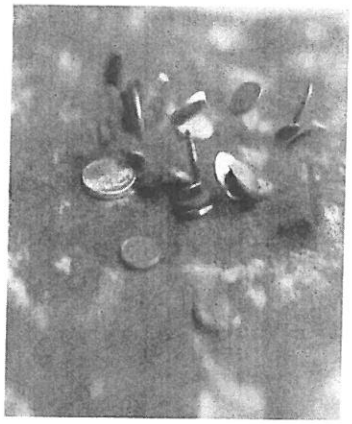
Ross: "Intro to Probability Models"

8

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 which are of great use in analyzing queueing models. We also introduce three different sets of limiting probabilities which 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 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 multiservers 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 which assumes Poisson arrivals but allows for a general service distribution.

8.2. Preliminaries

In this section we will derive certain identities which 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:

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

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 (and 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

*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.

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.

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.*

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

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

$$a_n = \text{proportion of customers that find } n \text{ in the system when they arrive, and}$$

$$d_n = \text{proportion of customers leaving behind } n \text{ in the system when they depart}$$

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.

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 one at a time and are served one at a time

$$a_n = d_n, \qquad n \geq 0$$

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. The result now follows since the overall arrival rate must equal the overall departure rate (what goes in eventually goes out.) ■

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 the 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) $P_n T$. 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) $\lambda P_n T$. 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.

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 time 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 which will enable us to determine the state probabilities. Namely, for each $n \geq 0$, *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 $(\lambda + \mu)P_1$.

*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.

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 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)

The set of Equations (8.5) which balances the rate at which the process enters each state with the rate at which it leaves that state is known as *balance equations*.

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

$$P_1 = \frac{\lambda}{\mu} P_0,$$

$$P_{n+1} = \frac{\lambda}{\mu} P_n + \left(P_n - \frac{\lambda}{\mu} P_{n-1} \right), \qquad n \geqslant 1$$

Solving in terms of P_0 yields

$$P_0 = P_0,$$

$$P_1 = \frac{\lambda}{\mu} P_0,$$

$$P_2 = \frac{\lambda}{\mu} P_1 + \left(P_1 - \frac{\lambda}{\mu} P_0 \right) = \frac{\lambda}{\mu} P_1 = \left(\frac{\lambda}{\mu} \right)^2 P_0,$$

$$P_3 = \frac{\lambda}{\mu} P_2 + \left(P_2 - \frac{\lambda}{\mu} P_1 \right) = \frac{\lambda}{\mu} P_2 = \left(\frac{\lambda}{\mu} \right)^3 P_0,$$

$$P_4 = \frac{\lambda}{\mu} P_3 + \left(P_3 - \frac{\lambda}{\mu} P_2 \right) = \frac{\lambda}{\mu} P_3 = \left(\frac{\lambda}{\mu} \right)^4 P_0,$$

$$P_{n+1} = \frac{\lambda}{\mu} P_n + \left(P_n - \frac{\lambda}{\mu} P_{n-1} \right) = \frac{\lambda}{\mu} P_n = \left(\frac{\lambda}{\mu} \right)^{n+1} P_0$$

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

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

$$P_n = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right), \qquad n \geq 1 \tag{8.6}$$

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.

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

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

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 two, 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 (8.7), 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 (8.7) as

$$L = \frac{\lambda/\mu}{1 - \lambda/\mu},$$
$$W = \frac{1/\mu}{1 - \lambda/\mu}$$

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

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

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. ■

Let W^* denote the amount of time an arbitrary customer spends in the system. To obtain the distribution of W^*, we condition on the number in the system when

the customer arrives. This yields

$$P\{W^* \leqslant a\} = \sum_{n=0}^{\infty} P\{W^* \leqslant a \mid n \text{ in the system when he arrives}\} \times P\{n \text{ in the system when he arrives}\} \tag{8.9}$$

Now consider the amount of time that our customer must spend in the system if there are already n customers present when he arrives. If $n = 0$, then his time in the system will just be his service time. When $n \geqslant 1$, there will be one customer in service and $n - 1$ waiting in line ahead of our arrival. The customer in service might have been in service for some time, but due to the lack of memory of the exponential distribution (see Section 5.2), it follows that our arrival would have to wait an exponential amount of time with rate μ for this customer to complete service. As he also would have to wait an exponential amount of time for each of the other $n - 1$ customers in line, it follows, upon adding his own service time, that the amount of time that a customer must spend in the system if there are already n customers present when he arrives is the sum of $n + 1$ independent and identically distributed exponential random variables with rate μ. But it is known (see Section 5.2.3) that such a random variable has a gamma distribution with parameters $(n + 1, \mu)$. That is,

$$P\{W^* \leqslant a \mid n \text{ in the system when he arrives}\} = \int_0^a \mu e^{-\mu t} \frac{(\mu t)^n}{n!}\, dt$$

Because

$$P\{n \text{ in the system when he arrives}\} = P_n \qquad \text{(since Poisson arrivals)} = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right)$$

we have from Equation (8.9) and the preceding that

$$\begin{aligned} P\{W^* \leqslant a\} &= \sum_{n=0}^{\infty} \int_0^a \mu e^{-\mu t} \frac{(\mu t)^n}{n!}\, dt \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right) \\ &= \int_0^a (\mu - \lambda) e^{-\mu t} \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!}\, dt \qquad \text{(by interchanging)} \\ &= \int_0^a (\mu - \lambda) e^{-\mu t} e^{\lambda t}\, dt \end{aligned}$$

$$= \int_0^a (\mu - \lambda) e^{-(\mu-\lambda)t}\, dt$$
$$= 1 - e^{-(\mu-\lambda)a}$$

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,

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

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 $\mu - \lambda$.

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, $0 \leq n \leq N$, 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	$\lambda P_0 = \mu P_1$
$1 \leq n \leq N - 1$	$(\lambda + \mu) P_n = \lambda P_{n-1} + \mu P_{n+1}$
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 states 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.

To solve, we again rewrite the preceding system of equations:

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

which, solving in terms of P_0, yields

$$\begin{aligned} P_1 &= \frac{\lambda}{\mu} P_0, \\ P_2 &= \frac{\lambda}{\mu} P_1 + \left(P_1 - \frac{\lambda}{\mu} P_0 \right) = \frac{\lambda}{\mu} P_1 = \left(\frac{\lambda}{\mu} \right)^2 P_0, \\ P_3 &= \frac{\lambda}{\mu} P_2 + \left(P_2 - \frac{\lambda}{\mu} P_1 \right) = \frac{\lambda}{\mu} P_2 = \left(\frac{\lambda}{\mu} \right)^3 P_0, \\ &\vdots \\ P_{N-1} &= \frac{\lambda}{\mu} P_{N-2} + \left(P_{N-2} - \frac{\lambda}{\mu} P_{N-3} \right) = \left(\frac{\lambda}{\mu} \right)^{N-1} P_0, \\ P_N &= \frac{\lambda}{\mu} P_{N-1} = \left(\frac{\lambda}{\mu} \right)^N P_0 \end{aligned} \tag{8.10}$$

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 Equation (8.10) we obtain

$$P_n = \frac{(\lambda/\mu)^n(1 - \lambda/\mu)}{1 - (\lambda/\mu)^{N+1}}, \qquad n = 0, 1, \ldots, N \tag{8.11}$$

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

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

$$= \frac{(1 - \lambda/\mu)}{1 - (\lambda/\mu)^{N+1}} \sum_{n=0}^{N} n\left(\frac{\lambda}{\mu}\right)^n$$

which after some algebra yields

$$L = \frac{\lambda[1 + N(\lambda/\mu)^{N+1} - (N + 1)(\lambda/\mu)^N]}{(\mu - \lambda)(1 - (\lambda/\mu)^{N+1})} \tag{8.12}$$

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 8.3 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. ■

In the previous two models, it has been quite easy to define the state of the system. Namely, it was defined as the number of people in the system. Now we shall consider some examples where a more detailed state space is necessary.

8.3.3. A Shoeshine Shop

Consider a shoeshine 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 $(1, 0)$ which 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.

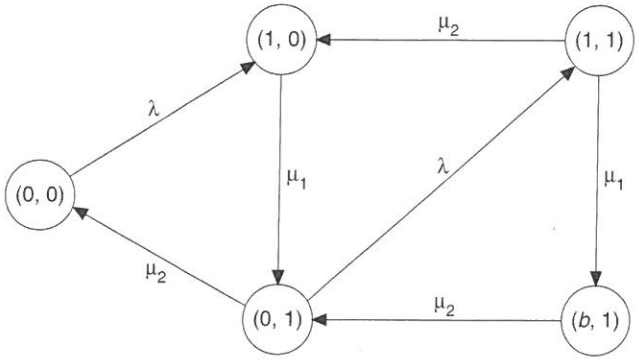


Figure 8.1. A transition diagram.

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 and 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.

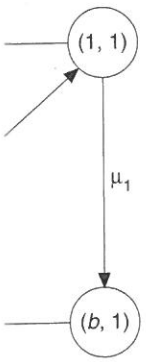
To write the balance equations we equate the sum of the arrows (multiplied by the probability of the states where they originate) coming into a state with the sum of the arrows (multiplied by the probability of the state) going out of that state. This gives

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 in state (0, 0) or (0, 1), it follows that $\lambda_a = \lambda(P_{00} + P_{01})$ and hence

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

Example 8.4 (a) If $\lambda = 1$, $\mu_1 = 1$, $\mu_2 = 2$, then calculate the preceding quantities of interest.

(b) If $\lambda = 1$, $\mu_1 = 2$, $\mu_2 = 1$, then calculate the preceding.

Solution: (a) Solving the balance equations yields

$$P_{00} = \tfrac{12}{37}, \quad P_{10} = \tfrac{16}{37}, \quad P_{11} = \tfrac{2}{37}, \quad P_{01} = \tfrac{6}{37}, \quad P_{b1} = \tfrac{1}{37}$$

Hence,

$$L = \tfrac{28}{37}, \quad W = \tfrac{28}{18}$$

(b) Solving the balance equations yields

$$P_{00} = \tfrac{3}{11}, \quad P_{10} = \tfrac{2}{11}, \quad P_{11} = \tfrac{1}{11}, \quad P_{b1} = \tfrac{2}{11}, \quad P_{01} = \tfrac{3}{11}$$

Hence,

$$L = 1, \quad W = \tfrac{11}{6} \quad \blacksquare$$

8.3.4. 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