

Chapter 6

Queueing Models

6.1 Queueing Systems

Queues (or waiting lines) are an unavoidable component of modern life. We are required to stand physically in queues in grocery stores, banks, department stores, amusement parks, movie theaters, etc. Although we don't like standing in a queue, we appreciate the fairness that it imposes. Even when we use phones to conduct business, often we are put on hold and served in a first-come–first-served fashion. Thus we face a queue even if we are in our own home!

Queues are not just for humans, however. Modern communication systems transmit messages (like emails) from one computer to another by queueing them up inside the network in a complicated fashion. Modern manufacturing systems maintain queues (called inventories) of raw materials, partly finished goods, and finished goods throughout the manufacturing process. “Supply chain management” is nothing but the management of these queues!

In this chapter, we shall study some simple models of queues. Typically, a queueing system consists of a stream of customers (humans, finished goods, messages) that arrive at a service facility, get served according to a given service discipline, and then depart. The service facility may have one or more servers and finite or infinite capacity. In practice, we are interested in designing a queueing system, namely its capacity, number of servers, service discipline, etc. These models will help us do this by answering the following questions (and many others):

- (1) How many customers are there in the queue on average?
- (2) How long does a typical customer spend in the queue?
- (3) How many customers are rejected or lost due to capacity limitations?
- (4) How busy are the servers?

We start by introducing standard nomenclature for single-station queues; i.e., queueing systems where customers form a single queue. Such a queue is described as follows:

1. Arrival Process

We assume that customers arrive one at a time and that the successive inter-arrival times are iid. (Thus the arrival process is a renewal process, studied

in Chapter 5.) It is described by the distribution of the inter-arrival times, represented by special symbols as follows:

- M : exponential;
- G : general;
- D : deterministic;
- E_k : Erlang with k phases, etc.

Note that the Poisson arrival process is represented by M (for memoryless; i.e., exponential inter-arrival times).

2. Service Times

We assume that the service times of successive customers are iid. They are represented by the same letters as the inter-arrival times.

3. Number of Servers

Typically denoted by s . All the servers are assumed to be identical, and it is assumed that any customer can be served by any server.

4. Capacity

Typically denoted by K . It includes the customers in service. If an arriving customer finds K customers in the system, he or she is permanently lost. If capacity is not mentioned, it is assumed to be infinite.

Example 6.1. (Nomenclature). In an $M/M/1$ queue, customers arrive according to a Poisson process, request iid exponential service times, and are served by a single server. The capacity is infinite. In an $M/G/2/10$ queue, the customers arrive according to a Poisson process and demand iid service times with a general distribution. The service facility has two servers and the capacity to hold ten customers. ■

In this chapter, we shall always assume a first-come–first-served (FCFS) service discipline.

6.2 Single-Station Queues: General Results

In this section, we state several general results for single-station queueing systems. Let A_n be the time of the n th arrival. We allow for the possibility that not all arrivals may actually enter the system. Hence we define E_n to be the time of the n th entry. (If all arriving customers enter the system, we have $A_n = E_n$ for all $n = 0, 1, 2, \dots$.) Let D_n be the time of the n th departure. If the service discipline is FCFS, then the n th entering customer enters at time E_n and leaves at time D_n . Hence the time spent by the n th customer in the system is given by $W_n = D_n - E_n$. (See Figure 6.1.)

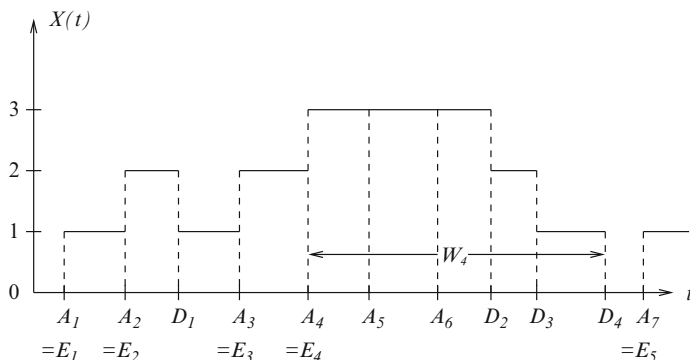


Fig. 6.1 A typical sample path of the queue-length process.

Let $N(t)$ be the number of customers that arrived by time t , let $D(t)$ be the number of customers that departed by time t , and let $X(t)$ be the number of customers in the system at time t . Furthermore, let

$$X_n = X(D_n+),$$

$$X_n^* = X(E_n-),$$

$$\hat{X}_n = X(A_n-).$$

Thus, X_n is the number of customers left behind by the n th departure, X_n^* is the number of customers in the system as seen by the n th entry (not including the entering customer) into the system, and $\hat{X}_n$ is the number of customers in the system as seen by the n th arrival (not including the arriving customer) to the system. Now (for $j = 0, 1, 2, \dots$) define the following limits, assuming they exist:

$$p_j = \lim_{t \rightarrow \infty} \mathbf{P}(X(t) = j), \quad (6.1)$$

$$\pi_j = \lim_{n \rightarrow \infty} \mathbf{P}(X_n = j), \quad (6.2)$$

$$\pi_j^* = \lim_{n \rightarrow \infty} \mathbf{P}(X_n^* = j), \quad (6.3)$$

$$\hat{\pi}_j = \lim_{n \rightarrow \infty} \mathbf{P}(\hat{X}_n = j). \quad (6.4)$$

We can interpret p_j as the long-run fraction of the time that the system has j customers in it, π_j as the long-run fraction of departures that leave behind j customers in the system, π_j^* as the long-run fraction of entering customers that see j customers ahead of them in the system, and $\hat{\pi}_j$ as the long-run fraction of arrivals that see j customers ahead of them in the system. Since all these quantities relate to a common queueing system, we expect them to be related to each other. The next two theorems state these relations.

Theorem 6.1. (When is $\pi_j = \pi_j^*$?). *Suppose the arrivals and departures take place one at a time. If the limits in (6.2) or (6.3) exist,*

$$\pi_j = \pi_j^*, \quad j \geq 0.$$

Idea of Proof. The theorem is a deterministic result with no probabilistic assumptions. Suppose $X(0) = i$. Then, using a careful counting argument (using the fact that the arrivals and departures take place one at a time), we can show that

$$\{X_{n+i} \leq j\} \Leftrightarrow \{X_{n+j+1}^* \leq j\}$$

for all $n \geq 0$ and all $j \geq 0$. (See Kulkarni [2010].) Hence

$$P(X_{n+i} \leq j) = P(X_{n+j+1}^* \leq j).$$

Assuming the limit exists on either the left or the right side, we get

$$\lim_{n \rightarrow \infty} P(X_{n+i} \leq j) = \lim_{n \rightarrow \infty} P(X_{n+j+1}^* \leq j),$$

which yields

$$\lim_{n \rightarrow \infty} P(X_n \leq j) = \lim_{n \rightarrow \infty} P(X_n^* \leq j).$$

However, this implies

$$\sum_{i=0}^j \pi_i = \sum_{i=0}^j \pi_i^*, \quad j \geq 0,$$

and hence

$$\pi_j = \pi_j^*, \quad j \geq 0.$$

Hence the theorem follows. ■

Theorem 6.2. (Poisson Arrivals See Time Averages: PASTA). *Suppose the arrival process is Poisson and $\{N(t+s), s \geq 0\}$ is independent of $\{X(u), 0 \leq u \leq t\}$. If the limits in (6.1) or (6.4) exist,*

$$p_j = \hat{\pi}_j, \quad j \geq 0.$$

Intuition. The technical condition about independence says that the arrival process from any point t onward is independent of the history of the system up to time t . Now, assuming this condition and using the properties of the Poisson processes given in Section 3.3, we see that the probability that exactly one arrival occurs in the interval $[t, t+h]$ is independent of the history up to time t . If h is sufficiently small, we can say that the arrival occurs at time t . By assumption, $X(t)$ is independent of what happens to the arrival process over $(t, t+h)$. Hence the conditional distribution of $X(t)$, given that there is an arrival in the interval $(t, t+h)$, is the same as the distribution of $X(t)$. Thus the distribution of the state as seen by an arrival at time t

is the same as the distribution of $X(t)$. Thus the long-run fraction of the arrivals that see the system in state j (namely $\hat{\pi}_j$) must be the same as the fraction of the time the system spends in state j (namely p_j). This result is sometimes called PASTA (Poisson Arrivals See Time Averages). ■

We illustrate the theorems above by means of an example.

Example 6.2. (M/M/1/1 Queue). Consider an $M/M/1/1$ queue with arrival rate λ and service rate μ . In such a queue, the arrival process is Poisson with rate λ and service times are iid exponential with parameter μ . However, since the capacity is 1, an arriving customer can enter the system only if the server is idle; otherwise the arrival is lost. The $\{X(t), t \geq 0\}$ process for such a queue is modeled as a CTMC on state space $\{0, 1\}$ in Example 4.7 with $K = 1$. The steady-state distribution can be computed from the results of Example 4.25 as

$$p_0 = \frac{\mu}{\lambda + \mu}, \quad p_1 = \frac{\lambda}{\lambda + \mu}.$$

Since the arrival process is Poisson, it follows from Theorem 6.2 that

$$\hat{\pi}_0 = p_0, \quad \hat{\pi}_1 = p_1.$$

Thus the long-run probability that a customer is turned away is given by $\hat{\pi}_1 = \lambda/(\lambda + \mu)$. Since a departing customer always leaves an empty system, we have $X_n = 0$ for all $n \geq 0$. Hence

$$\pi_0 = 1, \quad \pi_1 = 0.$$

Similarly, since a customer can enter only if the system is empty, we have $X_n^* = 0$ for all $n \geq 0$. Hence,

$$\pi_0^* = 1, \quad \pi_1^* = 0.$$

Thus Theorem 6.1 is verified. Note that the last two observations are valid for any $G/G/1/1$ queue. ■

Now define

- L = the expected number of customers in the system in steady state;
- λ = the arrival rate of customers;
- W = the expected time spent in the system by a customer in steady state.

The relation between these quantities is the celebrated Little's Law and is stated in the following theorem.

Theorem 6.3. (Little's Law). *If the quantities L , λ , and W defined above exist and are finite, they satisfy*

$$L = \lambda W.$$

Intuition. Suppose each arriving customer pays the system \$1 per unit time that the customer spends in the system. First we compute the long-run rate, R_{customer} , at which the money is spent by the customers. Customers arrive at the rate of λ per unit time. Each customer spends on average W time in the system and hence pays $\$W$ to the system. Hence we must have

$$R_{\text{customer}} = \lambda W.$$

Next we compute the long-run rate, R_{system} , at which the system earns money. Since each customer pays the system \$1 per unit time and there are L customers in the system on average, we must have

$$R_{\text{system}} = L.$$

Now we expect that R_{customer} will equal R_{system} since no money is created or destroyed in the transaction. Hence we get

$$L = \lambda W,$$

which is Little's Law. ■

Little's Law is also a sample path result; i.e., it does not make any probabilistic assumptions about the system. Care has to be taken in applying it in a consistent manner, as described in the following example.

Example 6.3. (Little's Law and $M/M/1/1$ Queue). Consider the $M/M/1/1$ queue of Example 6.2. The expected number of customers in the system in steady state can be computed as

$$L = 0 \cdot p_0 + 1 \cdot p_1 = \frac{\lambda}{\lambda + \mu}. \quad (6.5)$$

Little's Law can be applied to the stream of arriving customers or entering customers. We discuss these two cases below.

Arriving Customers. With probability p_0 , the arriving customer finds the server idle, in which case he or she spends $1/\mu$ amount of time in the system; with probability p_1 , the arriving customer finds the server busy, in which case he spends 0 amount of time in the system (since he leaves instantaneously). Hence the expected time spent by a customer in the system is

$$W = \frac{1}{\mu} \cdot p_0 + 0 \cdot p_1 = \frac{1}{\lambda + \mu}. \quad (6.6)$$

The customers arrive according to a $PP(\lambda)$; hence the arrival rate is λ . Hence, from (6.5) and (6.6), we see that $L = \lambda W$; i.e., Little's Law holds.

Entering Customers. An entering customer always spends $1/\mu$ amount of time in the system on average. Hence

$$W_e = \frac{1}{\mu}. \quad (6.7)$$

In order to verify Little's Law, we need to compute λ_e , the rate at which customers enter the system. Since the successive inter-entry times are iid with mean $\tau = 1/\mu + 1/\lambda = (\lambda + \mu)/\lambda\mu$ (mean service time followed by mean idle time), we see that the rate at which the customers enter the system per unit time is given by (see Theorem 5.2) $\lambda_e = 1/\tau = \lambda\mu/(\lambda + \mu)$. Hence, we have

$$L = \frac{\lambda}{\lambda + \mu} = \frac{\lambda\mu}{\lambda + \mu} \cdot \frac{1}{\mu} = \lambda_e W_e.$$

Thus Little's Law is verified. ■

Example 6.4. (Variation of Little's Law). We can use the intuition behind Little's Law to create many variations by changing the payment rules for the customers. For example, suppose each customer pays \$1 per unit time that he spends in the queue (but not in service). Then the expected amount paid by a customer is W_q , the expected time spent in the queue (not including service). Also, the rate at which the system collects revenue is L_q , the expected number of customers in the queue (not including those in service). Hence the same intuitive argument that yielded Little's Law yields

$$L_q = \lambda W_q,$$

where λ is the arrival rate of customers. ■

Example 6.5. (Another Variation of Little's Law). Now consider the following payment scheme for the customers. Each customer pays \$1 per unit time that he spends in service. Then the expected amount paid by a customer is τ , the expected time spent in service. Also, the rate at which the system collects revenue is B , the expected number of customers in service, which is the same as the expected number of busy servers. (In a single-server queue, this is also the fraction of the time that the server is busy.) Hence the same intuitive argument that yielded Little's Law yields

$$B = \lambda \tau,$$

where λ is the arrival rate of customers. However, B cannot exceed s , the number of servers at the station. Hence we must have

$$B = \min(s, \lambda \tau). \quad \blacksquare$$

The example above creates an interesting question: what happens when $\lambda \tau > s$? We discuss this question in Theorem 6.4 below. When there is unlimited waiting room space (i.e., when the capacity is infinite), the state space of the $\{X(t), t \geq 0\}$ process is $\{0, 1, 2, \dots\}$. So far in this book we have always considered finite state spaces. New issues arise in the study of stochastic processes with infinite state space that we did not encounter in the finite state space stochastic processes. In particular, it is possible that $X(t)$ may tend to infinity as t goes to infinity. In such a case, customers may end up waiting an infinite amount of time in the system in the long

run. If this occurs, we call the process unstable. For example, if the arrival rate of customers at a queue is larger than the server's ability to serve, we would expect the queue length to build up without bounds, thus making the queue-length process unstable. We formally define the concept below.

Definition 6.1. (Stability). A single-station queue with unlimited capacity is called *stable* if

$$\sum_{j=0}^{\infty} p_j = 1,$$

where p_j is as defined in (6.1). Otherwise it is called *unstable*.

In an unstable queue, we have

$$\sum_{j=0}^{\infty} p_j < 1.$$

This can be interpreted to mean that, in the long run, the queue length goes to infinity with a positive probability given by $1 - \sum_{j=0}^{\infty} p_j$. Obviously, having infinite queue lengths is bad for the customers. Hence it makes sense to call such a system unstable. The next theorem gives a condition for stability.

Theorem 6.4. (Condition of Stability). *Consider a single-station queue with s servers and infinite capacity. Suppose the customers enter at rate λ and the mean service time is τ . The queue is stable if*

$$\lambda \tau < s. \quad (6.8)$$

Intuition. From Example 6.5, we get

$$B = \text{expected number of busy servers} = \lambda \tau, \quad (6.9)$$

where λ is the arrival rate of entering customers. However, the number of busy servers cannot exceed s , the total number of servers. Hence, for the argument to be valid, we must have

$$\lambda \tau \leq s.$$

What goes wrong when $\lambda \tau > s$? It was implicitly assumed in the derivation of (6.9) that each entering customer eventually enters service. Intuitively, when $\lambda \tau > s$, some of the customers never get to enter service; i.e., they wait in the system indefinitely. This implies that the number in the system must be infinity; i.e., the system must be unstable. It is more difficult to prove that the system can be unstable even if $\lambda \tau = s$. Essentially, this is due to the variance of either the inter-arrival times or the service times. Thus we can safely say that the queue is stable if $\lambda \tau < s$. ■

6.3 Birth and Death Queues with Finite Capacity

Let $X_K(t)$ be the number of customers at time t in a single-station queue with finite capacity K . In this section, we consider the case where $\{X_K(t), t \geq 0\}$ is a birth and death process on state space $\{0, 1, 2, \dots, K\}$ with birth parameters $\{\lambda_i, i = 0, 1, 2, \dots, K-1\}$ and death parameters $\{\mu_i, i = 1, 2, 3, \dots, K\}$. We have studied such processes in Example 4.8 and their limiting behavior in Example 4.23. We restate the result about the limiting probabilities as

$$p_i(K) = \lim_{t \rightarrow \infty} P(X_K(t) = i), \quad 0 \leq i \leq K.$$

Let $\rho_0 = 1$ and

$$\rho_i = \frac{\lambda_0 \lambda_1 \cdots \lambda_{i-1}}{\mu_1 \mu_2 \cdots \mu_i}, \quad 1 \leq i \leq K. \quad (6.10)$$

Then, from Example 4.23, we get

$$p_i(K) = \frac{\rho_i}{\sum_{j=0}^K \rho_j}, \quad 0 \leq i \leq K. \quad (6.11)$$

We consider several special cases below.

6.3.1 M/M/1/K Queue

Consider a service station where customers arrive according to a $PP(\lambda)$. The service times are iid random variables with $\text{Exp}(\mu)$ distribution. (Equivalently, we say the service rate is μ .) They are served in an FCFS fashion by a single server. The system has the capacity to hold K customers. Thus an incoming customer who sees K customers in the system ahead of him is permanently lost.

We have already seen this system in Example 4.7 as a model of a bank queue. Let $X_K(t)$ be the number of customers in the system at time t . Then we know (see the discussion following Example 4.8) that $\{X_K(t), t \geq 0\}$ is a birth and death process on state space $\{0, 1, \dots, K\}$ with birth parameters

$$\lambda_i = \lambda, \quad i = 0, 1, \dots, K-1,$$

and death parameters

$$\mu_i = \mu, \quad i = 1, 2, \dots, K.$$

Substituting in (6.10), we get

$$\rho_i = \left(\frac{\lambda}{\mu}\right)^i, \quad 0 \leq i \leq K. \quad (6.12)$$

Using

$$\rho = \frac{\lambda}{\mu},$$

we get

$$\sum_{i=0}^K \rho_i = \begin{cases} \frac{1 - \rho^{K+1}}{1 - \rho} & \text{if } \rho \neq 1, \\ K + 1 & \text{if } \rho = 1. \end{cases}$$

Substituting in (6.11), we get

$$p_i(K) = \begin{cases} \frac{1 - \rho}{1 - \rho^{K+1}} \rho^i & \text{if } \rho \neq 1, \\ \frac{1}{K + 1} & \text{if } \rho = 1, \end{cases} \quad 0 \leq i \leq K. \quad (6.13)$$

We compute several relevant quantities with the help of this limiting distribution.

Server Idleness. The long-run fraction of the time that the server is idle is given by

$$p_0(K) = \frac{1 - \rho}{1 - \rho^{K+1}}.$$

Blocking Probability. The long-run fraction of the time that the system is full is given by

$$p_K(K) = \frac{1 - \rho}{1 - \rho^{K+1}} \rho^K. \quad (6.14)$$

Since the arrival process is Poisson, we can use Theorem 6.2 to get the system distribution as seen by an arriving (not necessarily entering) customer as

$$\hat{\pi}_i(K) = p_i(K), \quad i = 0, 1, 2, \dots, K.$$

Thus an arriving customer sees the system full with probability $p_K(K)$. Hence, $p_K(K)$ also represents the long-run blocking probability and long-run fraction of the arriving customers that are blocked (or lost).

Entering Customers. Now we compute $\pi_i^*(K)$, the probability that an entering customer sees i customers in the system ahead of him. We have $\pi_K^*(K) = 0$ since a customer cannot enter if the system is full. For $0 \leq i \leq K - 1$, we have

$$\begin{aligned} \pi_i^*(K) &= \text{P}(\text{entering customer sees } i \text{ in the system}) \\ &= \text{P}(\text{arriving customer sees } i \text{ in the system} | \text{arriving customer enters}) \\ &= \frac{\text{P}(\text{arriving customer sees } i \text{ in the system and arriving customer enters})}{\text{P}(\text{arriving customer enters})} \end{aligned}$$

$$\begin{aligned}
&= \frac{\text{P(arriving customer sees } i \text{ in the system)}}{\text{P(arriving customer enters)}} \\
&= \frac{p_i(K)}{1 - p_K(K)}, \quad 0 \leq i \leq K-1,
\end{aligned}$$

where $p_i(K)$ are as given in (6.13).

Expected Number in the System. The expected number in the system in steady state is computed as follows. If $\rho = 1$, we have

$$L = \sum_{i=0}^K i p_i(K) = \frac{1}{K+1} \sum_{i=0}^K i = \frac{K}{2}.$$

If $\rho \neq 1$, more tedious calculations yield

$$\begin{aligned}
L &= \sum_{i=0}^K i p_i(K) \\
&= \frac{1-\rho}{1-\rho^{K+1}} \sum_{i=0}^K i \rho^i \\
&= \frac{\rho}{1-\rho} \cdot \frac{1-(K+1)\rho^K + K\rho^{K+1}}{1-\rho^{K+1}}. \tag{6.15}
\end{aligned}$$

Note that L increases from 0 at $\rho = 0$ to K as $\rho \rightarrow \infty$. This stands to reason since as $\rho \rightarrow \infty$ the arrival rate is increasing relative to the service rate. Hence, eventually, when the arrival rate becomes very large, the system will always be full, thus leading to $L = K$.

Expected Waiting Time: Arrivals. We can compute the expected time spent by an arrival in the system by using Little's Law as stated in Theorem 6.3. Here the arrival rate is λ and the mean number in the system is L as given in (6.15). Hence W , the expected time spent in the system by an arriving customer, is given by

$$W = L/\lambda.$$

Note that an arriving customer spends no time in the system if he finds the system full.

Expected Waiting Time: Entries. Next we compute the expected time spent by an entering customer in the system by using Little's Law as stated in Theorem 6.3. Here the arrival rate of entering customers is

$$\lambda \text{P(arriving customer enters)} = \lambda(1 - p_K(K)).$$

The mean number in the system is L as given in (6.15). Hence W , the expected time spent in the system by an entering customer, is given by

$$W = L/(\lambda(1 - p_K(K))). \quad (6.16)$$

Example 6.6. (ATM Queue). Consider the model of the queue in front of an ATM as described in Example 4.7. The data given in Example 4.20 imply that this is an $M/M/1/5$ queue with an arrival rate of $\lambda = 10$ per hour and a service rate of $\mu = 15$ per hour. Thus

$$\rho = \lambda/\mu = 2/3.$$

1. What is the probability that an incoming customer finds the system full?

Using (6.14), we get the desired answer

$$p_5(5) = .0481.$$

2. What is the expected amount of time an entering customer spends in the system?

Using (6.16), we get the desired answer

$$W = .1494 \text{ hr} = 8.97 \text{ min.} \blacksquare$$

6.3.2 $M/M/s/K$ Queue

Let $X_K(t)$ be the number of customers in an $M/M/s/K$ queue at time t . (We shall assume that $K \geq s$ since it does not make sense to have more servers than there is room for customers.) We saw this queue in Example 4.10 as a model of a call center. We saw there that $\{X_K(t), t \geq 0\}$ is a birth and death process on $\{0, 1, \dots, K\}$ with birth parameters

$$\lambda_i = \lambda, \quad 0 \leq i \leq K-1,$$

and death parameters

$$\mu_i = \min(i, s)\mu, \quad 0 \leq i \leq K.$$

The next theorem gives the limiting distribution of an $M/M/s/K$ queue in terms of the dimensionless quantity defined below,

$$\rho = \frac{\lambda}{s\mu}. \quad (6.17)$$

Theorem 6.5. (Limiting Distribution of an $M/M/s/K$ Queue). *The limiting distribution of an $M/M/s/K$ queue is given by*

$$p_i(K) = p_0(K)\rho^i, \quad i = 0, 1, 2, \dots, K,$$

where

$$\rho_i = \begin{cases} \frac{1}{i!} \left(\frac{\lambda}{\mu} \right)^i, & \text{if } 0 \leq i \leq s-1, \\ \frac{s^s}{s!} \rho^i, & \text{if } s \leq i \leq K, \end{cases}$$

and

$$p_0(K) = \left[\sum_{i=0}^{s-1} \frac{1}{i!} \left(\frac{\lambda}{\mu} \right)^i + \frac{1}{s!} \left(\frac{\lambda}{\mu} \right)^s \cdot \frac{1 - \rho^{K-s+1}}{1 - \rho} \right]^{-1}.$$

Proof. The expressions for ρ_i follow from (6.10), and that for $p_i(K)$ follows from (6.11). Furthermore, we have

$$\begin{aligned} \sum_{i=0}^K \rho_i &= \sum_{i=0}^{s-1} \frac{1}{i!} \left(\frac{\lambda}{\mu} \right)^i + \sum_{i=s}^K \frac{s^s}{s!} \rho^i \\ &= \sum_{i=0}^{s-1} \frac{1}{i!} \left(\frac{\lambda}{\mu} \right)^i + \frac{s^s}{s!} \rho^s \sum_{i=s}^K \rho^{i-s} \\ &= \sum_{i=0}^{s-1} \frac{1}{i!} \left(\frac{\lambda}{\mu} \right)^i + \frac{s^s}{s!} \rho^s \sum_{i=0}^{K-s} \rho^i \\ &= \sum_{i=0}^{s-1} \frac{1}{i!} \left(\frac{\lambda}{\mu} \right)^i + \frac{s^s}{s!} \rho^s \frac{1 - \rho^{K-s+1}}{1 - \rho}. \end{aligned}$$

Substituting in (6.11), we get the theorem. ■

Using the limiting distribution above, we can study the same quantities as in the $M/M/1/K$ case.

Example 6.7. (Call Center). Consider the call center described in Example 4.10. Using the data in Computational Problem 4.9, we get an $M/M/6/10$ queue with an arrival rate of $\lambda = 60$ per hour and a service rate of $\mu = 10$ per hour per server.

1. Compute the limiting distribution of the number of calls in the system in steady state.

From Theorem 6.5, we get

$$p(K) = [.0020 .0119 .0357 .0715 .1072 .1286 .1286 .1286 .1286 .1286 .1286 .1286].$$

2. Compute the expected number of calls on hold.

This is the same as the number of customers in the system not being served. It is given by

$$\sum_{i=7}^{10} (i-6) p_i(K) = 1.2862.$$

3. What fraction of the calls are lost?

The desired quantity is given by

$$p_{10}(10) = .1286.$$

Thus 12.86% of the incoming calls are lost. ■

6.3.3 *M/M/K/K Queue*

An $M/M/s/K$ queue with $s = K$ produces the $M/M/K/K$ queue as a special case. Such a system is called a *loss system* since any customer that cannot enter service immediately upon arrival is lost. We have seen such a queue as a model of a telephone switch in Example 4.9. By setting $s = K$ in Theorem 6.5, we get the limiting distribution of an $M/M/K/K$ queue as given in the following theorem.

Theorem 6.6. (Limiting Distribution of an $M/M/K/K$ Queue). *Let $X(t)$ be the number of customers in an $M/M/K/K$ queue at time t and*

$$p_i(K) = \lim_{t \rightarrow \infty} P(X(t) = i), \quad 0 \leq i \leq K.$$

Then

$$p_i(K) = \frac{\rho_i}{\sum_{j=0}^K \rho_j}, \quad 0 \leq i \leq K,$$

where

$$\rho_i = \frac{(\lambda/\mu)^i}{i!}, \quad 0 \leq i \leq K.$$

It is possible to interpret the limiting distribution in the theorem above as the distribution of a truncated (at K) Poisson random variable. To be precise, let Y be a Poisson random variable with parameter λ/μ . Then, for $0 \leq i \leq K$,

$$\begin{aligned} P(Y = i | Y \leq K) &= \frac{P(Y = i, Y \leq K)}{P(Y \leq K)} \\ &= \frac{P(Y = i)}{P(Y \leq K)} \end{aligned}$$

$$\begin{aligned}
&= \frac{e^{-\lambda/\mu} (\lambda/\mu)^i / i!}{\sum_{j=0}^K e^{-\lambda/\mu} (\lambda/\mu)^j / j!} \\
&= \frac{(\lambda/\mu)^i / i!}{\sum_{j=0}^K (\lambda/\mu)^j / j!} = p_i(K).
\end{aligned}$$

The quantity $p_K(K)$ is an important quantity, called the blocking probability. It is the fraction of the time that the system is full, or the fraction of the customers that are lost, in the long run.

Example 6.8. (Telephone Switch). Recall the model of a telephone switch developed in Example 4.9. Using the data in Example 4.31, we see that the number of calls handled by the switch is an $M/M/6/6$ queue with an arrival rate of $\lambda = 4$ per minute and a service rate of $\mu = .5$ per minute per server.

1. What is the expected number of calls in the switch in steady state?

Using Theorem 6.6, we compute the limiting distribution as

$$p(K) = [.0011 \ .0086 \ .0343 \ .0913 \ .1827 \ .2923 \ .3898].$$

Hence the expected number of calls is given by

$$\sum_{i=0}^6 i p_i(K) = 4.8820.$$

2. How large should the switch capacity be if we desire to lose no more than 10% of the incoming calls?

The fraction of calls lost is given by $p_K(K)$. At $K = 6$, this quantity is .3898. By computing $p_K(K)$ for larger values of K , we get

$$\begin{aligned}
p_7(7) &= .3082, \quad p_8(8) = .2356, \quad p_9(9) = .1731, \\
p_{10}(10) &= .1217, \quad p_{11}(11) = .0813.
\end{aligned}$$

Hence the telephone switch needs at least 11 lines to lose less than 10% of the traffic. ■

Remark. One can show that Theorem 6.6 remains valid for the $M/G/K/K$ queue where the service times are iid generally distributed random variables with mean $1/\mu$. The proof of this remarkable result is beyond the scope of this book, and we refer the reader to any advanced textbook on queueing theory.

6.4 Birth and Death Queues with Infinite Capacity

Let $X(t)$ be the number of customers in a single-station queue with infinite capacity. Here we consider the cases where $\{X(t), t \geq 0\}$ is a birth and death process on state space $\{0, 1, 2, \dots\}$ with birth parameters $\{\lambda_i, i \geq 0\}$ and death parameters $\{\mu_i, i \geq 1\}$. We are interested in the condition under which such a queue is stable. If the queue is stable, we are interested in its limiting distribution

$$p_i = \lim_{t \rightarrow \infty} \mathbf{P}(X(t) = i), \quad i \geq 0.$$

We have seen a birth and death queue with finite state space $\{0, 1, 2, \dots, K\}$ in the previous section. We treat the infinite-state birth and death queue as the limiting case of it as $K \rightarrow \infty$.

Define $\rho_0 = 1$ and

$$\rho_i = \frac{\lambda_0 \lambda_1 \cdots \lambda_{i-1}}{\mu_1 \mu_2 \cdots \mu_i}, \quad i \geq 1. \quad (6.18)$$

Theorem 6.7. (Infinite Birth and Death Process). *A birth and death queue with infinite state space is stable if and only if*

$$\sum_{j=0}^{\infty} \rho_j < \infty. \quad (6.19)$$

When the queue is stable, its limiting distribution is given by

$$p_i = \frac{\rho_i}{\sum_{j=0}^{\infty} \rho_j}, \quad i \geq 0. \quad (6.20)$$

Proof. Let $p_i(K)$ be as in (6.11). We have

$$\begin{aligned} p_i &= \lim_{K \rightarrow \infty} p_i(K) \\ &= \begin{cases} \frac{\rho_i}{\sum_{j=0}^{\infty} \rho_j} & \text{if } \sum_{j=0}^{\infty} \rho_j < \infty, \\ 0 & \text{if } \sum_{j=0}^{\infty} \rho_j = \infty. \end{cases} \end{aligned}$$

Thus we have

$$\begin{aligned} \sum_{i=0}^{\infty} p_i &= \begin{cases} \sum_{i=0}^{\infty} \rho_i / \sum_{j=0}^{\infty} \rho_j & \text{if } \sum_{j=0}^{\infty} \rho_j < \infty, \\ 0 & \text{if } \sum_{j=0}^{\infty} \rho_j = \infty, \end{cases} \\ &= \begin{cases} 1 & \text{if } \sum_{j=0}^{\infty} \rho_j < \infty, \\ 0 & \text{if } \sum_{j=0}^{\infty} \rho_j = \infty. \end{cases} \end{aligned}$$

Hence it follows from Definition 6.1 that the queue is stable if and only if (6.19) holds and that the limiting distribution of a stable birth and death queue is given by (6.20). ■

We study several special cases of the birth and death queues with infinite state spaces below.

6.4.1 *M/M/1 Queue*

In an $M/M/1$ queueing system, the customers arrive according to a $PP(\lambda)$ and the service times are iid $\text{Exp}(\mu)$ random variables. There is a single server and an unlimited waiting room capacity. The next theorem gives the main result for the $M/M/1$ queue. Define

$$\rho = \frac{\lambda}{\mu}. \quad (6.21)$$

The quantity ρ is called the *traffic intensity* of the $M/M/1$ queue.

Theorem 6.8. (Limiting Distribution of an $M/M/1$ Queue). *An $M/M/1$ queue with traffic intensity ρ is stable if and only if*

$$\rho < 1. \quad (6.22)$$

If the queue is stable, its limiting distribution is given by

$$p_i = (1 - \rho)\rho^i, \quad i \geq 0. \quad (6.23)$$

Proof. Let $X(t)$ be the number of customers in the system at time t . $\{X(t), t \geq 0\}$ is a birth and death process on $\{0, 1, 2, \dots\}$ with birth parameters

$$\lambda_i = \lambda, \quad i \geq 0,$$

and death parameters

$$\mu_i = \mu, \quad i \geq 1.$$

Substituting in (6.18), we get

$$\rho_i = \left(\frac{\lambda}{\mu}\right)^i = \rho^i, \quad i \geq 0.$$

Now

$$\sum_{i=0}^{\infty} \rho_i = \sum_{i=0}^{\infty} \rho^i = \begin{cases} \frac{1}{1-\rho} & \text{if } \rho < 1, \\ \infty & \text{if } \rho \geq 1. \end{cases}$$

Thus, from Theorem 6.7, we see that the condition of stability is as given in (6.22). Substituting in (6.20), we get the limiting distribution in (6.23). ■

Note that the condition of stability in (6.22) is consistent with the general condition in (6.8) (using $\tau = 1/\mu$). Using the definition of traffic intensity ρ , the condition of stability can also be written as

$$\lambda < \mu.$$

In this form, it makes intuitive sense: if the arrival rate (λ) is less than the service rate (μ), the system is able to handle the load and the queue does not build up to infinity. Hence it is stable. If not, the system is unable to handle the load and the queue builds to infinity! Hence it is unstable. It is curious that even in the case $\lambda = \mu$ (or, equivalently, $\rho = 1$) the queue is unstable. This is due to the effect of randomness. Even if the average service capacity of the system is perfectly matched with the average load on the system, it has no leftover capacity to absorb the random buildups that will invariably occur in a stochastic service system.

We shall compute several quantities using the limiting distribution given in Theorem 6.8.

Server Idleness. The probability that the server is idle is given by

$$p_0 = 1 - \rho, \quad (6.24)$$

and the probability that the server is busy is given by

$$1 - p_0 = \rho.$$

Hence the expected number of busy servers is ρ . This is consistent with the general result in Example 6.5.

Embedded Distributions. The queue length in an $M/M/1$ queue changes by ± 1 at a time. The arrival process is Poisson, and all arriving customers enter the system. Hence Theorems 6.1 and 6.2 imply that

$$\hat{\pi}_i = \pi_i^* = \pi_i = p_i = (1 - \rho)\rho^i, \quad i \geq 0.$$

Expected Number in the System. The expected number in the system is given by

$$\begin{aligned} L &= \sum_{i=0}^{\infty} ip_i \\ &= \sum_{i=0}^{\infty} i(1 - \rho)\rho^i \\ &= \frac{\rho}{1 - \rho}. \end{aligned} \quad (6.25)$$

The last equation is a result of tedious calculations.

Expected Waiting Time. Using Little's Law, we get

$$W = \frac{L}{\lambda} = \frac{1}{\mu} \frac{1}{1 - \rho} = \frac{1}{\mu - \lambda}. \quad (6.26)$$

Note that, as expected, both L and W tend to infinity as $\rho \rightarrow 1$.

Example 6.9. (Taxi Stand). Customers arrive at a taxi stand according to a Poisson process at a rate of 15 per hour and form a queue for the taxis. Taxis arrive at the stand according to a Poisson process at a rate of 18 per hour. Each taxi carries away one passenger. If no customers are waiting when a taxi arrives, it leaves right away without picking up any customers. Let $X(t)$ be the number of customers waiting at the taxi stand at time t . Due to the memoryless property of the exponential distribution, we see that the customer at the head of the line has to wait for an $\exp(18)$ amount of time for the next taxi. Thus we can think of $\{X(t), t \geq 0\}$ as an $M/M/1$ queue with an arrival rate of $\lambda = 15$ per hour and a service rate of $\mu = 18$ per hour.

1. Is this queue stable?

The traffic intensity is

$$\rho = \lambda/\mu = 15/18 = 5/6.$$

Since this is less than 1, the queue is stable, from Theorem 6.8.

2. On average, how long does a customer wait before getting into a taxi?

This is given by the average waiting time

$$W = 1/(\mu - \lambda) = .3333 \text{ hours.}$$

Thus the average wait for a taxi is 20 minutes. ■

6.4.2 $M/M/s$ Queue

Next we study an $M/M/s$ queueing system. Customers arrive according to a $PP(\lambda)$. The service times are iid $\text{Exp}(\mu)$ random variables. There are s identical servers and there is unlimited waiting room capacity. The customers form a single line and get served by the next available server on an FCFS basis. The next theorem gives the main result for the $M/M/s$ queue.

Theorem 6.9. (Limiting Distribution of an $M/M/s$ Queue). *An $M/M/s$ queue with an arrival rate of λ and s identical servers each with service rate μ is stable if and only if*

$$\rho = \frac{\lambda}{s\mu} < 1. \quad (6.27)$$

If the queue is stable, its limiting distribution is given by

$$p_i = p_0 \rho_i, \quad i \geq 0,$$

where

$$\rho_i = \begin{cases} \frac{1}{i!} \left(\frac{\lambda}{\mu} \right)^i & \text{if } 0 \leq i \leq s-1, \\ \frac{s^s}{s!} \rho^i & \text{if } i \geq s, \end{cases} \quad (6.28)$$

and

$$p_0 = \left[\sum_{i=0}^{s-1} \frac{1}{i!} \left(\frac{\lambda}{\mu} \right)^i + \frac{s^s}{s!} \cdot \frac{\rho^s}{1-\rho} \right]^{-1}.$$

Proof. Let $X(t)$ be the number of customers in the $M/M/s$ queue at time t . $\{X(t), t \geq 0\}$ is a birth and death process on $\{0, 1, 2, \dots\}$ with birth parameters

$$\lambda_i = \lambda, \quad i \geq 0,$$

and death parameters

$$\mu_i = \min(s, i)\mu, \quad i \geq 1.$$

Using

$$\rho = \frac{\lambda}{s\mu} \quad (6.29)$$

in (6.18), we get (6.28). Now

$$\begin{aligned} \sum_{i=0}^{\infty} \rho_i &= \sum_{i=0}^{s-1} \rho_i + \sum_{i=s}^{\infty} \rho_i \\ &= \sum_{i=0}^{s-1} \rho_i + \frac{s^s}{s!} \frac{\rho^s}{1-\rho} \quad \text{if } \rho < 1. \end{aligned}$$

If $\rho \geq 1$, the sum above diverges. Thus, from Theorem 6.7, we see that the condition of stability is as given in (6.27). The limiting distribution then follows from (6.20). ■

Note that the condition of stability in (6.27) can also be written as

$$\lambda < s\mu.$$

In this form, it makes intuitive sense: if the arrival rate λ is less than the maximum service rate $s\mu$ (which occurs when all servers are busy), the system is able to handle the load and the queue does not build up to infinity. Hence it is stable. If not, the

system is unable to handle the load and the queue builds to infinity! Hence it is unstable. As in the $M/M/1$ queue, the queue is unstable when $\lambda = s\mu$.

We shall compute several quantities using the limiting distribution given in Theorem 6.9.

Embedded Distributions. The queue length in an $M/M/s$ queue changes by ± 1 at a time. The arrival process is Poisson, and all arriving customers enter the system. Hence Theorems 6.1 and 6.2 imply that

$$\hat{\pi}_i = \pi_i^* = \pi_i = p_i.$$

Probability of Waiting. An incoming customer has to wait for service if all the servers are busy. This probability is given by

$$\begin{aligned} \sum_{i=s}^{\infty} p_i &= \sum_{i=s}^{\infty} p_0 \frac{s^s}{s!} \rho^i \\ &= p_0 \frac{s^s}{s!} \rho^s \sum_{i=s}^{\infty} \rho^{i-s} \\ &= p_s \sum_{i=0}^{\infty} \rho^i \\ &= \frac{p_s}{1-\rho}. \end{aligned}$$

Expected Number of Busy Servers. If there are i customers in the system, the number of busy servers is given by $\min(i, s)$. Hence B , the expected number of busy servers, is given by

$$B = \sum_{i=0}^{\infty} \min(i, s) p_i.$$

The sum above can be simplified to

$$B = \frac{\lambda}{\mu}. \quad (6.30)$$

This matches the general result in Example 6.5.

Expected Number in the System. The expected number in the system is given by

$$L = \sum_{i=0}^{\infty} i p_i.$$

This can be simplified to

$$L = \frac{\lambda}{\mu} + p_s \frac{\rho}{(1-\rho)^2}. \quad (6.31)$$

Expected Waiting Time. Using Little's Law, we get

$$W = \frac{L}{\lambda} = \frac{1}{\mu} + \frac{1}{s\mu} \frac{p_s}{(1-\rho)^2}. \quad (6.32)$$

Note that, as expected, both L and W tend to infinity as $\rho \rightarrow 1$.

Example 6.10. (How Many Service Windows?). The U.S. Postal Service has announced that no customer will have to wait more than 5 minutes in line in a post office. It wants to know how many service windows should be kept open to keep this promise. You are hired as a consultant to help decide this. You realize immediately that this is an impossible promise to keep since the service times and arrival rates of customers are beyond the control of the post office. Hence you decide to concentrate on the average wait in the queue. Thus you want to decide how many service windows to keep open so that the mean queueing time (excluding service time) is less than 5 minutes.

Suppose the arrival rate is λ , service times are iid $\text{Exp}(\mu)$, and s windows are kept open. Then, using (6.32), the queueing time is given by

$$W_q(s) = W - 1/\mu = \frac{1}{s\mu} \frac{p_s}{(1-\rho)^2}.$$

You estimate the average service time to be 3.8 minutes and the arrival rate to be 47.2 per hour. Thus

$$\mu = 60/3.8 = 15.7895, \quad \lambda = 47.2.$$

Using the stability condition of Theorem 6.9, we see that the queue is unstable for $s = 1$ and 2. Using the formula above for $s \geq 3$, we get

$$W_q(3) = 5.8977 \text{ hr} = 353.86 \text{ min}, \quad W_q(4) = .0316 \text{ hr} = 1.8986 \text{ min}.$$

Hence you recommend that the post office should keep four windows open. ■

6.4.3 $M/M/\infty$ Queue

By letting $s \rightarrow \infty$, we get the $M/M/\infty$ queue as a limiting case of the $M/M/s$ queue. The limiting distribution of the $M/M/\infty$ queue thus follows from Theorem 6.9 as given below.

Theorem 6.10. (Limiting Distribution of an $M/M/\infty$ Queue). *An $M/M/\infty$ queue with an arrival rate of λ and an unlimited number of identical servers each with service rate μ is always stable, and its limiting distribution is given by*

$$p_i = e^{-\lambda/\mu} \frac{(\lambda/\mu)^i}{i!}, \quad i \geq 0.$$