&= \frac{\lambda}{\lambda + \mu}(v_n + v_{n-1}) + (\lambda/\mu)^{2n-2}\left(\frac{\lambda}{\mu} + 1\right)^2 \frac{\lambda\mu}{(\lambda + \mu)^2} \\
&= \frac{\lambda}{\lambda + \mu}(v_n + v_{n-1}) + (\lambda/\mu)^{2n-1}
\end{aligned}$$

Hence,

$$\mu v_n = \lambda v_{n-1} + (\lambda + \mu)(\lambda/\mu)^{2n-1}$$

or, with $\rho = \lambda/\mu$

$$v_n = \rho v_{n-1} + \rho^{2n-1} + \rho^{2n}$$

Therefore,

$$\begin{aligned}
v_1 &= \rho + \rho^2, \\
v_2 &= \rho^2 + 2\rho^3 + \rho^4, \\
v_3 &= \rho^3 + 2\rho^4 + 2\rho^5 + \rho^6, \\
v_4 &= \rho^4 + 2\rho^5 + 2\rho^6 + 2\rho^7 + \rho^8
\end{aligned}$$

and, in general,

$$v_n = \rho^n + 2 \sum_{j=n+1}^{2n-1} \rho^j + \rho^{2n}$$

8.3.3 Birth and Death Queueing Models

An exponential queueing system in which the arrival rates and the departure rates depend on the number of customers in the system is known as a *birth and death* queueing model. Let λ_n denote the arrival rate and let μ_n denote the departure rate when there are n customers in the system. Loosely speaking, when there are n customers in the system then the time until the next arrival is exponential with rate λ_n and is independent of the time of the next departure, which is exponential with rate μ_n. Equivalently, and more formally, whenever there are n customers in the system, the time until either the next arrival or the next departure occurs is an exponential random variable with rate $\lambda_n + \mu_n$ and, independent of how long it takes for this occurrence, it will be an arrival with probability $\frac{\lambda_n}{\lambda_n+\mu_n}$. We now give some examples of birth and death queues.

(a) **The $M/M/1$ Queueing System**
Because the arrival rate is always λ, and the departure rate is μ when the system is nonempty, the $M/M/1$ is a birth and death model with

$$\lambda_n = \lambda, \quad n \geqslant 0$$
$$\mu_n = \mu, \quad n \geqslant 1$$

(b) **The $M/M/1$ Queueing System with Balking**
Consider the $M/M/1$ system but now suppose that a customer that finds n others in the system upon its arrival will only join the system with probability α_n. (That is, with probability $1 - \alpha_n$ it balks at joining the system.) Then this system is a birth and death model with

$$\lambda_n = \lambda\alpha_n, \quad n \geqslant 0$$
$$\mu_n = \mu, \quad n \geqslant 1$$

The $M/M/1$ with finite capacity N is the special case where

$$\alpha_n = \begin{cases} 1, & \text{if } n < N \\ 0, & \text{if } n \geqslant N \end{cases}$$

(c) **The $M/M/k$ Queueing System**
Consider a k server system in which customers arrive according to a Poisson process with rate λ. An arriving customer immediately enters service if any of the k servers are free. If all k servers are busy, then the arrival joins the queue. When a server completes a service the customer served departs the system and if there are any customers in

queue then the one who has been waiting longest enters service with that server. All service times are exponential random variables with rate μ. Because customers are always arriving at rate λ,

$$\lambda_n = \lambda, \quad n \geqslant 0$$

Now, when there are $n \leqslant k$ customers in the system then each customer will be receiving service and so the time until a departure will be the minimum of n independent exponentials each having rate μ, and so will be exponential with rate $n\mu$. On the other hand if there are $n > k$ in the system then only k of the n will be in service, and so the departure rate in this case is $k\mu$. Hence, the $M/M/k$ is a birth and death queueing model with arrival rates

$$\lambda_n = \lambda, \quad n \geqslant 0$$

and departure rates

$$\mu_n = \begin{cases} n\mu, & \text{if } n \leqslant k \\ k\mu, & \text{if } n \geqslant k \end{cases} \qquad \blacksquare$$

To analyze the general birth and death queueing model, let P_n denote the long-run proportion of time there are n in the system. Then, either as a consequence of the balance equations given by

state	*rate at which process leaves = rate at which process enters*
$n = 0$	$\lambda_0 P_0 = \mu_1 P_1$
$n \geqslant 1$	$(\lambda_n + \mu_n)P_n = \lambda_{n-1}P_{n-1} + \mu_{n+1}P_{n+1}$

or by directly using the result that the rate at which arrivals find n in the system is equal to the rate at which departures leave behind n, we obtain

$$\lambda_n P_n = \mu_{n+1} P_{n+1}, \qquad n \geqslant 0$$

or, equivalently, that

$$P_{n+1} = \frac{\lambda_n}{\mu_{n+1}} P_n, \qquad n \geqslant 0$$

Thus,

$$P_0 = P_0,$$
$$P_1 = \frac{\lambda_0}{\mu_1} P_0,$$
$$P_2 = \frac{\lambda_1}{\mu_2} P_1 = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} P_0,$$

$$P_3 = \frac{\lambda_2}{\mu_3} P_2 = \frac{\lambda_2 \lambda_1 \lambda_0}{\mu_3 \mu_2 \mu_1} P_0$$

and, in general

$$P_n = \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} P_0, \quad n \geqslant 1$$

Using that $\sum_{n=0}^{\infty} P_n = 1$ shows that

$$1 = P_0 \left[1 + \sum_{n=1}^{\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} \right]$$

Hence,

$$P_0 = \frac{1}{1 + \sum_{n=1}^{\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n}}$$

and

$$P_n = \frac{\frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n}}{1 + \sum_{n=1}^{\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n}}, \quad n \geqslant 1$$

The necessary and sufficient conditions for the long-run probabilities to exist is that the denominator in the preceding is finite. That is, we need have that

$$\sum_{n=1}^{\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} < \infty$$

Example 8.5 For the $M/M/k$ system

$$\frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = \begin{cases} \frac{(\lambda/\mu)^n}{n!}, & \text{if } n \leqslant k \\ \frac{\lambda^n}{\mu^n k! k^{n-k}}, & \text{if } n > k \end{cases}$$

Hence, using that $\frac{\lambda^n}{\mu^n k! k^{n-k}} = (\lambda/k\mu)^n k^k/k!$ we see that

$$P_0 = \frac{1}{1 + \sum_{n=1}^{k} (\lambda/\mu)^n/n! + \sum_{n=k+1}^{\infty} (\lambda/k\mu)^n k^k/k!},$$

$$P_n = P_0 (\lambda/\mu)^n/n!, \qquad \text{if} \quad n \leqslant k$$

$$P_n = P_0 (\lambda/k\mu)^n k^k/k!, \quad \text{if} \quad n > k$$

It follows from the preceding that the condition needed for the limiting probabilities to exist is $\lambda < k\mu$. Because $k\mu$ is the service rate when all servers are busy, the preceding is just the intuitive condition that for limiting probabilities to exist the service rate needs to be larger than the arrival rate when there are many customers in the system. ■

To determine W, the average time that a customer spends in the system, for the birth and death queueing system, we employ the fundamental queueing identity $L = \lambda_a W$. Because L is the average number of customers in the system,

$$L = \sum_{n=0}^{\infty} nP_n$$

Also, because the arrival rate when there are n in the system is λ_n and the proportion of time in which there are n in the system is P_n, we see that the average arrival rate of customers is

$$\lambda_a = \sum_{n=0}^{\infty} \lambda_n P_n$$

Consequently,

$$W = \frac{\sum_{n=0}^{\infty} nP_n}{\sum_{n=0}^{\infty} \lambda_n P_n}$$

Now consider a_n equal to the proportion of arrivals that find n in the system. Since arrivals are at rate λ_n whenever there are n in system it follows that the rate at which arrivals find n is $\lambda_n P_n$. Hence, in a large time T approximately $\lambda_n P_n T$ of the approximately $\lambda_a T$ arrivals will encounter n. Letting T go to infinity shows that the long-run proportion of arrivals finding n in the system is

$$a_n = \frac{\lambda_n P_n}{\lambda_a}$$

Let us now consider the average length of a busy period, where we say that the system alternates between idle periods when there are no customers in the system and busy periods in which there is at least one customer in the system. Now, an idle period begins when the system is empty and ends when the next customer arrives. Because the arrival rate when the system is empty is λ_0, it thus follows that, independent of all that previously occurred, the length of an idle period is exponential with rate λ_0. Because a busy period always begins when there is one in the system and ends when the system is empty, it is easy to see that the lengths of successive busy periods are independent and identically distributed. Let I_j and B_j denote, respectively, the lengths of the j^{th} idle and the j^{th} busy period, $j \geqslant 1$.

Now, in the first $\sum_{j=1}^{n}(I_j + B_j)$ time units the system will be empty for a time $\sum_{j=1}^{n} I_j$. Consequently, P_0, the long-run proportion of time in which the system is empty, can be expressed as

$$\begin{aligned} P_0 &= \text{long-run proportion of time empty} \\ &= \lim_{n\to\infty} \frac{I_1 + \ldots + I_n}{I_1 + \ldots + I_n + B_1 + \ldots + B_n} \\ &= \lim_{n\to\infty} \frac{(I_1 + \ldots + I_n)/n}{(I_1 + \ldots + I_n)/n + (B_1 + \ldots + B_n)/n} \\ &= \frac{E[I]}{E[I] + E[B]} \end{aligned} \tag{8.11}$$

where I and B represent, respectively, the lengths of an idle and of a busy period, and where the final equality follows from the strong law of large numbers. Hence, using that $E[I] = 1/\lambda_0$, we see that

$$P_0 = \frac{1}{1 + \lambda_0 E[B]}$$

or,

$$E[B] = \frac{1 - P_0}{\lambda_0 P_0} \tag{8.12}$$

For instance, in the $M/M/1$ queue, this yields $E[B] = \frac{\lambda/\mu}{\lambda(1-\lambda/\mu)} = \frac{1}{\mu-\lambda}$.

Another quantity of interest is T_n, the amount of time during a busy period that there are n in the system. To determine its mean, note that $E[T_n]$ is the average amount of time there are n in the system in intervals between successive busy periods. Because the average time between successive busy periods is $E[B] + E[I]$, it follows that

$$\begin{aligned} P_n &= \text{long-run proportion of time there are } n \text{ in system} \\ &= \frac{E[T_n]}{E[I] + E[B]} \\ &= \frac{E[T_n] P_0}{E[I]} \qquad \text{from (8.11)} \end{aligned}$$

Hence,

$$E[T_n] = \frac{P_n}{\lambda_0 P_0} = \frac{\lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n}$$

As a check, note that

$$B = \sum_{n=1}^{\infty} T_n$$

and thus,

$$E[B] = \sum_{n=1}^{\infty} E\,[T_n] = \frac{1}{\lambda_0 P_0} \sum_{n=1}^{\infty} P_n = \frac{1 - P_0}{\lambda_0 P_0}$$

which is in agreement with (8.12).

For the $M/M/1$ system, the preceding gives $E\,[T_n] = \lambda^{n-1}/\mu^n$.

Whereas in exponential birth and death queueing models the state of the system is just the number of customers in the system, there are other exponential models in which a more detailed state space is needed. To illustrate, we consider some examples.

8.3.4 A Shoe Shine Shop

Consider a shoe shine shop consisting of two chairs. Suppose that an entering customer first will go to chair 1. When his work is completed in chair 1, he will go either to chair 2 if that chair is empty or else wait in chair 1 until chair 2 becomes empty. Suppose that a potential customer will enter this shop as long as chair 1 is empty. (Thus, for instance, a potential customer might enter even if there is a customer in chair 2.)

If we suppose that potential customers arrive in accordance with a Poisson process at rate λ, and that the service times for the two chairs are independent and have respective exponential rates of μ_1 and μ_2, then

(a) what proportion of potential customers enters the system?
(b) what is the mean number of customers in the system?
(c) what is the average amount of time that an entering customer spends in the system?

To begin we must first decide upon an appropriate state space. It is clear that the state of the system must include more information than merely the number of customers in the system. For instance, it would not be enough to specify that there is one customer in the system as we would also have to know which chair he was in. Further, if we only know that there are two customers in the system, then we would not know if the man in chair 1 is still being served or if he is just waiting for the person in chair 2 to finish. To account for these points, the following state space, consisting of the five states $(0,0)$, $(1,0)$, $(0,1)$, $(1,1)$, and $(b,1)$, will be used. The states have the following interpretation:

State	*Interpretation*
$(0,0)$	There are no customers in the system.
$(1,0)$	There is one customer in the system, and he is in chair 1.

$(0, 1)$	There is one customer in the system, and he is in chair 2.
$(1, 1)$	There are two customers in the system, and both are presently being served.
$(b, 1)$	There are two customers in the system, but the customer in the first chair has completed his work in that chair and is waiting for the second chair to become free.

It should be noted that when the system is in state $(b, 1)$, the person in chair 1, though not being served, is nevertheless "blocking" potential arrivals from entering the system.

As a prelude to writing down the balance equations, it is usually worthwhile to make a transition diagram. This is done by first drawing a circle for each state and then drawing an arrow labeled by the rate at which the process goes from one state to another. The transition diagram for this model is shown in Figure 8.1. The explanation for the diagram is as follows: The arrow from state $(0, 0)$ to state $(1, 0)$ that is labeled λ means that when the process is in state $(0, 0)$, that is, when the system is empty, then it goes to state $(1, 0)$ at a rate λ, that is, via an arrival. The arrow from $(0, 1)$ to $(1, 1)$ is similarly explained.

When the process is in state $(1, 0)$, it will go to state $(0, 1)$ when the customer in chair 1 is finished and this occurs at a rate μ_1; hence the arrow from $(1, 0)$ to $(0, 1)$ labeled μ_1. The arrow from $(1, 1)$ to $(b, 1)$ is similarly explained.

When in state $(b, 1)$ the process will go to state $(0, 1)$ when the customer in chair 2 completes his service (which occurs at rate μ_2); hence the arrow from $(b, 1)$ to $(0, 1)$ labeled μ_2. Also, when in state $(1, 1)$ the process will go to state $(1, 0)$ when the man in chair 2 finishes; hence the arrow from $(1, 1)$ to $(1, 0)$ labeled μ_2. Finally, if the process is in state $(0, 1)$, then it will go to state $(0, 0)$ when the man in chair 2 completes his service; hence the arrow from $(0, 1)$ to $(0, 0)$ labeled μ_2.

Because there are no other possible transitions, this completes the transition diagram.

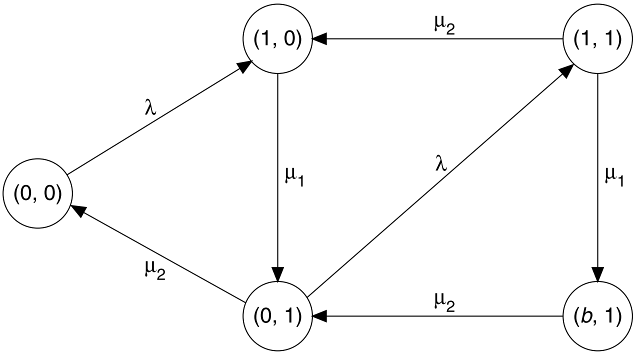

Figure 8.1 A transition diagram.

To write the balance equations we equate the sum of the arrows (multiplied by the probability of the states where they originate) coming into a state with the sum of the arrows (multiplied by the probability of the state) going out of that state. This gives

State	*Rate that the process leaves = rate that it enters*
$(0,0)$	$\lambda P_{00} = \mu_2 P_{01}$
$(1,0)$	$\mu_1 P_{10} = \lambda P_{00} + \mu_2 P_{11}$
$(0,1)$	$(\lambda + \mu_2) P_{01} = \mu_1 P_{10} + \mu_2 P_{b1}$
$(1,1)$	$(\mu_1 + \mu_2) P_{11} = \lambda P_{01}$
$(b,1)$	$\mu_2 P_{b1} = \mu_1 P_{11}$

These along with the equation

$$P_{00} + P_{10} + P_{01} + P_{11} + P_{b1} = 1$$

may be solved to determine the limiting probabilities. Though it is easy to solve the preceding equations, the resulting solutions are quite involved and hence will not be explicitly presented. However, it is easy to answer our questions in terms of these limiting probabilities. First, since a potential customer will enter the system when the state is either (0, 0) or (0, 1), it follows that the proportion of customers entering the system is $P_{00} + P_{01}$. Secondly, since there is one customer in the system whenever the state is (0 1) or (1, 0) and two customers in the system whenever the state is (1, 1) or $(b, 1)$, it follows that L, the average number in the system, is given by

$$L = P_{01} + P_{10} + 2(P_{11} + P_{b1})$$

To derive the average amount of time that an entering customer spends in the system, we use the relationship $W = L/\lambda_a$. Since a potential customer will enter the system when the state is either $(0,0)$ or $(0,1)$, it follows that $\lambda_a = \lambda(P_{00} + P_{01})$ and hence

$$W = \frac{P_{01} + P_{10} + 2(P_{11} + P_{b1})}{\lambda(P_{00} + P_{01})}$$

8.3.5 A Queueing System with Bulk Service

In this model, we consider a single-server exponential queueing system in which the server is able to serve two customers at the same time. Whenever the server completes a service, she then serves the next two customers at the same time. However, if there is only one customer in line, then she serves that customer by herself. We shall assume that her service time is exponential at rate μ whether she is serving one or two customers. As usual, we suppose that customers arrive at an exponential rate λ. One example of such a system might be an elevator or a cable car that can take at most two passengers at any time.

It would seem that the state of the system would have to tell us not only how many customers there are in the system, but also whether one or two are presently being served. However, it turns out that we can more easily solve the problem not by concentrating on the number of customers in the system, but rather on the number in *queue*. So let us define the state as the number of customers waiting in queue, with two states when there is no one in queue. That is, let us have as a state space $0', 0, 1, 2, \dots$, with the interpretation

State	*Interpretation*
$0'$	No one in service
0	Server busy; no one waiting
$n, n > 0$	n customers waiting

The transition diagram is shown in Figure 8.2 and the balance equations are

State	*Rate at which the process leaves = rate at which it enters*
$0'$	$\lambda P_{0'} = \mu P_0$
0	$(\lambda + \mu)P_0 = \lambda P_{0'} + \mu P_1 + \mu P_2$
$n, n \geqslant 1$	$(\lambda + \mu)P_n = \lambda P_{n-1} + \mu P_{n+2}$

Now the set of equations

$$(\lambda + \mu)P_n = \lambda P_{n-1} + \mu P_{n+2}, \qquad n = 1, 2, \dots \tag{8.13}$$

has a solution of the form

$$P_n = \alpha^n P_0$$

To see this, substitute the preceding in Equation (8.13) to obtain

$$(\lambda + \mu)\alpha^n P_0 = \lambda \alpha^{n-1} P_0 + \mu \alpha^{n+2} P_0$$

or

$$(\lambda + \mu)\alpha = \lambda + \mu \alpha^3$$

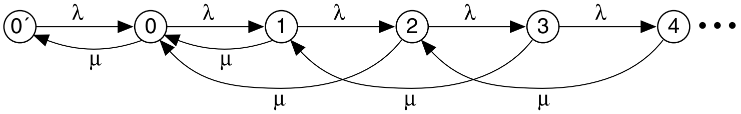

Figure 8.2 A transition diagram.

Solving this for α yields the following three roots:

$$\alpha = 1, \qquad \alpha = \frac{-1 - \sqrt{1 + 4\lambda/\mu}}{2}, \quad \text{and} \quad \alpha = \frac{-1 + \sqrt{1 + 4\lambda/\mu}}{2}$$

As the first two are clearly not possible, it follows that

$$\alpha = \frac{\sqrt{1 + 4\lambda/\mu} - 1}{2}$$

Hence,

$$P_n = \alpha^n P_0,$$
$$P_{0'} = \frac{\mu}{\lambda} P_0$$

where the bottom equation follows from the first balance equation. (We can ignore the second balance equation as one of these equations is always redundant.) To obtain P_0, we use

$$P_0 + P_{0'} + \sum_{n=1}^{\infty} P_n = 1$$

or

$$P_0 \left[1 + \frac{\mu}{\lambda} + \sum_{n=1}^{\infty} \alpha^n \right] = 1$$

or

$$P_0 \left[\frac{1}{1-\alpha} + \frac{\mu}{\lambda} \right] = 1$$

or

$$P_0 = \frac{\lambda(1-\alpha)}{\lambda + \mu(1-\alpha)}$$

and, thus

$$P_n = \frac{\alpha^n \lambda(1-\alpha)}{\lambda + \mu(1-\alpha)}, \qquad n \geqslant 0$$
$$P_{0'} = \frac{\mu(1-\alpha)}{\lambda + \mu(1-\alpha)} \tag{8.14}$$

where

$$\alpha = \frac{\sqrt{1 + 4\lambda/\mu} - 1}{2}$$

Note that for the preceding to be valid we need $\alpha < 1$, or equivalently $\lambda/\mu < 2$, which is intuitive since the maximum service rate is 2μ, which must be larger than the arrival rate λ to avoid overloading the system.

All the relevant quantities of interest now can be determined. For instance, to determine the proportion of customers that are served alone, we first note that the rate at which customers are served alone is $\lambda P_{0'} + \mu P_1$, since when the system is empty a customer will be served alone upon the next arrival and when there is one customer in queue he will be served alone upon a departure. As the rate at which customers are served is λ, it follows that

$$\begin{aligned}\text{proportion of customers that are served alone} &= \frac{\lambda P_{0'} + \mu P_1}{\lambda}\\ &= P_{0'} + \frac{\mu}{\lambda} P_1\end{aligned}$$

Also,

$$\begin{aligned}L_Q &= \sum_{n=1}^{\infty} n P_n\\ &= \frac{\lambda(1-\alpha)}{\lambda + \mu(1-\alpha)} \sum_{n=1}^{\infty} n\alpha^n \qquad \text{from Equation (8.14)}\\ &= \frac{\lambda\alpha}{(1-\alpha)[\lambda + \mu(1-\alpha)]} \qquad \text{by algebraic identity } \sum_{1}^{\infty} n\alpha^n = \frac{\alpha}{(1-\alpha)^2}\end{aligned}$$

and

$$\begin{aligned}W_Q &= \frac{L_Q}{\lambda},\\ W &= W_Q + \frac{1}{\mu},\\ L &= \lambda W\end{aligned}$$

8.4 Network of Queues

8.4.1 Open Systems

Consider a two-server system in which customers arrive at a Poisson rate λ at server 1. After being served by server 1 they then join the queue in front of server 2.

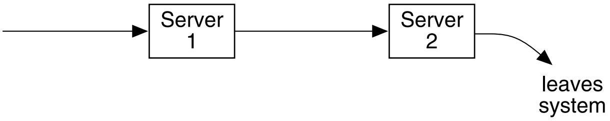

Figure 8.3 A tandem queue.

We suppose there is infinite waiting space at both servers. Each server serves one customer at a time with server i taking an exponential time with rate μ_i for a service, $i = 1, 2$. Such a system is called a *tandem* or *sequential* system (see Figure 8.3).

To analyze this system we need to keep track of the number of customers at server 1 and the number at server 2. So let us define the state by the pair (n, m)—meaning that there are n customers at server 1 and m at server 2. The balance equations are

State	*Rate that the process leaves = rate that it enters*
$0, 0$	$\lambda P_{0,0} = \mu_2 P_{0,1}$
$n, 0; n > 0$	$(\lambda + \mu_1) P_{n,0} = \mu_2 P_{n,1} + \lambda P_{n-1,0}$
$0, m; m > 0$	$(\lambda + \mu_2) P_{0,m} = \mu_2 P_{0,m+1} + \mu_1 P_{1,m-1}$
$n, m; nm > 0$	$(\lambda + \mu_1 + \mu_2) P_{n,m} = \mu_2 P_{n,m+1} + \mu_1 P_{n+1,m-1} + \lambda P_{n-1,m}$ (8.15)

Rather than directly attempting to solve these (along with the equation $\sum_{n,m} P_{n,m} = 1$) we shall guess at a solution and then verify that it indeed satisfies the preceding. We first note that the situation at server 1 is just as in an $M/M/1$ model. Similarly, as it was shown in Section 6.6 that the departure process of an $M/M/1$ queue is a Poisson process with rate λ, it follows that what server 2 faces is also an $M/M/1$ queue. Hence, the probability that there are n customers at server 1 is

$$P\{n \text{ at server } 1\} = \left(\frac{\lambda}{\mu_1}\right)^n \left(1 - \frac{\lambda}{\mu_1}\right)$$

and, similarly,

$$P\{m \text{ at server } 2\} = \left(\frac{\lambda}{\mu_2}\right)^m \left(1 - \frac{\lambda}{\mu_2}\right)$$

Now, if the numbers of customers at servers 1 and 2 were independent random variables, then it would follow that

$$P_{n,m} = \left(\frac{\lambda}{\mu_1}\right)^n \left(1 - \frac{\lambda}{\mu_1}\right) \left(\frac{\lambda}{\mu_2}\right)^m \left(1 - \frac{\lambda}{\mu_2}\right) \tag{8.16}$$

To verify that $P_{n,m}$ is indeed equal to the preceding (and thus that the number of customers at server 1 is independent of the number at server 2), all we need do is verify that the preceding satisfies Equations (8.15)—this suffices since we know that the $P_{n,m}$ are the unique solution of Equations (8.15). Now, for instance, if we consider the first equation of (8.15), we need to show that

$$\lambda\left(1-\frac{\lambda}{\mu_1}\right)\left(1-\frac{\lambda}{\mu_2}\right)=\mu_2\left(1-\frac{\lambda}{\mu_1}\right)\left(\frac{\lambda}{\mu_2}\right)\left(1-\frac{\lambda}{\mu_2}\right)$$

which is easily verified. We leave it as an exercise to show that the $P_{n,m}$, as given by Equation (8.16), satisfy all of the equations of (8.15), and are thus the limiting probabilities.

From the preceding we see that L, the average number of customers in the system, is given by

$$\begin{aligned}L&=\sum_{n,m}(n+m)P_{n,m}\\&=\sum_n n\left(\frac{\lambda}{\mu_1}\right)^n\left(1-\frac{\lambda}{\mu_1}\right)+\sum_m m\left(\frac{\lambda}{\mu_2}\right)^m\left(1-\frac{\lambda}{\mu_2}\right)\\&=\frac{\lambda}{\mu_1-\lambda}+\frac{\lambda}{\mu_2-\lambda}\end{aligned}$$

and from this we see that the average time a customer spends in the system is

$$W=\frac{L}{\lambda}=\frac{1}{\mu_1-\lambda}+\frac{1}{\mu_2-\lambda}$$

Remarks

(i) The result (Equations (8.15)) could have been obtained as a direct consequence of the time reversibility of an $M/M/1$ (see Section 6.6). For not only does time reversibility imply that the output from server 1 is a Poisson process, but it also implies (Exercise 26 of Chapter 6) that the number of customers at server 1 is independent of the past departure times from server 1. As these past departure times constitute the arrival process to server 2, the independence of the numbers of customers in the two systems follows.

(ii) Since a Poisson arrival sees time averages, it follows that in a tandem queue the numbers of customers an arrival (to server 1) sees at the two servers are independent random variables. However, it should be noted that this does not imply that the waiting times of a given customer at the two servers are independent. For a counterexample suppose that λ is very small with respect to $\mu_1=\mu_2$, and thus almost all customers have zero wait in queue at both servers. However, given that the wait in queue of a customer at server 1 is positive, his wait in queue at server 2 also will be positive with probability at least as large as $\frac{1}{2}$ (why?). Hence, the waiting times

> in queue are not independent. Remarkably enough, however, it turns out that the total times (that is, service time plus wait in queue) that an arrival spends at the two servers are indeed independent random variables.

The preceding result can be substantially generalized. To do so, consider a system of k servers. Customers arrive from outside the system to server i, $i = 1, \ldots, k$, in accordance with independent Poisson processes at rate r_i; they then join the queue at i until their turn at service comes. Once a customer is served by server i, he then joins the queue in front of server j, $j = 1, \ldots, k$, with probability P_{ij}. Hence, $\sum_{j=1}^{k} P_{ij} \leqslant 1$, and $1 - \sum_{j=1}^{k} P_{ij}$ represents the probability that a customer departs the system after being served by server i.

If we let λ_j denote the total arrival rate of customers to server j, then the λ_j can be obtained as the solution of

$$\lambda_j = r_j + \sum_{i=1}^{k} \lambda_i P_{ij}, \qquad i = 1, \ldots, k \tag{8.17}$$

Equation (8.17) follows since r_j is the arrival rate of customers to j coming from outside the system and, as λ_i is the rate at which customers depart server i (rate in must equal rate out), $\lambda_i P_{ij}$ is the arrival rate to j of those coming from server i.

It turns out that the number of customers at each of the servers is independent and of the form

$$P\{n \text{ customers at server } j\} = \left(\frac{\lambda_j}{\mu_j}\right)^n \left(1 - \frac{\lambda_j}{\mu_j}\right), \qquad n \geqslant 1$$

where μ_j is the exponential service rate at server j and the λ_j are the solution to Equation (8.17). Of course, it is necessary that $\lambda_j/\mu_j < 1$ for all j. To prove this, we first note that it is equivalent to asserting that the limiting probabilities $P(n_1, n_2, \ldots, n_k) = P\{n_j \text{ at server } j, j = 1, \ldots, k\}$ are given by

$$P(n_1, n_2, \ldots, n_k) = \prod_{j=1}^{k} \left(\frac{\lambda_j}{\mu_j}\right)^{n_j} \left(1 - \frac{\lambda_j}{\mu_j}\right) \tag{8.18}$$

which can be verified by showing that it satisfies the balance equations for this model.

The average number of customers in the system is

$$\begin{aligned} L &= \sum_{j=1}^{k} \text{average number at server } j \\ &= \sum_{j=1}^{k} \frac{\lambda_j}{\mu_j - \lambda_j} \end{aligned}$$

The average time a customer spends in the system can be obtained from $L = \lambda W$ with $\lambda = \sum_{j=1}^{k} r_j$. (Why not $\lambda = \sum_{j=1}^{k} \lambda_j$?) This yields

$$W = \frac{\sum_{j=1}^{k} \lambda_j/(\mu_j - \lambda_j)}{\sum_{j=1}^{k} r_j}$$

Remark The result embodied in Equation (8.18) is rather remarkable in that it says that the distribution of the number of customers at server i is the same as in an $M/M/1$ system with rates λ_i and μ_i. What is remarkable is that in the network model the arrival process at node i need *not* be a Poisson process. For if there is a possibility that a customer may visit a server more than once (a situation called *feedback*), then the arrival process will not be Poisson. An easy example illustrating this is to suppose that there is a single server whose service rate is very large with respect to the arrival rate from outside. Suppose also that with probability $p = 0.9$ a customer upon completion of service is fed back into the system. Hence, at an arrival time epoch there is a large probability of another arrival in a short time (namely, the feedback arrival); whereas at an arbitrary time point there will be only a very slight chance of an arrival occurring shortly (since λ is so very small). Hence, the arrival process does not possess independent increments and so cannot be Poisson.

Thus, we see that when feedback is allowed the steady-state probabilities of the number of customers at any given station have the same distribution as in an $M/M/1$ model even though the model is not $M/M/1$. (Presumably such quantities as the joint distribution of the number at the station at two different time points will not be the same as for an $M/M/1$.)

Example 8.6 Consider a system of two servers where customers from outside the system arrive at server 1 at a Poisson rate 4 and at server 2 at a Poisson rate 5. The service rates of 1 and 2 are respectively 8 and 10. A customer upon completion of service at server 1 is equally likely to go to server 2 or to leave the system (i.e., $P_{11} = 0$, $P_{12} = \frac{1}{2}$); whereas a departure from server 2 will go 25 percent of the time to server 1 and will depart the system otherwise (i.e., $P_{21} = \frac{1}{4}$, $P_{22} = 0$). Determine the limiting probabilities, L, and W.

Solution: The total arrival rates to servers 1 and 2—call them λ_1 and λ_2—can be obtained from Equation (8.17). That is, we have

$$\lambda_1 = 4 + \tfrac{1}{4}\lambda_2,$$
$$\lambda_2 = 5 + \tfrac{1}{2}\lambda_1$$

implying that

$$\lambda_1 = 6, \qquad \lambda_2 = 8$$

Hence,

$$\begin{aligned}P\{n \text{ at server } 1, m \text{ at server } 2\} &= \left(\tfrac{3}{4}\right)^n \tfrac{1}{4} \left(\tfrac{4}{5}\right)^m \tfrac{1}{5}\\ &= \tfrac{1}{20} \left(\tfrac{3}{4}\right)^n \left(\tfrac{4}{5}\right)^m\end{aligned}$$

and

$$\begin{aligned}L &= \frac{6}{8-6} + \frac{8}{10-8} = 7,\\ W &= \frac{L}{9} = \frac{7}{9}\end{aligned} \qquad \blacksquare$$

8.4.2 Closed Systems

The queueing systems described in Section 8.4.1 are called *open systems* since customers are able to enter and depart the system. A system in which new customers never enter and existing ones never depart is called a *closed system*.

Let us suppose that we have m customers moving among a system of k servers, where the service times at server i are exponential with rate μ_i, $i = 1, \ldots, k$. When a customer completes service at server i, she then joins the queue in front of server j, $j = 1, \ldots, k$, with probability P_{ij}, where we now suppose that $\sum_{j=1}^{k} P_{ij} = 1$ for all $i = 1, \ldots, k$. That is, $\mathbf{P} = [P_{ij}]$ is a Markov transition probability matrix, which we shall assume is irreducible. Let $\pi = (\pi_1, \ldots, \pi_k)$ denote the stationary probabilities for this Markov chain; that is, π is the unique positive solution of

$$\begin{aligned}\pi_j &= \sum_{i=1}^{k} \pi_i P_{ij},\\ \sum_{j=1}^{k} \pi_j &= 1\end{aligned} \tag{8.19}$$

If we denote the average arrival rate (or equivalently the average service completion rate) at server j by $\lambda_m(j)$, $j = 1, \ldots, k$ then, analogous to Equation (8.17), the $\lambda_m(j)$ satisfy

$$\lambda_m(j) = \sum_{i=1}^{k} \lambda_m(i) P_{ij}$$

Hence, from (8.19) we can conclude that

$$\lambda_m(j) = \lambda_m \pi_j, \qquad j = 1, 2, \ldots, k \tag{8.20}$$

where

$$\lambda_m = \sum_{j=1}^{k} \lambda_m(j) \tag{8.21}$$

From Equation (8.21), we see that λ_m is the average service completion rate of the entire system, that is, it is the system *throughput* rate.*

If we let $P_m(n_1, n_2, \ldots, n_k)$ denote the limiting probabilities

$$P_m(n_1, n_2, \ldots, n_k) = P\{n_j \text{ customers at server } j, j = 1, \ldots, k\}$$

then, by verifying that they satisfy the balance equation, it can be shown that

$$P_m(n_1, n_2, \ldots, n_k) = \begin{cases} K_m \prod_{j=1}^{k} (\lambda_m(j)/\mu_j)^{n_j}, & \text{if } \sum_{j=1}^{k} n_j = m \\ 0, & \text{otherwise} \end{cases}$$

But from Equation (8.20) we thus obtain

$$P_m(n_1, n_2, \ldots, n_k) = \begin{cases} C_m \prod_{j=1}^{k} (\pi_j/\mu_j)^{n_j}, & \text{if } \sum_{j=1}^{k} n_j = m \\ 0, & \text{otherwise} \end{cases} \tag{8.22}$$

where

$$C_m = \left[\sum_{\substack{n_1, \ldots, n_k: \\ \sum n_j = m}} \prod_{j=1}^{k} (\pi_j/\mu_j)^{n_j} \right]^{-1} \tag{8.23}$$

Equation (8.22) is not as useful as we might suppose, for in order to utilize it we must determine the normalizing constant C_m given by Equation (8.23), which requires summing the products $\Pi_{j=1}^{k} (\pi_j/\mu_j)^{n_j}$ over all the feasible vectors $(n_1, \ldots, n_k)$: $\sum_{j=1}^{k} n_j = m$. Hence, since there are $\binom{m+k-1}{m}$ vectors this is only computationally feasible for relatively small values of m and k.

We will now present an approach that will enable us to determine recursively many of the quantities of interest in this model without first computing the normalizing constants. To begin, consider a customer who has just left server i and is headed to server j, and let us determine the probability of the system as seen by this customer. In particular, let us determine the probability that this customer

* We are just using the notation $\lambda_m(j)$ and λ_m to indicate the dependence on the number of customers in the closed system. This will be used in recursive relations we will develop.

observes, at that moment, n_l customers at server $l, l = 1, \ldots, k, \sum_{l=1}^{k} n_l = m-1$. This is done as follows:

$$
\begin{aligned}
&P\{\text{customer observes } n_l \text{ at server } l, l = 1, \ldots, k \mid \text{ customer goes from } i \text{ to } j\}\\
&\quad = \frac{P\{\text{state is } (n_1, \ldots, n_i + 1, \ldots, n_j, \ldots, n_k), \text{customer goes from } i \text{ to } j\}}{P\{\text{customer goes from } i \text{ to } j\}}\\
&\quad = \frac{P_m(n_1, \ldots, n_i + 1, \ldots, n_j, \ldots, n_k)\mu_i P_{ij}}{\sum_{n:\sum n_j = m-1} P_m(n_1, \ldots, n_i + 1, \ldots, n_k)\mu_i P_{ij}}\\
&\quad = \frac{(\pi_i/\mu_i)\prod_{j=1}^{k}(\pi_j/\mu_j)^{n_j}}{K} \qquad \text{from (8.22)}\\
&\quad = C\prod_{j=1}^{k}(\pi_j/\mu_j)^{n_j}
\end{aligned}
$$

where C does not depend on $n_1, \ldots, n_k$. But because the preceding is a probability density on the set of vectors $(n_1, \ldots, n_k)$, $\sum_{j=1}^{k} n_j = m - 1$, it follows from (8.22) that it must equal $P_{m-1}(n_1, \ldots, n_k)$. Hence,

$$
\begin{aligned}
&P\{\text{customer observes } n_l \text{ at server } l, l = 1, \ldots, k \mid \text{customer goes from } i \text{ to } j\}\\
&\quad = P_{m-1}(n_1, \ldots, n_k), \qquad \sum_{i=1}^{k} n_i = m - 1 \qquad (8.24)
\end{aligned}
$$

As (8.24) is true for all i, we thus have proven the following proposition, known as the arrival theorem.

Proposition 8.3 (The Arrival Theorem) In the closed network system with m customers, the system as seen by arrivals to server j is distributed as the stationary distribution in the same network system when there are only $m - 1$ customers.

Denote by $L_m(j)$ and $W_m(j)$ the average number of customers and the average time a customer spends at server j when there are m customers in the network. Upon conditioning on the number of customers found at server j by an arrival to that server, it follows that

$$
\begin{aligned}
W_m(j) &= \frac{1 + E_m[\text{number at server } j \text{ as seen by an arrival}]}{\mu_j}\\
&= \frac{1 + L_{m-1}(j)}{\mu_j} \qquad (8.25)
\end{aligned}
$$

where the last equality follows from the arrival theorem. Now when there are $m - 1$ customers in the system, then, from Equation (8.20), $\lambda_{m-1}(j)$, the average arrival rate to server j, satisfies

$$\lambda_{m-1}(j) = \lambda_{m-1}\pi_j$$

Now, applying the basic cost identity Equation (8.1) with the cost rule being that each customer in the network system of $m-1$ customers pays one per unit time while at server j, we obtain

$$L_{m-1}(j) = \lambda_{m-1}\pi_j W_{m-1}(j) \tag{8.26}$$

Using Equation (8.25), this yields

$$W_m(j) = \frac{1 + \lambda_{m-1}\pi_j W_{m-1}(j)}{\mu_j} \tag{8.27}$$

Also using the fact that $\sum_{j=1}^{k} L_{m-1}(j) = m-1$ (why?) we obtain, from Equation (8.26), the following:

$$m - 1 = \lambda_{m-1}\sum_{j=1}^{k}\pi_j W_{m-1}(j)$$

or

$$\lambda_{m-1} = \frac{m-1}{\sum_{i=1}^{k}\pi_i W_{m-1}(i)} \tag{8.28}$$

Hence, from Equation (8.27), we obtain the recursion

$$W_m(j) = \frac{1}{\mu_j} + \frac{(m-1)\pi_j W_{m-1}(j)}{\mu_j \sum_{i=1}^{k}\pi_i W_{m-1}(i)} \tag{8.29}$$

Starting with the stationary probabilities $\pi_j, j = 1,\ldots,k$, and $W_1(j) = 1/\mu_j$ we can now use Equation (8.29) to determine recursively $W_2(j), W_3(j), \ldots, W_m(j)$. We can then determine the throughput rate λ_m by using Equation (8.28), and this will determine $L_m(j)$ by Equation (8.26). This recursive approach is called *mean value analysis*.

Example 8.7 Consider a k-server network in which the customers move in a cyclic permutation. That is,

$$P_{i,i+1} = 1, \qquad i = 1,2\ldots,k-1, \qquad P_{k,1} = 1$$

Let us determine the average number of customers at server j when there are two customers in the system. Now, for this network,

$$\pi_i = 1/k, \qquad i = 1,\ldots,k$$

and as

$$W_1(j) = \frac{1}{\mu_j}$$

we obtain from Equation (8.29) that

$$\begin{aligned} W_2(j) &= \frac{1}{\mu_j} + \frac{(1/k)(1/\mu_j)}{\mu_j \sum_{i=1}^k (1/k)(1/\mu_i)} \\ &= \frac{1}{\mu_j} + \frac{1}{\mu_j^2 \sum_{i=1}^k 1/\mu_i} \end{aligned}$$

Hence, from Equation (8.28),

$$\lambda_2 = \frac{2}{\sum\limits_{l=1}^{k} \frac{1}{k} W_2(l)} = \frac{2k}{\sum\limits_{l=1}^{k} \left(\frac{1}{\mu_l} + \frac{1}{\mu_l^2 \sum_{i=1}^k 1/\mu_i} \right)}$$

and finally, using Equation (8.26),

$$L_2(j) = \lambda_2 \frac{1}{k} W_2(j) = \frac{2\left(\frac{1}{\mu_j} + \frac{1}{\mu_j^2 \sum_{i=1}^k 1/\mu_i} \right)}{\sum\limits_{l=1}^{k} \left(\frac{1}{\mu_l} + \frac{1}{\mu_l^2 \sum_{i=1}^k 1/\mu_i} \right)} \qquad \blacksquare$$

Another approach to learning about the stationary probabilities specified by Equation (8.22), which finesses the computational difficulties of computing the constant C_m, is to use the Gibbs sampler of Section 4.9 to generate a Markov chain having these stationary probabilities. To begin, note that since there are always a total of m customers in the system, Equation (8.22) may equivalently be written as a joint mass function of the numbers of customers at each of the servers $1, \ldots, k-1$, as follows:

$$\begin{aligned} P_m(n_1, \ldots, n_{k-1}) &= C_m (\pi_k/\mu_k)^{m - \sum n_j} \prod_{j=1}^{k-1} (\pi_j/\mu_j)^{n_j} \\ &= K \prod_{j=1}^{k-1} (a_j)^{n_j}, \qquad \sum_{j=1}^{k-1} n_j \leqslant m \end{aligned}$$

where $a_j = (\pi_j \mu_k)/(\pi_k \mu_j), j = 1, \ldots, k-1$. Now, if $\mathbf{N} = (N_1, \ldots, N_{k-1})$ has the preceding joint mass function then

$$\begin{aligned}
&P\{N_i = n | N_1 = n_1, \ldots, N_{i-1} = n_{i-1}, N_{i+1} = n_{i+1}, \ldots, N_{k-1} = n_{k-1}\}\\
&\quad = \frac{P_m(n_1, \ldots, n_{i-1}, n, n_{i+1}, \ldots, n_{k-1})}{\sum_r P_m(n_1, \ldots, n_{i-1}, r, n_{i+1}, \ldots, n_{k-1})}\\
&\quad = Ca_i^n, \qquad n \leqslant m - \sum_{j \neq i} n_j
\end{aligned}$$

It follows from the preceding that we may use the Gibbs sampler to generate the values of a Markov chain whose limiting probability mass function is $P_m(n_1, \ldots, n_{k-1})$ as follows:

1. Let $(n_1, \ldots, n_{k-1})$ be arbitrary nonnegative integers satisfying $\sum_{j=1}^{k-1} n_j \leqslant m$.
2. Generate a random variable I that is equally likely to be any of $1, \ldots, k-1$.
3. If $I = i$, set $s = m - \sum_{j \neq i} n_j$, and generate the value of a random variable X having probability mass function

$$P\{X = n\} = Ca_i^n, \qquad n = 0, \ldots, s$$

4. Let $n_I = X$ and go to step 2.

The successive values of the state vector $(n_1, \ldots, n_{k-1}, m - \sum_{j=1}^{k-1} n_j)$ constitute the sequence of states of a Markov chain with the limiting distribution P_m. All quantities of interest can be estimated from this sequence. For instance, the average of the values of the jth coordinate of these vectors will converge to the mean number of individuals at station j, the proportion of vectors whose jth coordinate is less than r will converge to the limiting probability that the number of individuals at station j is less than r, and so on.

Other quantities of interest can also be obtained from the simulation. For instance, suppose we want to estimate W_j, the average amount of time a customer spends at server j on each visit. Then, as noted in the preceding, L_j, the average number of customers at server j, can be estimated. To estimate W_j, we use the identity

$$L_j = \lambda_j W_j$$

where λ_j is the rate at which customers arrive at server j. Setting λ_j equal to the service completion rate at server j shows that

$$\lambda_j = P\{j \text{ is busy}\} \mu_j$$

Using the Gibbs sampler simulation to estimate $P\{j \text{ is busy}\}$ then leads to an estimator of W_j.

8.5 The System $M/G/1$

8.5.1 Preliminaries: Work and Another Cost Identity

For an arbitrary queueing system, let us define the work in the system at any time t to be the sum of the remaining service times of all customers in the system at time t. For instance, suppose there are three customers in the system—the one in service having been there for three of his required five units of service time, and both people in queue having service times of six units. Then the work at that time is $2 + 6 + 6 = 14$. Let V denote the (time) average work in the system.

Now recall the fundamental cost equation (8.1), which states that the

$$\begin{aligned}&\text{average rate at which the system earns}\\ &\quad = \lambda_a \times \text{average amount a customer pays}\end{aligned}$$

and consider the following cost rule: *Each customer pays at a rate of y/unit time when his remaining service time is y, whether he is in queue or in service*. Thus, the rate at which the system earns is just the work in the system; so the basic identity yields

$$V = \lambda_a E[\text{amount paid by a customer}]$$

Now, let S and W_Q^* denote respectively the service time and the time a given customer spends waiting in queue. Then, since the customer pays at a constant rate of S per unit time while he waits in queue and at a rate of $S-x$ after spending an amount of time x in service, we have

$$E[\text{amount paid by a customer}] = E\left[SW_Q^* + \int_0^S (S-x)\,dx\right]$$

and thus

$$V = \lambda_a E[SW_Q^*] + \frac{\lambda_a E[S^2]}{2} \tag{8.30}$$

It should be noted that the preceding is a basic queueing identity (like Equations (8.2)–(8.4)) and as such is valid in almost all models. In addition, if a customer's service time is independent of his wait in queue (as is usually, but not always the case),* then we have from Equation (8.30) that

$$V = \lambda_a E[S] W_Q + \frac{\lambda_a E[S^2]}{2} \tag{8.31}$$

* For an example where it is not true, see Section 8.6.2.

8.5.2 Application of Work to M/G/1

The $M/G/1$ model assumes (i) Poisson arrivals at rate λ; (ii) a general service distribution; and (iii) a single server. In addition, we will suppose that customers are served in the order of their arrival.

Now, for an arbitrary customer in an $M/G/1$ system,

$$\text{customer's wait in queue} = \text{work in the system when he arrives} \tag{8.32}$$

This follows since there is only a single server (think about it!). Taking expectations of both sides of Equation (8.32) yields

$$W_Q = \text{average work as seen by an arrival}$$

But, due to Poisson arrivals, the average work as seen by an arrival will equal V, the time average work in the system. Hence, for the model $M/G/1$,

$$W_Q = V$$

The preceding in conjunction with the identity

$$V = \lambda E[S] W_Q + \frac{\lambda E[S^2]}{2}$$

yields the so-called *Pollaczek–Khintchine formula*,

$$W_Q = \frac{\lambda E[S^2]}{2(1 - \lambda E[S])} \tag{8.33}$$

where $E[S]$ and $E[S^2]$ are the first two moments of the service distribution.

The quantities L, L_Q, and W can be obtained from Equation (8.33) as

$$\begin{aligned} L_Q &= \lambda W_Q = \frac{\lambda^2 E[S^2]}{2(1 - \lambda E[S])}, \\ W &= W_Q + E[S] = \frac{\lambda E[S^2]}{2(1 - \lambda E[S])} + E[S], \\ L &= \lambda W = \frac{\lambda^2 E[S^2]}{2(1 - \lambda E[S])} + \lambda E[S] \end{aligned} \tag{8.34}$$

Remarks

(i) For the preceding quantities to be finite, we need $\lambda E[S] < 1$. This condition is intuitive since we know from renewal theory that if the server was always busy, then the departure rate would be $1/E[S]$ (see Section 7.3), which must be larger than the arrival rate λ to keep things finite.

(ii) Since $E[S^2] = \text{Var}(S) + (E[S])^2$, we see from Equations (8.33) and (8.34) that, for fixed mean service time, L, L_Q, W, and W_Q all increase as the variance of the service distribution increases.

(iii) Another approach to obtain W_Q is presented in Exercise 38.

8.5.3 Busy Periods

The system alternates between idle periods (when there are no customers in the system, and so the server is idle) and busy periods (when there is at least one customer in the system, and so the server is busy).

Let I and B represent, respectively, the length of an idle and of a busy period. Because I represents the time from when a customer departs and leaves the system empty until the next arrival, it follows, since arrivals are according to a Poisson process with rate λ, that I is exponential with rate λ and thus

$$E[I] = \frac{1}{\lambda} \tag{8.35}$$

To determine $E[B]$ we argue, as in Section 8.3.3, that the long-run proportion of time the system is empty is equal to the ratio of $E[I]$ to $E[I] + E[B]$. That is,

$$P_0 = \frac{E[I]}{E[I] + E[B]} \tag{8.36}$$

To compute P_0, we note from Equation (8.4) (obtained from the fundamental cost equation by supposing that a customer pays at a rate of one per unit time while in service) that

$$\text{average number of busy servers} = \lambda E[S]$$

However, as the left-hand side of the preceding equals $1 - P_0$ (why?), we have

$$P_0 = 1 - \lambda E[S] \tag{8.37}$$

and, from Equations (8.35)–(8.37),

$$1 - \lambda E[S] = \frac{1/\lambda}{1/\lambda + E[B]}$$

or

$$E[B] = \frac{E[S]}{1 - \lambda E[S]}$$

Another quantity of interest is C, the number of customers served in a busy period. The mean of C can be computed by noting that, on the average, for every $E[C]$ arrivals exactly one will find the system empty (namely, the first customer in the busy period). Hence,

$$a_0 = \frac{1}{E[C]}$$

and, as $a_0 = P_0 = 1 - \lambda E[S]$ because of Poisson arrivals, we see that

$$E[C] = \frac{1}{1 - \lambda E[S]}$$

8.6 Variations on the *M/G/1*

8.6.1 *The* M/G/1 *with Random-Sized Batch Arrivals*

Suppose that, as in the $M/G/1$, arrivals occur in accordance with a Poisson process having rate λ. But now suppose that each arrival consists not of a single customer but of a random number of customers. As before there is a single server whose service times have distribution G.

Let us denote by $\alpha_j, j \geqslant 1$, the probability that an arbitrary batch consists of j customers; and let N denote a random variable representing the size of a batch and so $P\{N = j\} = \alpha_j$. Since $\lambda_a = \lambda E(N)$, the basic formula for work (Equation (8.31)) becomes

$$V = \lambda E[N]\left[E(S)W_Q + \frac{E(S^2)}{2}\right] \tag{8.38}$$

To obtain a second equation relating V to W_Q, consider an average customer. We have that

$$\begin{aligned}\text{his wait in queue} &= \text{work in system when he arrives}\\ &\quad + \text{ his waiting time due to those in his batch}\end{aligned}$$

Taking expectations and using the fact that Poisson arrivals see time averages yields

$$\begin{aligned}W_Q &= V + E[\text{waiting time due to those in his batch}]\\ &= V + E[W_B] \end{aligned} \tag{8.39}$$

Now, $E(W_B)$ can be computed by conditioning on the number in the batch, but we must be careful because the probability that our average customer comes from a batch of size j is *not* α_j. For α_j is the proportion of batches that are of size j, and if we pick a customer at random, it is more likely that he comes from a larger rather than a smaller batch. (For instance, suppose $\alpha_1 = \alpha_{100} = \frac{1}{2}$, then half the batches are of size 1 but 100/101 of the customers will come from a batch of size 100!)

To determine the probability that our average customer came from a batch of size j we reason as follows: Let M be a large number. Then of the first M batches approximately $M\alpha_j$ will be of size j, $j \geqslant 1$, and thus there would have been approximately $jM\alpha_j$ customers that arrived in a batch of size j. Hence, the proportion of arrivals in the first M batches that were from batches of size j is approximately $jM\alpha_j / \sum_j jM\alpha_j$. This proportion becomes exact as $M \to \infty$, and so we see that

$$\begin{aligned}\text{proportion of customers from batches of size } j &= \frac{j\alpha_j}{\sum_j j\alpha_j}\\ &= \frac{j\alpha_j}{E[N]}\end{aligned}$$

We are now ready to compute $E(W_B)$, the expected wait in queue due to others in the batch:

$$E[W_B] = \sum_j E[W_B \mid \text{batch of size } j] \frac{j\alpha_j}{E[N]} \tag{8.40}$$

Now if there are j customers in his batch, then our customer would have to wait for $i - 1$ of them to be served if he was ith in line among his batch members. As he is equally likely to be either 1st, 2nd, ..., or jth in line we see that

$$\begin{aligned}E[W_B \mid \text{batch is of size } j] &= \sum_{i=1}^{j} (i-1)E(S)\frac{1}{j}\\ &= \frac{j-1}{2}E[S]\end{aligned}$$

Substituting this in Equation (8.40) yields

$$\begin{aligned}E[W_B] &= \frac{E[S]}{2E[N]} \sum_j (j-1)j\alpha_j\\ &= \frac{E[S](E[N^2] - E[N])}{2E[N]}\end{aligned}$$

and from Equations (8.38) and (8.39) we obtain

$$W_Q = \frac{E[S](E[N^2] - E[N])/2E[N] + \lambda E[N]E[S^2]/2}{1 - \lambda E[N]E[S]}$$

Remarks

(i) Note that the condition for W_Q to be finite is that

$$\lambda E(N) < \frac{1}{E[S]}$$

which again says that the arrival rate must be less than the service rate (when the server is busy).

(ii) For fixed value of $E[N]$, W_Q is increasing in $\text{Var}(N)$, again indicating that "single-server queues do not like variation."

(iii) The other quantities L, L_Q, and W can be obtained by using

$$\begin{aligned} W &= W_Q + E[S], \\ L &= \lambda_a W = \lambda E[N]W, \\ L_Q &= \lambda E[N]W_Q \end{aligned}$$

8.6.2 Priority Queues

Priority queueing systems are ones in which customers are classified into types and then given service priority according to their type. Consider the situation where there are two types of customers, which arrive according to independent Poisson processes with respective rates λ_1 and λ_2, and have service distributions G_1 and G_2. We suppose that type 1 customers are given service priority, in that service will never begin on a type 2 customer if a type 1 is waiting. However, if a type 2 is being served and a type 1 arrives, we assume that the service of the type 2 is continued until completion. That is, there is no preemption once service has begun.

Let W_Q^i denote the average wait in queue of a type i customer, $i = 1, 2$. Our objective is to compute the W_Q^i.

First, note that the total work in the system at any time would be exactly the same no matter what priority rule was employed (as long as the server is always busy whenever there are customers in the system). This is so since the work will always decrease at a rate of one per unit time when the server is busy (no matter who is in service) and will always jump by the service time of an arrival. Hence, the work in the system is exactly as it would be if there was no priority rule but rather a first-come, first-served (called FIFO) ordering. However, under FIFO the preceding model is just $M/G/1$ with

$$\lambda = \lambda_1 + \lambda_2,$$
$$G(x) = \frac{\lambda_1}{\lambda} G_1(x) + \frac{\lambda_2}{\lambda} G_2(x) \tag{8.41}$$

which follows since the combination of two independent Poisson processes is itself a Poisson process whose rate is the sum of the rates of the component processes. The service distribution G can be obtained by conditioning on which priority class the arrival is from—as is done in Equation (8.41).

Hence, from the results of Section 8.5, it follows that V, the average work in the priority queueing system, is given by

$$\begin{aligned} V &= \frac{\lambda E[S^2]}{2(1 - \lambda E[S])} \\ &= \frac{\lambda((\lambda_1/\lambda)E[S_1^2] + (\lambda_2/\lambda)E[S_2^2])}{2[1 - \lambda((\lambda_1/\lambda)E[S_1] + (\lambda_2/\lambda)E[S_2])]} \\ &= \frac{\lambda_1 E[S_1^2] + \lambda_2 E[S_2^2]}{2(1 - \lambda_1 E[S_1] - \lambda_2 E[S_2])} \end{aligned} \tag{8.42}$$

where S_i has distribution G_i, $i = 1, 2$.

Continuing in our quest for W_Q^i let us note that S and W_Q^*, the service and wait in queue of an arbitrary customer, are not independent in the priority model since knowledge about S gives us information as to the type of customer, which in turn gives us information about W_Q^*. To get around this we will compute separately the average amount of type 1 and type 2 work in the system. Denoting V^i as the average amount of type i work we have, exactly as in Section 8.5.1,

$$V^i = \lambda_i E[S_i] W_Q^i + \frac{\lambda_i E[S_i^2]}{2}, \qquad i = 1, 2 \tag{8.43}$$

If we define

$$V_Q^i \equiv \lambda_i E[S_i] W_Q^i,$$
$$V_S^i \equiv \frac{\lambda_i E[S_i^2]}{2}$$

then we may interpret V_Q^i as the average amount of type i work in queue, and V_S^i as the average amount of type i work in service (why?).

Now we are ready to compute W_Q^1. To do so, consider an arbitrary type 1 arrival. Then

$$\begin{aligned} \text{his delay} &= \text{amount of type 1 work in the system when he arrives} \\ &\quad + \text{amounts of type 2 work in service when he arrives} \end{aligned}$$

Taking expectations and using the fact that Poisson arrivals see time average yields

$$\begin{aligned} W_Q^1 &= V^1 + V_S^2 \\ &= \lambda_1 E[S_1] W_Q^1 + \frac{\lambda_1 E[S_1^2]}{2} + \frac{\lambda_2 E[S_2^2]}{2} \end{aligned} \tag{8.44}$$

or

$$W_Q^1 = \frac{\lambda_1 E[S_1^2] + \lambda_2 E[S_2^2]}{2(1 - \lambda_1 E[S_1])} \tag{8.45}$$

To obtain W_Q^2 we first note that since $V = V^1 + V^2$, we have from Equations (8.42) and (8.43) that

$$\begin{aligned} \frac{\lambda_1 E[S_1^2] + \lambda_2 E[S_2^2]}{2(1 - \lambda_1 E[S_1] - \lambda_2 E[S_2])} &= \lambda_1 E[S_1] W_Q^1 + \lambda_2 E[S_2] W_Q^2 \\ &\quad + \frac{\lambda_1 E[S_1^2]}{2} + \frac{\lambda_2 E[S_2^2]}{2} \\ &= W_Q^1 + \lambda_2 E[S_2] W_Q^2 \qquad \text{from Equation (8.44)} \end{aligned}$$

Now, using Equation (8.45), we obtain

$$\lambda_2 E[S_2] W_Q^2 = \frac{\lambda_1 E[S_1^2] + \lambda_2 E[S_2^2]}{2} \left[\frac{1}{1 - \lambda_1 E[S_1] - \lambda_2 E[S_2]} - \frac{1}{1 - \lambda_1 E[S_1]} \right]$$

or

$$W_Q^2 = \frac{\lambda_1 E[S_1^2] + \lambda_2 E[S_2^2]}{2(1 - \lambda_1 E[S_1] - \lambda_2 E[S_2])(1 - \lambda_1 E[S_1])} \tag{8.46}$$

Remarks

(i) Note that from Equation (8.45), the condition for W_Q^1 to be finite is that $\lambda_1 E[S_1] < 1$, which is independent of the type 2 parameters. (Is this intuitive?) For W_Q^2 to be finite, we need, from Equation (8.46), that

$$\lambda_1 E[S_1] + \lambda_2 E[S_2] < 1$$

Since the arrival rate of all customers is $\lambda = \lambda_1 + \lambda_2$, and the average service time of a customer is $(\lambda_1/\lambda)E[S_1] + (\lambda_2/\lambda)E[S_2]$, the preceding condition is just that the average arrival rate be less than the average service rate.

(ii) If there are n types of customers, we can solve for $V^j, j = 1, \ldots, n$, in a similar fashion. First, note that the total amount of work in the system of customers of types $1, \ldots, j$ is independent of the internal priority rule concerning types $1, \ldots, j$ and only depends on the fact that each of them is given priority over any customers of types $j + 1, \ldots, n$. (Why is this? Reason it out!) Hence, $V^1 + \cdots + V^j$ is the same as it would be if types $1, \ldots, j$ were considered as a single type I priority class and types $j + 1, \ldots, n$ as a single type II priority class. Now, from Equations (8.43) and (8.45),

$$V^{\mathrm{I}} = \frac{\lambda_{\mathrm{I}} E[S_{\mathrm{I}}^2] + \lambda_{\mathrm{I}} \lambda_{\mathrm{II}} E[S_{\mathrm{I}}] E[S_{\mathrm{II}}^2]}{2(1 - \lambda_{\mathrm{I}} E[S_{\mathrm{I}}])}$$

where

$$\begin{aligned}
\lambda_{\mathrm{I}} &= \lambda_1 + \cdots + \lambda_j, \\
\lambda_{\mathrm{II}} &= \lambda_{j+1} + \cdots + \lambda_n, \\
E[S_{\mathrm{I}}] &= \sum_{i=1}^{j} \frac{\lambda_i}{\lambda_{\mathrm{I}}} E[S_i], \\
E[S_{\mathrm{I}}^2] &= \sum_{i=1}^{j} \frac{\lambda_i}{\lambda_{\mathrm{I}}} E[S_i^2], \\
E[S_{\mathrm{II}}^2] &= \sum_{i=j+1}^{n} \frac{\lambda_i}{\lambda_{\mathrm{II}}} E[S_i^2]
\end{aligned}$$

Hence, as $V^{\mathrm{I}} = V^1 + \cdots + V^j$, we have an expression for $V^1 + \cdots + V^j$, for each $j = 1, \ldots, n$, which then can be solved for the individual $V^1, V^2, \ldots, V^n$. We now can obtain W_Q^i from Equation (8.43). The result of all this (which we leave for an exercise) is that

$$W_Q^i = \frac{\lambda_1 E[S_1^2] + \cdots + \lambda_n E[S_n^2]}{2 \prod_{j=i-1}^{i} (1 - \lambda_1 E[S_1] - \cdots - \lambda_j E[S_j])}, \qquad i = 1, \ldots, n \tag{8.47}$$

8.6.3 An M/G/1 Optimization Example

Consider a single-server system where customers arrive according to a Poisson process with rate λ, and where the service times are independent and have distribution function G. Let $\rho = \lambda E[S]$, where S represents a service time random variable, and suppose that $\rho < 1$. Suppose that the server departs whenever a busy period ends and does not return until there are n customers waiting. At that time the server returns and continues serving until the system is once again empty. If the system facility incurs costs at a rate of c per unit time per customer in the system, as well as a cost K each time the server returns, what value of $n, n \geqslant 1$, minimizes the long-run average cost per unit time incurred by the facility, and what is this minimal cost?

To answer the preceding, let us first determine $A(n)$, the average cost per unit time for the policy that returns the server whenever there are n customers waiting. To do so, say that a new cycle begins each time the server returns. As it is easy to see that everything probabilistically starts over when a cycle begins, it follows from the theory of renewal reward processes that if $C(n)$ is the cost incurred in a cycle and $T(n)$ is the time of a cycle, then

$$A(n) = \frac{E[C(n)]}{E[T(n)]}$$

To determine $E[C(n)]$ and $E[T(n)]$, consider the time interval of length, say, T_i, starting from the first time during a cycle that there are a total of i customers in the system until the first time afterward that there are only $i-1$. Therefore, $\sum_{i=1}^{n} T_i$ is the amount of time that the server is busy during a cycle. Adding the additional mean idle time until n customers are in the system gives

$$E[T(n)] = \sum_{i=1}^{n} E[T_i] + n/\lambda$$

Now, consider the system at the moment when a service is about to begin and there are $i-1$ customers waiting in queue. Since service times do not depend on the order in which customers are served, suppose that the order of service is last come first served, implying that service does not begin on the $i-1$ presently in queue until these $i-1$ are the only ones in the system. Thus, we see that the time that it takes to go from i customers in the system to $i-1$ has the same distribution as the time it takes the $M/G/1$ system to go from a single customer (just beginning service) to empty; that is, its distribution is that of B, the length of an $M/G/1$ busy period. (Essentially the same argument was made in Example 5.25.) Hence,

$$E[T_i] = E[B] = \frac{E[S]}{1-\rho}$$

implying that

$$E[T(n)] = \frac{nE[S]}{1-\lambda E[S]} + \frac{n}{\lambda} = \frac{n}{\lambda(1-\rho)} \tag{8.48}$$

To determine $E[C(n)]$, let C_i denote the cost incurred during the interval of length T_i that starts with $i-1$ in queue and a service just beginning and ends when the $i-1$ in queue are the only customers in the system. Thus, $K + \sum_{i=1}^{n} C_i$ represents the total cost incurred during the busy part of the cycle. In addition, during the idle part of the cycle there will be i customers in the system for an exponential time with rate $\lambda, i = 1, \ldots, n-1$, resulting in an expected cost of

$c(1 + \cdots + n - 1)/\lambda$. Consequently,

$$E[C(n)] = K + \sum_{i=1}^{n} E[C_i] + \frac{n(n-1)c}{2\lambda} \tag{8.49}$$

To find $E[C_i]$, consider the moment when the interval of length T_i begins, and let W_i be the sum of the initial service time plus the sum of the times spent in the system by all the customers that arrive (and are served) until the moment when the interval ends and there are only $i - 1$ customers in the system. Then,

$$C_i = (i-1)cT_i + cW_i$$

where the first term refers to the cost incurred due to the $i - 1$ customers in queue during the interval of length T_i. As it is easy to see that W_i has the same distribution as W_b, the sum of the times spent in the system by all arrivals in an $M/G/1$ busy period, we obtain

$$E[C_i] = (i-1)c\frac{E[S]}{1-\rho} + cE[W_b] \tag{8.50}$$

Using Equation (8.49), this yields

$$\begin{aligned}
E[C(n)] &= K + \frac{n(n-1)cE[S]}{2(1-\rho)} + ncE[W_b] + \frac{n(n-1)c}{2\lambda} \\
&= K + ncE[W_b] + \frac{n(n-1)c}{2\lambda}\left(\frac{\rho}{1-\rho} + 1\right) \\
&= K + ncE[W_b] + \frac{n(n-1)c}{2\lambda(1-\rho)}
\end{aligned}$$

Utilizing the preceding in conjunction with Equation (8.48) shows that

$$A(n) = \frac{K\lambda(1-\rho)}{n} + \lambda c(1-\rho)E[W_b] + \frac{c(n-1)}{2} \tag{8.51}$$

To determine $E[W_b]$, we use the result that the average amount of time spent in the system by a customer in the $M/G/1$ system is

$$W = W_Q + E[S] = \frac{\lambda E[S^2]}{2(1-\rho)} + E[S]$$

However, if we imagine that on day $j, j \geqslant 1$, we earn an amount equal to the total time spent in the system by the jth arrival at the $M/G/1$ system, then it follows from renewal reward processes (since everything probabilistically restarts at the end of a busy period) that

$$W = \frac{E[W_b]}{E[N]}$$

where N is the number of customers served in an $M/G/1$ busy period. Since $E[N] = 1/(1-\rho)$ we see that

$$(1-\rho)E[W_b] = W = \frac{\lambda E[S^2]}{2(1-\rho)} + E[S]$$

Therefore, using Equation (8.51), we obtain

$$A(n) = \frac{K\lambda(1-\rho)}{n} + \frac{c\lambda^2 E[S^2]}{2(1-\rho)} + c\rho + \frac{c(n-1)}{2}$$

To determine the optimal value of n, treat n as a continuous variable and differentiate the preceding to obtain

$$A'(n) = \frac{-K\lambda(1-\rho)}{n^2} + \frac{c}{2}$$

Setting this equal to 0 and solving yields that the optimal value of n is

$$n^* = \sqrt{\frac{2K\lambda(1-\rho)}{c}}$$

and the minimal average cost per unit time is

$$A(n^*) = \sqrt{2\lambda K(1-\rho)c} + \frac{c\lambda^2 E[S^2]}{2(1-\rho)} + c\rho - \frac{c}{2}$$

It is interesting to see how close we can come to the minimal average cost when we use a simpler policy of the following form: Whenever the server finds the system empty of customers she departs and then returns after a fixed time t has elapsed. Let us say that a new cycle begins each time the server departs. Both the expected costs incurred during the idle and the busy parts of a cycle are obtained by conditioning on $N(t)$, the number of arrivals in the time t that the server is gone. With $\bar{C}(t)$ being the cost incurred during a cycle, we obtain

$$\begin{aligned} E[\bar{C}(t) \mid N(t)] &= K + \sum_{i=1}^{N(t)} E[C_i] + cN(t)\frac{t}{2} \\ &= K + \frac{N(t)(N(t)-1)cE[S]}{2(1-\rho)} + N(t)cE[W_b] + cN(t)\frac{t}{2} \end{aligned}$$

The final term of the first equality is the conditional expected cost during the idle time in the cycle and is obtained by using that, given the number of arrivals in the time t, the arrival times are independent and uniformly distributed on $(0, t)$; the second equality used Equation (8.50). Since $N(t)$ is Poisson with mean λt, it follows that $E[N(t)(N(t)-1)] = E[N^2(t)] - E[N(t)] = \lambda^2 t^2$. Thus, taking the expected value of the preceding gives

$$\begin{aligned} E[\bar{C}(t)] &= K + \frac{\lambda^2 t^2 c E[S]}{2(1-\rho)} + \lambda t c E[W_b] + \frac{c\lambda t^2}{2} \\ &= K + \frac{c\lambda t^2}{2(1-\rho)} + \lambda t c E[W_b] \end{aligned}$$

Similarly, if $\bar{T}(t)$ is the time of a cycle, then

$$\begin{aligned} E[\bar{T}(t)] &= E[E[\bar{T}(t)|N(t)]] \\ &= E[t + N(t)E[B]] \\ &= t + \frac{\rho t}{1-\rho} \\ &= \frac{t}{1-\rho} \end{aligned}$$

Hence, the average cost per unit time, call it $\bar{A}(t)$, is

$$\begin{aligned} \bar{A}(t) &= \frac{E[\bar{C}(t)]}{E[\bar{T}(t)]} \\ &= \frac{K(1-\rho)}{t} + \frac{c\lambda t}{2} + c\lambda(1-\rho)E[W_b] \end{aligned}$$

Thus, from Equation (8.51), we see that

$$\bar{A}(n/\lambda) - A(n) = c/2$$

which shows that allowing the return decision to depend on the number presently in the system can reduce the average cost only by the amount $c/2$. ■

8.6.4 The M/G/1 Queue with Server Breakdown

Consider a single server queue in which customers arrive according to a Poisson process with rate λ, and where the amount of service time required by each customer has distribution G. Suppose, however, that when working the server breaks down at an exponential rate α. That is, the probability a working server will be able to work for an additional time t without breaking down is $e^{-\alpha t}$.

When the server breaks down, it immediately goes to the repair facility. The repair time is a random variable with distribution H. Suppose that the customer in service when a breakdown occurs has its service continue, when the sever returns, from the point it was at when the breakdown occurred. (Therefore, the total amount of time a customer is actually receiving service from a working server has distribution G.)

By letting a customer's "service time" include the time that the customer is waiting for the server to come back from being repaired, the preceding is an $M/G/1$ queue. If we let T denote the amount of time from when a customer first enters service until it departs the system, then T is a service time random variable of this $M/G/1$ queue. The average amount of time a customer spends waiting in queue before its service first commences is, thus,

$$W_Q = \frac{\lambda E[T^2]}{2(1 - \lambda E[T])}$$

To compute $E[T]$ and $E[T^2]$, let S, having distribution G, be the service requirement of the customer; let N denote the number of times that the server breaks down while the customer is in service; let $R_1, R_2, \ldots$ be the amounts of time the server spends in the repair facility on its successive visits. Then,

$$T = \sum_{i=1}^{N} R_i + S$$

Conditioning on S yields

$$E[T|S = s] = E\left[\sum_{i=1}^{N} R_i \middle| S = s\right] + s,$$
$$\text{Var}(T|S = s) = \text{Var}\left(\sum_{i=1}^{N} R_i \middle| S = s\right)$$

Now, a working server always breaks down at an exponential rate α. Therefore, given that a customer requires s units of service time, it follows that the number of server breakdowns while that customer is being served is a Poisson random variable with mean αs. Consequently, conditional on $S = s$, the random variable $\sum_{i=1}^{N} R_i$ is a compound Poisson random variable with Poisson mean αs. Using the results from Examples 3.10 and 3.17, we thus obtain

$$E\left[\sum_{i=1}^{N} R_i \middle| S = s\right] = \alpha s E[R], \qquad \text{Var}\left(\sum_{i=1}^{N} R_i \middle| S = s\right) = \alpha s E[R^2]$$

where R has the repair distribution H. Therefore,

$$\begin{aligned}E[T|S] &= \alpha SE[R] + S = S(1 + \alpha E[R]),\\ \text{Var}(T|S) &= \alpha SE[R^2]\end{aligned}$$

Thus,

$$E[T] = E[E[T|S]] = E[S](1 + \alpha E[R])$$

and, by the conditional variance formula,

$$\begin{aligned}\text{Var}(T) &= E[\text{Var}(T|S)] + \text{Var}(E[T|S])\\ &= \alpha E[S]E[R^2] + (1 + \alpha E[R])^2\text{Var}(S)\end{aligned}$$

Therefore,

$$\begin{aligned}E[T^2] &= \text{Var}(T) + (E[T])^2\\ &= \alpha E[S]E[R^2] + (1 + \alpha E[R])^2 E[S^2]\end{aligned}$$

Consequently, assuming that $\lambda E[T] = \lambda E[S](1 + \alpha E[R]) < 1$, we obtain

$$W_Q = \frac{\lambda\alpha E[S]E[R^2] + \lambda(1 + \alpha E[R])^2 E[S^2]}{2(1 - \lambda E[S](1 + \alpha E[R]))}$$

From the preceding, we can now obtain

$$\begin{aligned}L_Q &= \lambda W_Q,\\ W &= W_Q + E[T],\\ L &= \lambda W\end{aligned}$$

Some other quantities we might be interested in are

(i) P_w, the proportion of time the server is working;
(ii) P_r, the proportion of time the server is being repaired;
(iii) P_I, the proportion of time the server is idle.

These quantities can all be obtained by using the queueing cost identity. For instance, if we suppose that customers pay 1 per unit time while actually being served, then

$$\begin{aligned}\text{average rate at which system earns} &= P_w,\\ \text{average amount a customer pays} &= E[S]\end{aligned}$$

Therefore, the identity yields

$$P_w = \lambda E[S]$$

To determine P_r, suppose a customer whose service is interrupted pays 1 per unit time while the server is being repaired. Then,

$$\text{average rate at which system earns} = P_r,$$
$$\text{average amount a customer pays} = E\left[\sum_{i=1}^{N} R_i\right] = \alpha E[S]E[R]$$

yielding

$$P_r = \lambda\alpha E[S]E[R]$$

Of course, P_I can be obtained from

$$P_I = 1 - P_w - P_r$$

Remark The quantities P_w and P_r could also have been obtained by first noting that $1 - P_0 = \lambda E[T]$ is the proportion of time the server is either working or in repair. Thus,

$$P_w = \lambda E[T]\frac{E[S]}{E[T]} = \lambda E[S],$$
$$P_r = \lambda E[T]\frac{E[T] - E[S]}{E[T]} = \lambda E[S]\alpha E[R] \qquad \blacksquare$$

8.7 The Model *G/M/*1

The model $G/M/1$ assumes that the times between successive arrivals have an arbitrary distribution G. The service times are exponentially distributed with rate μ and there is a single server.

The immediate difficulty in analyzing this model stems from the fact that the number of customers in the system is not informative enough to serve as a state space. For in summarizing what has occurred up to the present we would need to know not only the number in the system, but also the amount of time that has elapsed since the last arrival (since G is not memoryless). (Why need we not be concerned with the amount of time the person being served has already spent in service?) To get around this problem we shall only look at the system when a customer arrives; and so let us define $X_n, n \geqslant 1$, by

$$X_n \equiv \text{the number in the system as seen by the } n\text{th arrival}$$

It is easy to see that the process $\{X_n, n \geqslant 1\}$ is a Markov chain. To compute the transition probabilities P_{ij} for this Markov chain let us first note that, as long as there are customers to be served, the number of services in any length of time t is a Poisson random variable with mean μt. This is true since the time between successive services is exponential and, as we know, this implies that the number of services thus constitutes a Poisson process. Hence,

$$P_{i,i+1-j} = \int_0^\infty e^{-\mu t} \frac{(\mu t)^j}{j!}\, dG(t), \qquad j = 0, 1, \ldots, i$$

which follows since if an arrival finds i in the system, then the next arrival will find $i + 1$ minus the number served, and the probability that j will be served is easily seen to equal the right side of the preceding (by conditioning on the time between the successive arrivals).

The formula for P_{i0} is a little different (it is the probability that *at least* $i + 1$ Poisson events occur in a random length of time having distribution G) and can be obtained from

$$P_{i0} = 1 - \sum_{j=0}^{i} P_{i,i+1-j}$$

The limiting probabilities $\pi_k, k = 0, 1, \ldots$, can be obtained as the unique solution of

$$\pi_k = \sum_i \pi_i P_{ik}, \qquad k \geqslant 0,$$
$$\sum_i \pi_k = 1$$

which, in this case, reduce to

$$\begin{aligned} \pi_k &= \sum_{i=k-1}^{\infty} \pi_i \int_0^\infty e^{-\mu t} \frac{(\mu t)^{i+1-k}}{(i+1-k)!}\, dG(t), \qquad k \geqslant 1, \\ \sum_0^\infty \pi_k &= 1 \end{aligned} \tag{8.52}$$

(We have not included the equation $\pi_0 = \sum \pi_i P_{i0}$ since one of the equations is always redundant.)

To solve the preceding, let us try a solution of the form $\pi_k = c\beta^k$. Substitution into Equation (8.52) leads to

$$\begin{aligned} c\beta^k &= c \sum_{i=k-1}^{\infty} \beta^i \int_0^{\infty} e^{-\mu t} \frac{(\mu t)^{i+1-k}}{(i+1-k)!} \, dG(t) \\ &= c \int_0^{\infty} e^{-\mu t} \beta^{k-1} \sum_{i=k-1}^{\infty} \frac{(\beta \mu t)^{i+1-k}}{(i+1-k)!} \, dG(t) \end{aligned} \tag{8.53}$$

However,

$$\begin{aligned} \sum_{i=k-1}^{\infty} \frac{(\beta \mu t)^{i+1-k}}{(i+1-k)!} &= \sum_{j=0}^{\infty} \frac{(\beta \mu t)^j}{j!} \\ &= e^{\beta \mu t} \end{aligned}$$

and thus Equation (8.53) reduces to

$$\beta^k = \beta^{k-1} \int_0^{\infty} e^{-\mu t(1-\beta)} \, dG(t)$$

or

$$\beta = \int_0^{\infty} e^{-\mu t(1-\beta)} \, dG(t) \tag{8.54}$$

The constant c can be obtained from $\sum_k \pi_k = 1$, which implies that

$$c \sum_0^{\infty} \beta^k = 1$$

or

$$c = 1 - \beta$$

As (π_k) is the *unique* solution to Equation (8.52), and $\pi_k = (1-\beta)\beta^k$ satisfies, it follows that

$$\pi_k = (1-\beta)\beta^k, \qquad k = 0, 1, \ldots$$

where β is the solution of Equation (8.54). (It can be shown that if the mean of G is greater than the mean service time $1/\mu$, then there is a unique value of

β satisfying Equation (8.54) which is between 0 and 1.) The exact value of β usually can only be obtained by numerical methods.

As π_k is the limiting probability that an arrival sees k customers, it is just the a_k as defined in Section 8.2. Hence,

$$\alpha_k = (1-\beta)\beta^k, \qquad k \geqslant 0 \tag{8.55}$$

We can obtain W by conditioning on the number in the system when a customer arrives. This yields

$$\begin{aligned} W &= \sum_k E[\text{time in system} \mid \text{arrival sees } k](1-\beta)\beta^k \\ &= \sum_k \frac{k+1}{\mu}(1-\beta)\beta^k \qquad \begin{array}{l}\text{(Since if an arrival sees } k \text{ then he spends} \\ \; k+1 \text{ service periods in the system)}\end{array} \\ &= \frac{1}{\mu(1-\beta)} \qquad \left(\text{by using } \sum_0^\infty kx^k = \frac{x}{(1-x)^2}\right) \end{aligned}$$

and

$$\begin{aligned} W_Q &= W - \frac{1}{\mu} = \frac{\beta}{\mu(1-\beta)}, \\ L &= \lambda W = \frac{\lambda}{\mu(1-\beta)}, \\ L_Q &= \lambda W_Q = \frac{\lambda\beta}{\mu(1-\beta)} \end{aligned} \tag{8.56}$$

where λ is the reciprocal of the mean interarrival time. That is,

$$\frac{1}{\lambda} = \int_0^\infty x\, dG(x)$$

In fact, in exactly the same manner as shown for the $M/M/1$ in Section 8.3.1 and Exercise 4 we can show that

$$W^* \text{ is exponential with rate } \mu(1-\beta),$$

$$W_Q^* = \begin{cases} 0 & \text{with probability } 1-\beta \\ \text{exponential with rate } \mu(1-\beta) & \text{with probability } \beta \end{cases}$$

where W^* and W_Q^* are the amounts of time that a customer spends in system and queue, respectively (their means are W and W_Q).

Whereas $a_k = (1-\beta)\beta^k$ is the probability that an arrival sees k in the system, it is not equal to the proportion of time during which there are k in the system (since the arrival process is not Poisson). To obtain the P_k we first note that the rate at which the number in the system changes from $k-1$ to k must equal the rate at which it changes from k to $k-1$ (why?). Now the rate at which it changes from $k-1$ to k is equal to the arrival rate λ multiplied by the proportion of arrivals finding $k-1$ in the system. That is,

$$\text{rate number in system goes from } k-1 \text{ to } k = \lambda a_{k-1}$$

Similarly, the rate at which the number in the system changes from k to $k-1$ is equal to the proportion of time during which there are k in the system multiplied by the (constant) service rate. That is,

$$\text{rate number in system goes from } k \text{ to } k-1 = P_k \mu$$

Equating these rates yields

$$P_k = \frac{\lambda}{\mu} a_{k-1}, \qquad k \geqslant 1$$

and so, from Equation (8.55),

$$P_k = \frac{\lambda}{\mu}(1-\beta)\beta^{k-1}, \qquad k \geqslant 1$$

and, as $P_0 = 1 - \sum_{k=1}^{\infty} P_k$, we obtain

$$P_0 = 1 - \frac{\lambda}{\mu}$$

Remarks In the foregoing analysis we guessed at a solution of the stationary probabilities of the Markov chain of the form $\pi_k = c\beta^k$, then verified such a solution by substituting in the stationary Equation (8.52). However, it could have been argued directly that the stationary probabilities of the Markov chain are of this form. To do so, define β_i to be the expected number of times that state $i+1$ is visited in the Markov chain between two successive visits to state $i, i \geqslant 0$. Now it is not difficult to see (and we will let you argue it out for yourself) that

$$\beta_0 = \beta_1 = \beta_2 = \cdots = \beta$$

Now it can be shown by using renewal reward processes that

$$\begin{aligned}\pi_{i+1} &= \frac{E[\text{number of visits to state } i+1 \text{ in an } i\text{–}i \text{ cycle}]}{E[\text{number of transitions in an } i\text{–}i \text{ cycle}]}\\ &= \frac{\beta_i}{1/\pi_i}\end{aligned}$$

and so,

$$\pi_{i+1} = \beta_i \pi_i = \beta \pi_i, \qquad i \geqslant 0$$

implying, since $\sum_0^\infty \pi_i = 1$, that

$$\pi_i = \beta^i (1 - \beta), \qquad i \geqslant 0$$

8.7.1 The G/M/1 Busy and Idle Periods

Suppose that an arrival has just found the system empty—and so initiates a busy period—and let N denote the number of customers served in that busy period. Since the Nth arrival (after the initiator of the busy period) will also find the system empty, it follows that N is the number of transitions for the Markov chain (of Section 8.7) to go from state 0 to state 0. Hence, $1/E[N]$ is the proportion of transitions that take the Markov chain into state 0; or equivalently, it is the proportion of arrivals that find the system empty. Therefore,

$$E[N] = \frac{1}{a_0} = \frac{1}{1-\beta}$$

Also, as the next busy period begins after the Nth interarrival, it follows that the cycle time (that is, the sum of a busy and idle period) is equal to the time until the Nth interarrival. In other words, the sum of a busy and idle period can be expressed as the sum of N interarrival times. Thus, if T_i is the ith interarrival time after the busy period begins, then

$$\begin{aligned} E[\text{Busy}] + E[\text{Idle}] &= E\left[\sum_{i=1}^{N} T_i\right] \\ &= E[N]E[T] \qquad \text{(by Wald's equation)} \\ &= \frac{1}{\lambda(1-\beta)} \end{aligned} \tag{8.57}$$

For a second relation between $E[\text{Busy}]$ and $E[\text{Idle}]$, we can use the same argument as in Section 8.5.3 to conclude that

$$1 - P_0 = \frac{E[\text{Busy}]}{E[\text{Idle}] + E[\text{Busy}]}$$

and since $P_0 = 1 - \lambda/\mu$, we obtain, upon combining this with (8.57), that

$$\begin{aligned} E[\text{Busy}] &= \frac{1}{\mu(1-\beta)}, \\ E[\text{Idle}] &= \frac{\mu - \lambda}{\lambda\mu(1-\beta)} \end{aligned}$$

8.8 A Finite Source Model

Consider a system of m machines, whose working times are independent exponential random variables with rate λ. Upon failure, a machine instantly goes to a repair facility that consists of a single repairperson. If the repairperson is free, repair begins on the machine; otherwise, the machine joins the queue of failed machines. When a machine is repaired it becomes a working machine, and repair begins on a new machine from the queue of failed machines (provided the queue is nonempty). The successive repair times are independent random variables having density function g, with mean

$$\mu_R = \int_0^\infty xg(x)\, dx$$

To analyze this system, so as to determine such quantities as the average number of machines that are down and the average time that a machine is down, we will exploit the exponentially distributed working times to obtain a Markov chain. Specifically, let X_n denote the number of failed machines immediately after the nth repair occurs, $n \geqslant 1$. Now, if $X_n = i > 0$, then the situation when the nth repair has just occurred is that repair is about to begin on a machine, there are $i-1$ other machines waiting for repair, and there are $m-i$ working machines, each of which will (independently) continue to work for an exponential time with rate λ. Similarly, if $X_n = 0$, then all m machines are working and will (independently) continue to do so for exponentially distributed times with rate λ. Consequently, any information about earlier states of the system will not affect the probability distribution of the number of down machines at the moment of the next repair completion; hence, $\{X_n, n \geqslant 1\}$ is a Markov chain. To determine its transition probabilities $P_{i,j}$, suppose first that $i > 0$. Conditioning on R, the length of the next repair time, and making use of the independence of the $m - i$ remaining working times, yields that for $j \leqslant m - i$

$$\begin{aligned} P_{i,i-1+j} &= P\{j \text{ failures during } R\} \\ &= \int_0^\infty P\{j \text{ failures during } R \mid R = r\} g(r)\, dr \\ &= \int_0^\infty \binom{m-i}{j} (1 - e^{-\lambda r})^j (e^{-\lambda r})^{m-i-j} g(r)\, dr \end{aligned}$$

If $i = 0$, then, because the next repair will not begin until one of the machines fails,

$$P_{0,j} = P_{1,j}, \qquad j \leqslant m - 1$$

Let $\pi_j, j = 0, \ldots, m-1$, denote the stationary probabilities of this Markov chain. That is, they are the unique solution of

$$\pi_j = \sum_i \pi_i P_{i,j},$$

$$\sum_{j=0}^{m-1} \pi_j = 1$$

Therefore, after explicitly determining the transition probabilities and solving the preceding equations, we would know the value of π_0, the proportion of repair completions that leaves all machines working. Let us say that the system is "on" when all machines are working and "off" otherwise. (Thus, the system is on when the repairperson is idle and off when he is busy.) As all machines are working when the system goes back on, it follows from the lack of memory property of the exponential that the system probabilistically starts over when it goes on. Hence, this on–off system is an alternating renewal process. Suppose that the system has just become on, thus starting a new cycle, and let $R_i, i \geqslant 1$, be the time of the ith repair from that moment. Also, let N denote the number of repairs in the off (busy) time of the cycle. Then, it follows that B, the length of the off period, can be expressed as

$$B = \sum_{i=1}^{N} R_i$$

Although N is not independent of the sequence $R_1, R_2, \ldots$, it is easy to check that it is a stopping time for this sequence, and thus by Wald's equation (see Exercise 13 of Chapter 7) we have

$$E[B] = E[N]E[R] = E[N]\mu_R$$

Also, since an on time will last until one of the machines fails, and since the minimum of independent exponential random variables is exponential with a rate equal to the sum of their rates, it follows that $E[I]$, the mean on (idle) time in a cycle, is given by

$$E[I] = 1/(m\lambda)$$

Hence, P_B, the proportion of time that the repairperson is busy, satisfies

$$P_B = \frac{E[N]\mu_R}{E[N]\mu_R + 1/(m\lambda)}$$

However, since, on average, one out of every $E[N]$ repair completions will leave all machines working, it follows that

$$\pi_0 = \frac{1}{E[N]}$$

Consequently,

$$P_B = \frac{\mu_R}{\mu_R + \pi_0/(m\lambda)} \tag{8.58}$$

Now focus attention on one of the machines, call it machine number 1, and let $P_{1,R}$ denote the proportion of time that machine 1 is being repaired. Since the proportion of time that the repairperson is busy is P_B, and since all machines fail at the same rate and have the same repair distribution, it follows that

$$P_{1,R} = \frac{P_B}{m} = \frac{\mu_R}{m\mu_R + \pi_0/\lambda} \tag{8.59}$$

However, machine 1 alternates between time periods when it is working, when it is waiting in queue, and when it is in repair. Let W_i, Q_i, S_i denote, respectively, the ith working time, the ith queueing time, and the ith repair time of machine 1, $i \geqslant 1$. Then, the proportion of time that machine 1 is being repaired during its first n working–queue–repair cycles is as follows:

$$\begin{aligned}
&\text{proportion of time in the first } n \text{ cycles that machine 1 is being repaired}\\
&= \frac{\sum_{i=1}^n S_i}{\sum_{i=1}^n W_i + \sum_{i=1}^n Q_i + \sum_{i=1}^n S_i}\\
&= \frac{\sum_{i=1}^n S_i/n}{\sum_{i=1}^n W_i/n + \sum_{i=1}^n Q_i/n + \sum_{i=1}^n S_i/n}
\end{aligned}$$

Letting $n \to \infty$ and using the strong law of large numbers to conclude that the averages of the W_i and of the S_i converge, respectively, to $1/\lambda$ and μ_R, yields

$$P_{1,R} = \frac{\mu_R}{1/\lambda + \bar{Q} + \mu_R}$$

where $\bar{Q}$ is the average amount of time that machine 1 spends in queue when it fails. Using Equation (8.59), the preceding gives

$$\frac{\mu_R}{m\mu_R + \pi_0/\lambda} = \frac{\mu_R}{1/\lambda + \bar{Q} + \mu_R}$$

or, equivalently, that

$$\bar{Q} = (m-1)\mu_R - (1-\pi_0)/\lambda$$

Moreover, since all machines are probabilistically equivalent it follows that $\bar{Q}$ is equal to W_Q, the average amount of time that a failed machine spends in queue. To determine the average number of machines in queue, we will make use of the basic queueing identity

$$L_Q = \lambda_a W_Q = \lambda_a \bar{Q}$$

where λ_a is the average rate at which machines fail. To determine λ_a, again focus attention on machine 1 and suppose that we earn one per unit time whenever machine 1 is being repaired. It then follows from the basic cost identity of Equation (8.1) that

$$P_{1,R} = r_1 \mu_R$$

where r_1 is the average rate at which machine 1 fails. Thus, from Equation (8.59), we obtain

$$r_1 = \frac{1}{m\mu_R + \pi_0/\lambda}$$

Because all m machines fail at the same rate, the preceding implies that

$$\lambda_a = m r_1 = \frac{m}{m\mu_R + \pi_0/\lambda}$$

which gives that the average number of machines in queue is

$$L_Q = \frac{m(m-1)\mu_R - m(1-\pi_0)/\lambda}{m\mu_R + \pi_0/\lambda}$$

Since the average number of machines being repaired is P_B, the preceding, along with Equation (8.58), shows that the average number of down machines is

$$L = L_Q + P_B = \frac{m^2\mu_R - m(1-\pi_0)/\lambda}{m\mu_R + \pi_0/\lambda}$$

8.9 Multiserver Queues

By and large, systems that have more than one server are much more difficult to analyze than those with a single server. In Section 8.9.1 we start first with a Poisson arrival system in which no queue is allowed, and then consider in Section 8.9.2 the infinite capacity $M/M/k$ system. For both of these models we are able to present the limiting probabilities. In Section 8.9.3 we consider the model $G/M/k$. The analysis here is similar to that of the $G/M/1$ (Section 8.7) except that in place of a single quantity β given as the solution of an integral equation, we have k such quantities. We end in Section 8.9.4 with the model $M/G/k$ for which unfortunately our previous technique (used in $M/G/1$) no longer enables us to derive W_Q, and we content ourselves with an approximation.

8.9.1 Erlang's Loss System

A loss system is a queueing system in which arrivals that find all servers busy do not enter but rather are lost to the system. The simplest such system is the $M/M/k$ loss system in which customers arrive according to a Poisson process having rate λ, enter the system if at least one of the k servers is free, and then spend an exponential amount of time with rate μ being served. The balance equations for this system are

State	*Rate leave = rate enter*
0	$\lambda P_0 = \mu P_1$
1	$(\lambda + \mu)P_1 = 2\mu P_2 + \lambda P_0$
2	$(\lambda + 2\mu)P_2 = 3\mu P_3 + \lambda P_1$
$i, 0 < i < k$	$(\lambda + i\mu)P_i = (i+1)\mu P_{i+1} + \lambda P_{i-1}$
k	$k\mu P_k = \lambda P_{k-1}$

Rewriting gives

$$\begin{aligned} \lambda P_0 &= \mu P_1, \\ \lambda P_1 &= 2\mu P_2, \\ \lambda P_2 &= 3\mu P_3, \\ &\ \vdots \\ \lambda P_{k-1} &= k\mu P_k \end{aligned}$$

or

$$\begin{aligned} P_1 &= \frac{\lambda}{\mu} P_0, \\ P_2 &= \frac{\lambda}{2\mu} P_1 = \frac{(\lambda/\mu)^2}{2} P_0, \end{aligned}$$

$$P_3 = \frac{\lambda}{3\mu} P_2 = \frac{(\lambda/\mu)^3}{3!} P_0,$$

$$\vdots$$

$$P_k = \frac{\lambda}{k\mu} P_{k-1} = \frac{(\lambda/\mu)^k}{k!} P_0$$

and using $\sum_0^k P_i = 1$, we obtain

$$P_i = \frac{(\lambda/\mu)^i/i!}{\sum_{j=0}^k (\lambda/\mu)^j/j!}, \qquad i = 0, 1, \ldots, k$$

Since $E[S] = 1/\mu$, where $E[S]$ is the mean service time, the preceding can be written as

$$P_i = \frac{(\lambda E[S])^i/i!}{\sum_{j=0}^k (\lambda E[S])^j/j!}, \qquad i = 0, 1, \ldots, k \tag{8.60}$$

Consider now the same system except that the service distribution is general—that is, consider the $M/G/k$ with no queue allowed. This model is sometimes called the *Erlang loss system*. It can be shown (though the proof is advanced) that Equation (8.60) (which is called *Erlang's loss formula*) remains valid for this more general system.

8.9.2 The M/M/k Queue

The $M/M/k$ infinite capacity queue can be analyzed by the balance equation technique. We leave it for you to verify that

$$P_i = \begin{cases} \dfrac{\dfrac{(\lambda/\mu)^i}{i!}}{\displaystyle\sum_{i=0}^{k-1} \frac{(\lambda/\mu)^i}{i!} + \frac{(\lambda/\mu)^k}{k!} \frac{k\mu}{k\mu - \lambda}}, & i \leqslant k \\[2ex] \dfrac{(\lambda/k\mu)^i k^k}{k!} P_0, & i > k \end{cases}$$

We see from the preceding that we need to impose the condition $\lambda < k\mu$.

8.9.3 The G/M/k *Queue*

In this model we again suppose that there are k servers, each of whom serves at an exponential rate μ. However, we now allow the time between successive arrivals to have an arbitrary distribution G. To ensure that a steady-state (or limiting) distribution exists, we assume the condition $1/\mu_G < k\mu$ where μ_G is the mean of G.*

The analysis for this model is similar to that presented in Section 8.7 for the case $k = 1$. Namely, to avoid having to keep track of the time since the last arrival, we look at the system only at arrival epochs. Once again, if we define X_n as the number in the system at the moment of the nth arrival, then $\{X_n, n \geqslant 0\}$ is a Markov chain.

To derive the transition probabilities of the Markov chain, it helps to first note the relationship

$$X_{n+1} = X_n + 1 - Y_n, \qquad n \geqslant 0$$

where Y_n denotes the number of departures during the interarrival time between the nth and $(n + 1)$st arrival. The transition probabilities P_{ij} can now be calculated as follows:

Case 1: $j > i + 1$.

In this case it easily follows that $P_{ij} = 0$.

Case 2: $j \leqslant i + 1 \leqslant k$.

In this case if an arrival finds i in the system, then as $i < k$ the new arrival will also immediately enter service. Hence, the next arrival will find j if of the $i + 1$ services exactly $i + 1 - j$ are completed during the interarrival time. Conditioning on the length of this interarrival time yields

$$\begin{aligned} P_{ij} &= P\{i + 1 - j \text{ of } i + 1 \text{ services are completed in an interarrival time}\} \\ &= \int_0^\infty P\{i + 1 - j \text{ of } i + 1 \text{ are completed} | \text{interarrival time is } t\}\, dG(t) \\ &= \int_0^\infty \binom{i+1}{j} (i - e^{-\mu t})^{i+1-j} (e^{-\mu t})^j \, dG(t) \end{aligned}$$

where the last equality follows since the number of service completions in a time t will have a binomial distribution.

* It follows from the renewal theory (Proposition 7.1) that customers arrive at rate $1/\mu_G$, and as the maximum service rate is $k\mu$, we clearly need that $1/\mu_G < k\mu$ for limiting probabilities to exist.

Case 3: $i+1 \geqslant j \geqslant k$.

To evaluate P_{ij} in this case we first note that when all servers are busy, the departure process is a Poisson process with rate $k\mu$ (why?). Hence, again conditioning on the interarrival time we have

$$\begin{aligned}P_{ij} &= P\{i+1-j \text{ departures}\}\\ &= \int_0^\infty P\{i+1-j \text{ departures in time } t\}\, dG(t)\\ &= \int_0^\infty e^{-k\mu t}\frac{(k\mu t)^{i+1-j}}{(i+1-j)!}\, dG(t)\end{aligned}$$

Case 4: $i+1 \geqslant k > j$.

In this case since when all servers are busy the departure process is a Poisson process, it follows that the length of time until there will only be k in the system will have a gamma distribution with parameters $i+1-k, k\mu$ (the time until $i+1-k$ events of a Poisson process with rate $k\mu$ occur is gamma distributed with parameters $i+1-k, k\mu$). Conditioning first on the interarrival time and then on the time until there are only k in the system (call this latter random variable T_k) yields

$$\begin{aligned}P_{ij} &= \int_0^\infty P\{i+1-j \text{ departures in time } t\}\, dG(t)\\ &= \int_0^\infty \int_0^t P\{i+1-j \text{ departures in } t \mid T_k = s\} k\mu e^{-k\mu s}\frac{(k\mu s)^{i-k}}{(i-k)!}\, ds\, dG(t)\\ &= \int_0^\infty \int_0^t \binom{k}{j}\left(1-e^{-\mu(t-s)}\right)^{k-j}\left(e^{-\mu(t-s)}\right)^{j} k\mu e^{-k\mu s}\frac{(k\mu s)^{i-k}}{(i-k)!}\, ds\, dG(t)\end{aligned}$$

where the last equality follows since of the k people in service at time s the number whose service will end by time t is binomial with parameters k and $1-e^{-\mu(t-s)}$.

We now can verify either by a direct substitution into the equations $\pi_j = \sum_i \pi_i P_{ij}$, or by the same argument as presented in the remark at the end of Section 8.7, that the limiting probabilities of this Markov chain are of the form

$$\pi_{k-1+j} = c\beta^j, \qquad j = 0, 1, \ldots.$$

Substitution into any of the equations $\pi_j = \sum_i \pi_i P_{ij}$ when $j > k$ yields that β is given as the solution of

$$\beta = \int_0^\infty e^{-k\mu t(1-\beta)}\, dG(t)$$

The values $\pi_0, \pi_1, \ldots, \pi_{k-2}$ can be obtained by recursively solving the first $k-1$ of the steady-state equations, and c can then be computed by using $\sum_0^\infty \pi_i = 1$.

If we let W_Q^* denote the amount of time that a customer spends in queue, then in exactly the same manner as in $G/M/1$ we can show that

$$W_Q^* = \begin{cases} 0, & \text{with probability } \sum_0^{k-1} \pi_i = 1 - \frac{c\beta}{1-\beta} \\ \text{Exp}(k\mu(1-\beta)), & \text{with probability } \sum_k^{\infty} \pi_i = \frac{c\beta}{1-\beta} \end{cases}$$

where $\text{Exp}(k\mu(1-\beta))$ is an exponential random variable with rate $k\mu(1-\beta)$.

8.9.4 The M/G/k Queue

In this section we consider the $M/G/k$ system in which customers arrive at a Poisson rate λ and are served by any of k servers, each of whom has the service distribution G. If we attempt to mimic the analysis presented in Section 8.5 for the $M/G/1$ system, then we would start with the basic identity

$$V = \lambda E[S] W_Q + \lambda E[S^2]/2 \tag{8.61}$$

and then attempt to derive a second equation relating V and W_Q.

Now if we consider an arbitrary arrival, then we have the following identity:

$$\begin{aligned} &\text{work in system when customer arrives} \\ &\quad = k \times \text{time customer spends in queue} + R \end{aligned} \tag{8.62}$$

where R is the sum of the remaining service times of all other customers in service at the moment when our arrival enters service.

The foregoing follows because while the arrival is waiting in queue, work is being processed at a rate k per unit time (since all servers are busy). Thus, an amount of work $k \times$ time in queue is processed while he waits in queue. Now, all of this work was present when he arrived and in addition the remaining work on those still being served when he enters service was also present when he arrived—so we obtain Equation (8.62). For an illustration, suppose that there are three servers all of whom are busy when the customer arrives. Suppose, in addition, that there are no other customers in the system and also that the remaining service times of the three people in service are 3, 6, and 7. Hence, the work seen by the arrival is $3 + 6 + 7 = 16$. Now the arrival will spend 3 time units in queue, and at the moment he enters service, the remaining times of the other two customers are $6-3 = 3$ and $7-3 = 4$. Hence, $R = 3 + 4 = 7$ and as a check of Equation (8.62) we see that $16 = 3 \times 3 + 7$.

Taking expectations of Equation (8.62) and using the fact that Poisson arrivals see time averages, we obtain

$$V = kW_Q + E[R]$$

which, along with Equation (8.61), would enable us to solve for W_Q if we could compute $E[R]$. However there is no known method for computing $E[R]$ and in fact, there is no known exact formula for W_Q. The following approximation for W_Q was obtained in Reference 6 by using the foregoing approach and then approximating $E[R]$:

$$W_Q \approx \frac{\lambda^k E[S^2](E[S])^{k-1}}{2(k-1)!(k-\lambda E[S])^2 \left[\sum_{n=0}^{k-1} \frac{(\lambda E[S])^n}{n!} + \frac{(\lambda E[S])^k}{(k-1)!(k-\lambda E[S])}\right]} \tag{8.63}$$

The preceding approximation has been shown to be quite close to W_Q when the service distribution is gamma. It is also exact when G is exponential.

Exercises

1. For the $M/M/1$ queue, compute
   (a) the expected number of arrivals during a service period and
   (b) the probability that no customers arrive during a service period.

   Hint: "Condition."

*2. Machines in a factory break down at an exponential rate of six per hour. There is a single repairman who fixes machines at an exponential rate of eight per hour. The cost incurred in lost production when machines are out of service is $10 per hour per machine. What is the average cost rate incurred due to failed machines?

3. The manager of a market can hire either Mary or Alice. Mary, who gives service at an exponential rate of 20 customers per hour, can be hired at a rate of $3 per hour. Alice, who gives service at an exponential rate of 30 customers per hour, can be hired at a rate of $C per hour. The manager estimates that, on the average, each customer's time is worth $1 per hour and should be accounted for in the model. Assume customers arrive at a Poisson rate of 10 per hour
   (a) What is the average cost per hour if Mary is hired? If Alice is hired?
   (b) Find C if the average cost per hour is the same for Mary and Alice.

4. Suppose that a customer of the $M/M/1$ system spends the amount of time $x > 0$ waiting in queue before entering service.
   (a) Show that, conditional on the preceding, the number of other customers that were in the system when the customer arrived is distributed as $1 + P$, where P is a Poisson random variable with mean λ.
   (b) Let W_Q^* denote the amount of time that an $M/M/1$ customer spends in queue. As a by-product of your analysis in part (a), show that

$$P\{W_Q^* \leqslant x\} = \begin{cases} 1 - \frac{\lambda}{\mu} & \text{if } x = 0 \\ 1 - \frac{\lambda}{\mu} + \frac{\lambda}{\mu}(1 - e^{-(\mu-\lambda)x}) & \text{if } x > 0 \end{cases}$$

5. It follows from Exercise 4 that if, in the $M/M/1$ model, W_Q^* is the amount of time that a customer spends waiting in queue, then

$$W_Q^* = \begin{cases} 0, & \text{with probability } 1 - \lambda/\mu \\ \text{Exp}(\mu - \lambda), & \text{with probability } \lambda/\mu \end{cases}$$

where $\text{Exp}(\mu - \lambda)$ is an exponential random variable with rate $\mu - \lambda$. Using this, find $\text{Var}(W_Q^*)$.

6. Suppose we want to find the covariance between the times spent in the system by the first two customers in an $M/M/1$ queueing system. To obtain this covariance, let S_i be the service time of customer i, $i = 1, 2$, and let Y be the time between the two arrivals.
   (a) Argue that $(S_1 - Y)^+ + S_2$ is the amount of time that customer 2 spends in the system, where $x^+ = \max(x, 0)$.
   (b) Find $\text{Cov}(S_1, (S_1 - Y)^+ + S_2)$.

   Hint: Compute both $E[(S - Y)^+]$ and $E[S_1(S_1 - Y)^+]$ by conditioning on whether $S_1 > Y$.

*7. Show that W is smaller in an $M/M/1$ model having arrivals at rate λ and service at rate 2μ than it is in a two-server $M/M/2$ model with arrivals at rate λ and with each server at rate μ. Can you give an intuitive explanation for this result? Would it also be true for W_Q?

8. A facility produces items according to a Poisson process with rate λ. However, it has shelf space for only k items and so it shuts down production whenever k items are present. Customers arrive at the facility according to a Poisson process with rate μ. Each customer wants one item and will immediately depart either with the item or empty handed if there is no item available.
   (a) Find the proportion of customers that go away empty handed.
   (b) Find the average time that an item is on the shelf.
   (c) Find the average number of items on the shelf.

9. A group of n customers moves around among two servers. Upon completion of service, the served customer then joins the queue (or enters service if the server is free) at the other server. All service times are exponential with rate μ. Find the proportion of time that there are j customers at server 1, $j = 0, \ldots, n$.

10. A group of m customers frequents a single-server station in the following manner. When a customer arrives, he or she either enters service if the server is free or joins the queue otherwise. Upon completing service the customer departs the system, but then returns after an exponential time with rate θ. All service times are exponentially distributed with rate μ.
   (a) Find the average rate at which customers enter the station.
   (b) Find the average time that a customer spends in the station per visit.

11. Consider a single-server queue with Poisson arrivals and exponential service times having the following variation: Whenever a service is completed a departure occurs only with probability α. With probability $1 - \alpha$ the customer, instead of leaving, joins the end of the queue. Note that a customer may be serviced more than once.
   (a) Set up the balance equations and solve for the steady-state probabilities, stating conditions for it to exist.
   (b) Find the expected waiting time of a customer from the time he arrives until he enters service for the first time.
   (c) What is the probability that a customer enters service exactly n times, $n = 1, 2, \ldots$?
   (d) What is the expected amount of time that a customer spends in service (which does not include the time he spends waiting in line)?

Hint: Use part (c).

(e) What is the distribution of the total length of time a customer spends being served?

Hint: Is it memoryless?

*12. A supermarket has two exponential checkout counters, each operating at rate μ. Arrivals are Poisson at rate λ. The counters operate in the following way:

(i) One queue feeds both counters.

(ii) One counter is operated by a permanent checker and the other by a stock clerk who instantaneously begins checking whenever there are two or more customers in the system. The clerk returns to stocking whenever he completes a service, and there are fewer than two customers in the system.

(a) Find P_n, proportion of time there are n in the system.

(b) At what rate does the number in the system go from 0 to 1? From 2 to 1?

(c) What proportion of time is the stock clerk checking?

Hint: Be a little careful when there is one in the system.

13. Two customers move about among three servers. Upon completion of service at server i, the customer leaves that server and enters service at whichever of the other two servers is free. (Therefore, there are always two busy servers.) If the service times at server i are exponential with rate μ_i, $i = 1, 2, 3$, what proportion of time is server i idle?

14. Consider a queueing system having two servers and no queue. There are two types of customers. Type 1 customers arrive according to a Poisson process having rate λ_1, and will enter the system if either server is free. The service time of a type 1 customer is exponential with rate μ_1. Type 2 customers arrive according to a Poisson process having rate λ_2. A type 2 customer requires the simultaneous use of both servers; hence, a type 2 arrival will only enter the system if both servers are free. The time that it takes (the two servers) to serve a type 2 customer is exponential with rate μ_2. Once a service is completed on a customer, that customer departs the system.

(a) Define states to analyze the preceding model.

(b) Give the balance equations.

In terms of the solution of the balance equations, find

(c) the average amount of time an entering customer spends in the system;

(d) the fraction of served customers that are type 1.

15. Consider a sequential-service system consisting of two servers, A and B. Arriving customers will enter this system only if server A is free. If a customer does enter, then he is immediately served by server A. When his service by A is completed, he then goes to B if B is free, or if B is busy, he leaves the system. Upon completion of service at server B, the customer departs. Assume that the (Poisson) arrival rate is two customers an hour, and that A and B serve at respective (exponential) rates of four and two customers an hour.

(a) What proportion of customers enter the system?

(b) What proportion of entering customers receive service from B?

(c) What is the average number of customers in the system?

(d) What is the average amount of time that an entering customer spends in the system?

16. Customers arrive at a two-server system according to a Poisson process having rate $\lambda = 5$. An arrival finding server 1 free will begin service with that server. An arrival finding server 1 busy and server 2 free will enter service with server 2. An arrival finding both servers busy goes away. Once a customer is served by either server, he departs the system. The service times at server i are exponential with rates μ_i, where $\mu_1 = 4$, $\mu_2 = 2$.
    (a) What is the average time an entering customer spends in the system?
    (b) What proportion of time is server 2 busy?

17. Customers arrive at a two-server station in accordance with a Poisson process with a rate of two per hour. Arrivals finding server 1 free begin service with that server. Arrivals finding server 1 busy and server 2 free begin service with server 2. Arrivals finding both servers busy are lost. When a customer is served by server 1, she then either enters service with server 2 if 2 is free or departs the system if 2 is busy. A customer completing service at server 2 departs the system. The service times at server 1 and server 2 are exponential random variables with respective rates of four and six per hour.
    (a) What fraction of customers do not enter the system?
    (b) What is the average amount of time that an entering customer spends in the system?
    (c) What fraction of entering customers receives service from server 1?

18. Arrivals to a three-server system are according to a Poisson process with rate λ. Arrivals finding server 1 free enter service with 1. Arrivals finding 1 busy but 2 free enter service with 2. Arrivals finding both 1 and 2 busy do not join the system. After completion of service at either 1 or 2 the customer will then either go to server 3 if 3 is free or depart the system if 3 is busy. After service at 3 customers depart the system. The service times at i are exponential with rate μ_i, $i = 1, 2, 3$.
    (a) Define states to analyze the above system.
    (b) Give the balance equations.
    (c) In terms of the solution of the balance equations, what is the average time that an entering customer spends in the system?
    (d) Find the probability that a customer who arrives when the system is empty is served by server 3.

19. The economy alternates between good and bad periods. During good times customers arrive at a certain single-server queueing system in accordance with a Poisson process with rate λ_1, and during bad times they arrive in accordance with a Poisson process with rate λ_2. A good time period lasts for an exponentially distributed time with rate α_1, and a bad time period lasts for an exponential time with rate α_2. An arriving customer will only enter the queueing system if the server is free; an arrival finding the server busy goes away. All service times are exponential with rate μ.
    (a) Define states so as to be able to analyze this system.
    (b) Give a set of linear equations whose solution will yield the long-run proportion of time the system is in each state.

    In terms of the solutions of the equations in part (b),
    (c) what proportion of time is the system empty?
    (d) what is the average rate at which customers enter the system?

20. There are two types of customers. Type 1 and 2 customers arrive in accordance with independent Poisson processes with respective rate λ_1 and λ_2. There are two servers. A type 1 arrival will enter service with server 1 if that server is free; if server 1 is busy and server 2 is free, then the type 1 arrival will enter service with server 2. If both servers are busy, then the type 1 arrival will go away. A type 2 customer can only be served by server 2; if server 2 is free when a type 2 customer arrives, then the customer enters service with that server. If server 2 is busy when a type 2 arrives, then that customer goes away. Once a customer is served by either server, he departs the system. Service times at server i are exponential with rate μ_i, $i = 1, 2$.

    Suppose we want to find the average number of customers in the system.

    (a) Define states.
    (b) Give the balance equations. Do not attempt to solve them.

    In terms of the long-run probabilities, what is

    (c) the average number of customers in the system?
    (d) the average time a customer spends in the system?

*21. Suppose in Exercise 20 we want to find out the proportion of time there is a type 1 customer with server 2. In terms of the long-run probabilities given in Exercise 20, what is

    (a) the rate at which a type 1 customer enters service with server 2?
    (b) the rate at which a type 2 customer enters service with server 2?
    (c) the fraction of server 2's customers that are type 1?
    (d) the proportion of time that a type 1 customer is with server 2?

22. Customers arrive at a single-server station in accordance with a Poisson process with rate λ. All arrivals that find the server free immediately enter service. All service times are exponentially distributed with rate μ. An arrival that finds the server busy will leave the system and roam around "in orbit" for an exponential time with rate θ at which time it will then return. If the server is busy when an orbiting customer returns, then that customer returns to orbit for another exponential time with rate θ before returning again. An arrival that finds the server busy and N other customers in orbit will depart and not return. That is, N is the maximum number of customers in orbit.

    (a) Define states.
    (b) Give the balance equations.

    In terms of the solution of the balance equations, find

    (c) the proportion of all customers that are eventually served;
    (d) the average time that a served customer spends waiting in orbit.

23. Consider the $M/M/1$ system in which customers arrive at rate λ and the server serves at rate μ. However, suppose that in any interval of length h in which the server is busy there is a probability $\alpha h + o(h)$ that the server will experience a breakdown, which causes the system to shut down. All customers that are in the system depart, and no additional arrivals are allowed to enter until the breakdown is fixed. The time to fix a breakdown is exponentially distributed with rate β.

    (a) Define appropriate states.
    (b) Give the balance equations.

    In terms of the long-run probabilities,

    (c) what is the average amount of time that an entering customer spends in the system?

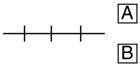

Figure 8.4

(d) what proportion of entering customers complete their service?
(e) what proportion of customers arrive during a breakdown?

*24. Reconsider Exercise 23, but this time suppose that a customer that is in the system when a breakdown occurs remains there while the server is being fixed. In addition, suppose that new arrivals during a breakdown period are allowed to enter the system. What is the average time a customer spends in the system?

25. Poisson (λ) arrivals join a queue in front of two parallel servers A and B, having exponential service rates μ_A and μ_B (see Figure 8.4). When the system is empty, arrivals go into server A with probability α and into B with probability $1-\alpha$. Otherwise, the head of the queue takes the first free server.
(a) Define states and set up the balance equations. Do not solve.
(b) In terms of the probabilities in part (a), what is the average number in the system? Average number of servers idle?
(c) In terms of the probabilities in part (a), what is the probability that an arbitrary arrival will get serviced in A?

26. In a queue with unlimited waiting space, arrivals are Poisson (parameter λ) and service times are exponentially distributed (parameter μ). However, the server waits until K people are present before beginning service on the first customer; thereafter, he services one at a time until all K units, and all subsequent arrivals, are serviced. The server is then “idle” until K new arrivals have occurred.
(a) Define an appropriate state space, draw the transition diagram, and set up the balance equations.
(b) In terms of the limiting probabilities, what is the average time a customer spends in queue?
(c) What conditions on λ and μ are necessary?

27. Consider a single-server exponential system in which ordinary customers arrive at a rate λ and have service rate μ. In addition, there is a special customer who has a service rate μ_1. Whenever this special customer arrives, she goes directly into service (if anyone else is in service, then this person is bumped back into queue). When the special customer is not being serviced, she spends an exponential amount of time (with mean $1/\theta$) out of the system.
(a) What is the average arrival rate of the special customer?
(b) Define an appropriate state space and set up balance equations.
(c) Find the probability that an ordinary customer is bumped n times.

*28. Let D denote the time between successive departures in a stationary $M/M/1$ queue with $\lambda < \mu$. Show, by conditioning on whether or not a departure has left the system empty, that D is exponential with rate λ.

Hint: By conditioning on whether or not the departure has left the system empty we see that

$$D = \begin{cases} \text{Exponential}(\mu), & \text{with probability } \lambda/\mu \\ \text{Exponential}(\lambda) * \text{Exponential}(\mu), & \text{with probability } 1-\lambda/\mu \end{cases}$$

where Exponential(λ) $*$ Exponential(μ) represents the sum of two independent exponential random variables having rates μ and λ. Now use moment-generating functions to show that D has the required distribution.

Note that the preceding does not prove that the departure process is Poisson. To prove this we need show not only that the interdeparture times are all exponential with rate λ, but also that they are independent.

29. Potential customers arrive to a single-server hair salon according to a Poisson process with rate λ. A potential customer who finds the server free enters the system; a potential customer who finds the server busy goes away. Each potential customer is type i with probability p_i, where $p_1 + p_2 + p_3 = 1$. Type 1 customers have their hair washed by the server; type 2 customers have their hair cut by the server; and type 3 customers have their hair first washed and then cut by the server. The time that it takes the server to wash hair is exponentially distributed with rate μ_1, and the time that it takes the server to cut hair is exponentially distributed with rate μ_2.
   (a) Explain how this system can be analyzed with four states.
   (b) Give the equations whose solution yields the proportion of time the system is in each state.

   In terms of the solution of the equations of (b), find
   (c) the proportion of time the server is cutting hair;
   (d) the average arrival rate of entering customers.

30. For the tandem queue model verify that

$$P_{n,m} = (\lambda/\mu_1)^n(1 - \lambda/\mu_1)(\lambda/\mu_2)^m(1 - \lambda/\mu_2)$$

   satisfies the balance equation (8.15).

31. Consider a network of three stations with a single server at each station. Customers arrive at stations 1, 2, 3 in accordance with Poisson processes having respective rates 5, 10, and 15. The service times at the three stations are exponential with respective rates 10, 50, and 100. A customer completing service at station 1 is equally likely to (i) go to station 2, (ii) go to station 3, or (iii) leave the system. A customer departing service at station 2 always goes to station 3. A departure from service at station 3 is equally likely to either go to station 2 or leave the system.
   (a) What is the average number of customers in the system (consisting of all three stations)?
   (b) What is the average time a customer spends in the system?

32. Consider a closed queueing network consisting of two customers moving among two servers, and suppose that after each service completion the customer is equally likely to go to either server—that is, $P_{1,2} = P_{2,1} = \frac{1}{2}$. Let μ_i denote the exponential service rate at server $i, i = 1, 2$.
   (a) Determine the average number of customers at each server.
   (b) Determine the service completion rate for each server.

33. Explain how a Markov chain Monte Carlo simulation using the Gibbs sampler can be utilized to estimate
   (a) the distribution of the amount of time spent at server j on a visit.

   Hint: Use the arrival theorem.

(b) the proportion of time a customer is with server j (i.e., either in server j's queue or in service with j).

34. For open queueing networks
    (a) state and prove the equivalent of the arrival theorem;
    (b) derive an expression for the average amount of time a customer spends waiting in queues.

35. Customers arrive at a single-server station in accordance with a Poisson process having rate λ. Each customer has a value. The successive values of customers are independent and come from a uniform distribution on $(0, 1)$. The service time of a customer having value x is a random variable with mean $3 + 4x$ and variance 5.
    (a) What is the average time a customer spends in the system?
    (b) What is the average time a customer having value x spends in the system?

*36. Compare the $M/G/1$ system for first-come, first-served queue discipline with one of last-come, first-served (for instance, in which units for service are taken from the top of a stack). Would you think that the queue size, waiting time, and busy-period distribution differ? What about their means? What if the queue discipline was always to choose at random among those waiting? Intuitively, which discipline would result in the smallest variance in the waiting time distribution?

37. In an $M/G/1$ queue,
    (a) what proportion of departures leave behind 0 work?
    (b) what is the average work in the system as seen by a departure?

38. For the $M/G/1$ queue, let X_n denote the number in the system left behind by the nth departure.
    (a) If

$$X_{n+1} = \begin{cases} X_n - 1 + Y_n, & \text{if } X_n \geqslant 1 \\ Y_n, & \text{if } X_n = 0 \end{cases}$$

    what does Y_n represent?

    (b) Rewrite the preceding as

$$X_{n+1} = X_n - 1 + Y_n + \delta_n \tag{8.64}$$

    where

$$\delta_n = \begin{cases} 1, & \text{if } X_n = 0 \\ 0, & \text{if } X_n \geqslant 1 \end{cases}$$

    Take expectations and let $n \to \infty$ in Equation (8.64) to obtain

$$E[\delta_\infty] = 1 - \lambda E[S]$$

    (c) Square both sides of Equation (8.64), take expectations, and then let $n \to \infty$ to obtain

$$E[X_\infty] = \frac{\lambda^2 E[S^2]}{2(1 - \lambda E[S])} + \lambda E[S]$$

    (d) Argue that $E[X_\infty]$, the average number as seen by a departure, is equal to L.

*39. Consider an $M/G/1$ system in which the first customer in a busy period has the service distribution G_1 and all others have distribution G_2. Let C denote the number of customers in a busy period, and let S denote the service time of a customer chosen at random.

Argue that

(a) $a_0 = P_0 = 1 - \lambda E[S]$.

(b) $E[S] = a_0 E[S_1] + (1 - a_0)E[S_2]$ where S_i has distribution G_i.

(c) Use (a) and (b) to show that $E[B]$, the expected length of a busy period, is given by

$$E[B] = \frac{E[S_1]}{1 - \lambda E[S_2]}$$

(d) Find $E[C]$.

40. Consider a $M/G/1$ system with $\lambda E[S] < 1$.

(a) Suppose that service is about to begin at a moment when there are n customers in the system.

(i) Argue that the additional time until there are only $n - 1$ customers in the system has the same distribution as a busy period.

(ii) What is the expected additional time until the system is empty?

(b) Suppose that the work in the system at some moment is A. We are interested in the expected additional time until the system is empty—call it $E[T]$. Let N denote the number of arrivals during the first A units of time.

(i) Compute $E[T|N]$.

(ii) Compute $E[T]$.

41. Carloads of customers arrive at a single-server station in accordance with a Poisson process with rate 4 per hour. The service times are exponentially distributed with rate 20 per hour. If each carload contains either 1, 2, or 3 customers with respective probabilities $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$, compute the average customer delay in queue.

42. In the two-class priority queueing model of Section 8.6.2, what is W_Q? Show that W_Q is less than it would be under FIFO if $E[S_1] < E[S_2]$ and greater than under FIFO if $E[S_1] > E[S_2]$.

43. In a two-class priority queueing model suppose that a cost of C_i per unit time is incurred for each type i customer that waits in queue, $i = 1, 2$. Show that type 1 customers should be given priority over type 2 (as opposed to the reverse) if

$$\frac{E[S_1]}{C_1} < \frac{E[S_2]}{C_2}$$

44. Consider the priority queueing model of Section 8.6.2 but now suppose that if a type 2 customer is being served when a type 1 arrives then the type 2 customer is bumped out of service. This is called the preemptive case. Suppose that when a bumped type 2 customer goes back in service his service begins at the point where it left off when he was bumped.

(a) Argue that the work in the system at any time is the same as in the nonpreemptive case.

(b) Derive W_Q^1.

Hint: How do type 2 customers affect type 1s?

(c) Why is it not true that

$$V_Q^2 = \lambda_2 E[S_2] W_Q^2$$

(d) Argue that the work seen by a type 2 arrival is the same as in the nonpreemptive case, and so

$$W_Q^2 = W_Q^2(\text{nonpreemptive}) + E[\text{extra time}]$$

where the extra time is due to the fact that he may be bumped.

(e) Let N denote the number of times a type 2 customer is bumped. Why is

$$E[\text{extra time}|N] = \frac{NE[S_1]}{1 - \lambda_1 E[S_1]}$$

Hint: When a type 2 is bumped, relate the time until he gets back in service to a "busy period."

(f) Let S_2 denote the service time of a type 2. What is $E[N|S_2]$?

(g) Combine the preceding to obtain

$$W_Q^2 = W_Q^2(\text{nonpreemptive}) + \frac{\lambda_1 E[S_1] E[S_2]}{1 - \lambda_1 E[S_1]}$$

*45. Calculate explicitly (not in terms of limiting probabilities) the average time a customer spends in the system in Exercise 24.

46. In the $G/M/1$ model if G is exponential with rate λ show that $\beta = \lambda/\mu$.

47. Verify Erlang's loss formula, Equation (8.60), when $k = 1$.

48. Verify the formula given for the P_i of the $M/M/k$.

49. In the Erlang loss system suppose the Poisson arrival rate is $\lambda = 2$, and suppose there are three servers, each of whom has a service distribution that is uniformly distributed over $(0, 2)$. What proportion of potential customers is lost?

50. In the $M/M/k$ system,

(a) what is the probability that a customer will have to wait in queue?

(b) determine L and W.

51. Verify the formula for the distribution of W_Q^* given for the $G/M/k$ model.

*52. Consider a system where the interarrival times have an arbitrary distribution F, and there is a single server whose service distribution is G. Let D_n denote the amount of time the nth customer spends waiting in queue. Interpret S_n, T_n so that

$$D_{n+1} = \begin{cases} D_n + S_n - T_n, & \text{if } D_n + S_n - T_n \geqslant 0 \\ 0, & \text{if } D_n + S_n - T_n < 0 \end{cases}$$

53. Consider a model in which the interarrival times have an arbitrary distribution F, and there are k servers each having service distribution G. What condition on F and G do you think would be necessary for there to exist limiting probabilities?

References

[1] J. Cohen, "The Single Server Queue," North-Holland, Amsterdam, 1969.

[2] R. B. Cooper, "Introduction to Queueing Theory," Second Edition, Macmillan, New York, 1984.

[3] D. R. Cox and W. L. Smith, "Queues," Wiley, New York, 1961.

[4] F. Kelly, "Reversibility and Stochastic Networks," Wiley, New York, 1979.

[5] L. Kleinrock, "Queueing Systems," Vol. I, Wiley, New York, 1975.

[6] S. Nozaki and S. Ross, "Approximations in Finite Capacity Multiserver Queues with Poisson Arrivals," *J. Appl. Prob.* **13**, 826–834 (1978).

[7] L. Takacs, "Introduction to the Theory of Queues," Oxford University Press, London and New York, 1962.

[8] H. Tijms, "Stochastic Models: An Algorithmic Approach," Wiley, New York, 1994.

[9] P. Whittle, "Systems in Stochastic Equilibrium," Wiley, New York, 1986.

[10] Wolff, "Stochastic Modeling and the Theory of Queues," Prentice Hall, New Jersey, 1989.