

Age	Fraction of Population
0	$1/(1 + s + s^2) = .6667$
1	$s/(1 + s + s^2) = .2440$
2	$s^2/(1 + s + s^2) = .0893$

### Problems 8.7

1. Determine the long run population growth rate for a population whose individual net maternity function is  $m_2 = m_3 = 2$ , and  $m_k = 0$ , otherwise. Why does delaying the age at which offspring are first produced cause a reduction in the population growth rate? (The population growth rate when  $m_1 = m_2 = 2$ , and  $m_k = 0$ , otherwise, was determined on page 339.)
2. Determine the long run population growth rate for a population whose individual net maternity function is  $m_0 = m_1 = 0$  and  $m_2 = m_3 = \dots = a > 0$ . Compare this with the population growth rate when  $m_2 = a$ , and  $m_k = 0$  for  $k \neq 2$ .

## Chapter 9 | Queueing Systems

Taylor + Karlin

An Introduction to Stochastic Modeling

Academic Press 1984

### 9.1 Queueing Processes

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

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

While a variety of input processes may arise in practice, two simple and frequently occurring types are mathematically tractable and give insights into more complex cases. First is the scheduled input where customers arrive at fixed times  $T, 2T, 3T, \dots$ . The second most common model is the "completely random" arrival process where the times of customer arrivals form a Poisson process. Understanding the axiomatic development of the Poisson process in Chapter 5 may help one to evaluate the validity of the Poisson assumption in any given application. Many theoretical results are available when the times of customer arrivals form a renewal process.

Exponentially distributed interarrival times then correspond to a Poisson process of arrivals as a special case.

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

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

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

- (1) *The probability distribution of the number of customers in the system.* Not only do customers in the system often incur costs, but in many systems, physical space for waiting customers must be planned for and provided. Large numbers of waiting customers can also adversely affect the input process by turning potential new customers away. (See Section 9.4.1 on queueing with balking.)
- (2) *The utilization of the server(s).* Idle servers may incur costs without contributing to system performance.
- (3) *System throughput.* The long run number of customers passing through the system is a direct measure of system performance.
- (4) *Customer waiting time.* Long waits for service are annoying in the simplest queueing situations and directly associated with major costs in many large systems such as those describing ships waiting to unload at a port facility or patients awaiting emergency care at a hospital.

#### The Queueing Formula $L = \lambda W$

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

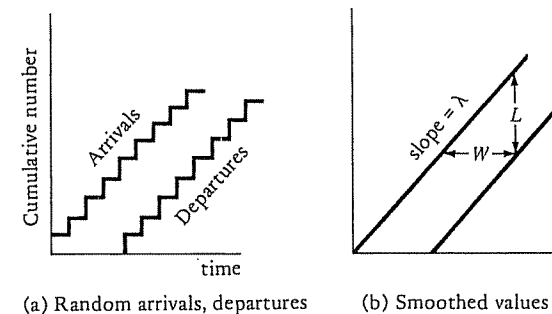
- $L$  = the average number of customers in the system;
- $\lambda$  = the rate of arrival of customers to the system; and
- $W$  = the average time spent by a customer in the system.

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

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

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

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



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

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

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

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

$$L_0 = \lambda W_0. \quad (9.1)$$

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

$$W = W_0 + \text{Mean Service Time}. \quad (9.2)$$

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

Some examples discussed in the sequel are:

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

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

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

More elaborate variations will also be set forth. *Balking* is the refusal of new customers to enter the system if the waiting line is too long. More generally, in a queueing system with balking, an arriving customer enters the system with a probability that depends on the size of the queue. Here it is important to distinguish between the *arrival process* and the *input process* as shown in Figure 9.2. A special case is a queue with *overflow* in which an arriving customer enters the queue if and only if there is at least one server free to begin service immediately.

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

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

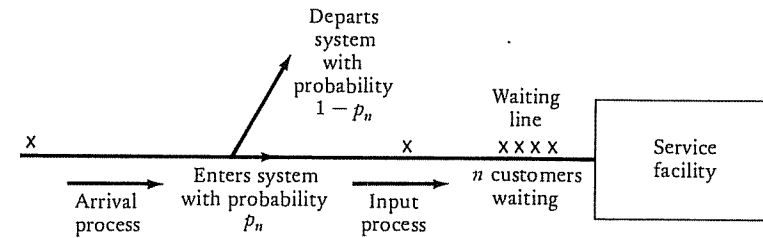


Figure 9.2 If  $n$  customers are waiting in a queueing system with balking, an arriving customer enters the system with probability  $p_n$ , and does not enter with probability  $1 - p_n$ .

### Problems 9.1

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

<i>The Customer</i>	<i>The Server</i>
(a) Arriving airplanes	The runway
(b) Cars	A parking lot
(c) Broken TV's	Repairman
(d) Patients	Doctor
(e) Fires	Fire engine company

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

2. Consider a system, such as a barber shop, where the service required is essentially identical for each customer. Then actual service times would tend to cluster near the mean service time. Argue that the exponential distribution would not be appropriate in this case. For what types of service situations might the exponential distribution be quite plausible?
3. Two dump trucks cycle between a gravel loader and a gravel unloader. Suppose that the travel times are insignificant relative to the load and unload times, which are exponentially distributed with parameters  $\mu$  and  $\lambda$ , respectively. Model the system as a closed queueing network. Determine the long run gravel loads moved per unit time. *Hint:* Refer to the example entitled "Repairman Models" in Section 6.4.

### 9.2 Poisson Arrivals and Exponentially Distributed Service Times

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

We let  $\lambda$  denote the intensity or rate of the Poisson arrival process and assume that the service time distribution is exponential with parameter  $\mu$ . The corresponding density function is

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

For the Poisson arrival process we have

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

and

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

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

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

and

$$\begin{aligned} \Pr\{\text{Service not completed in } [t, t + h] | \text{Service in progress at time } t\} \\ = 1 - \mu h + o(h). \end{aligned} \quad (9.7)$$

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

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

#### The M/M/1 System

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

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

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

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

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

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

and death parameters

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

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

Let

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

be the limiting or equilibrium distribution of queue length. Section 6.4 describes a straightforward procedure for determining the limiting distribution  $\pi_k$  from the birth and death parameters  $\lambda_k$  and  $\mu_k$ . The technique is to first obtain intermediate quantities  $\theta_j$  defined by

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

and then

$$\pi_0 = \frac{1}{\sum_{j=0}^{\infty} \theta_j} \quad \text{and} \quad \pi_k = \theta_k \pi_0 = \frac{\theta_k}{\sum_{j=0}^{\infty} \theta_j} \quad \text{for } k \geq 1. \quad (9.9)$$

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

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

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

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

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

$$\pi_0 = \frac{1}{\sum_{j=0}^{\infty} \theta_j} = 1 - \frac{\lambda}{\mu} \quad (9.10)$$

and

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

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

$$L = \frac{\lambda}{\mu - \lambda} \quad (9.12)$$

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

$$\rho = \frac{\text{Arrival rate}}{\text{System service rate}} = \frac{\lambda}{\mu} \quad (9.13)$$

As the traffic intensity approaches one, the mean queue length  $L = \rho/(1 - \rho)$  becomes infinite. Again using (9.8), the probability of being served immediately upon arrival is

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

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

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

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

By the law of total probabilities, we have

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

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

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

$$= \int_0^t \left(1 - \frac{\lambda}{\mu}\right) \mu \exp\left\{-\tau\mu\left(1 - \frac{\lambda}{\mu}\right)\right\} d\tau = 1 - \exp[-t(\mu - \lambda)],$$

which is also an exponential distribution.

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

$$W = \frac{1}{\mu - \lambda} \quad (9.15)$$

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

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

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

A busy period is an uninterrupted duration in which the system is not empty. When arrivals to a queue follow a Poisson process, then the successive durations  $X_k$  from the commencement of the  $k$ th busy period to the start of the next busy period form a renewal process (see Figure 9.3). Each  $X_k$  is comprised of a busy portion  $B_k$  and an idle portion  $I_k$ . Then the renewal theorem (see "A Queuing Model," p. 294) applies to tell us that  $p_0(t)$ , the probability that the system is empty at time  $t$ , converges to

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

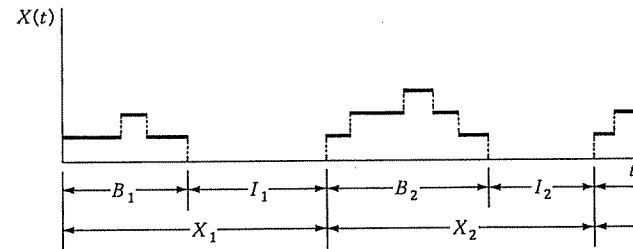


Figure 9.3 The busy periods  $B_k$  and idle periods  $I_k$  of a queuing system. When arrivals form a Poisson process, then  $X_k = B_k + I_k$ ,  $k = 1, 2, \dots$  are independent identically distributed non-negative random variables, and thus form a renewal process.

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

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

which solves to give

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

for the mean length of a busy period.

In Section 9.3 in studying the  $M/G/1$  system we will reverse this reasoning, calculate the mean busy period directly, and then use renewal theory to determine the server idle fraction  $\pi_0$ .

### The $M/M/\infty$ System

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

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

The auxiliary quantities of (9.8) are

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

which sum to

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

whence

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

and

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

a Poisson distribution with mean queue length

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

Since a customer in this system begins service immediately upon arrival, customer waiting time consists only of the exponentially distributed

service time, and the mean waiting time is  $W = 1/\mu$ . Again, the basic queueing formula  $L = \lambda W$  is verified.

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

### The $M/M/s$ System

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

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

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

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

The auxiliary quantities are given by

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

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

$$\sum_{j=0}^{\infty} \theta_j = \sum_{j=0}^{s-1} \frac{1}{j!} \left(\frac{\lambda}{\mu}\right)^j + \sum_{j=s}^{\infty} \frac{1}{s!} \left(\frac{\lambda}{\mu}\right)^s \left(\frac{\lambda}{s\mu}\right)^{j-s} \quad (9.17)$$

$$= \sum_{j=0}^{s-1} \frac{1}{j!} \left(\frac{\lambda}{\mu}\right)^j + \frac{(\lambda/\mu)^s}{s!(1 - \lambda/s\mu)} \quad \text{for} \quad \lambda < s\mu.$$

The traffic intensity in an  $M/M/s$  system is  $\rho = \lambda/s\mu$ . Again as the

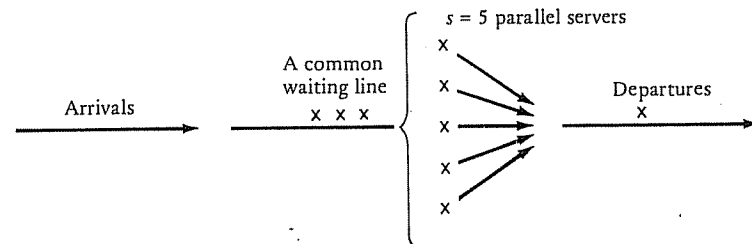


Figure 9.4 A queueing system with  $s$  servers

traffic intensity approaches one, the mean queue length becomes unbounded. When  $\lambda < s\mu$ , then from (9.9) and (9.17),

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

and

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

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

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

Then

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

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

and

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

## Problems 9.2

1. On a single graph, plot the server utilization  $1 - \pi_0 = \rho$  and the mean queue length  $L = \rho/(1 - \rho)$  for the  $M/M/1$  queue as a function of the traffic intensity  $\rho = \lambda/\mu$  for  $0 < \rho < 1$ .
2. Determine explicit expressions for  $\pi_0$  and  $L$  for the  $M/M/s$  queue when  $s = 2$ . Plot  $1 - \pi_0$  and  $L$  as a function of the traffic intensity  $\rho = \lambda/2\mu$ .
3. Determine the mean waiting time  $W$  for an  $M/M/2$  system when  $\lambda = 2$  and  $\mu = 1.2$ . Compare this with the mean waiting time in an  $M/M/1$

system whose arrival rate is  $\lambda = 1$  and service rate is  $\mu = 1.2$ . Why is there a difference when the arrival rate per server is the same in both cases?

4. The problem is to model a queueing system having finite capacity. We assume arrivals according to a Poisson process of rate  $\lambda$ , independent exponentially distributed service times having mean  $1/\mu$ , a single server, and a finite system capacity  $N$ . By this we mean that if an arriving customer finds that there are already  $N$  customers in the system, then that customer does not enter the system and is lost. Let  $X(t)$  be the number of customers in the system at time  $t$ . Suppose that  $N = 3$  (2 waiting, 1 being served).
  - (a) Specify the birth and death parameters for  $X(t)$ .
  - (b) In the long run, what fraction of time is the system idle?
  - (c) In the long run, what fraction of customers are lost?

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

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

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

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

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

## 9.3 The $M/G/1$ and $M/G/\infty$ Systems

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

### The M/G/1 System

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

$$\begin{aligned} \lim_{t \rightarrow \infty} p_0(t) &= \pi_0 = \frac{E[I_1]}{E[X_1]} \\ &= \frac{E[I_1]}{E[I_1] + E[B_1]} \end{aligned} \quad (9.20)$$

by the renewal theorem (see "A Queueing Model," p. 294).

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

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

$$E[B_1|A = 0, Y_1 = \gamma] = \gamma$$

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

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

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

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

and then, using the law of total probability, that

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

Finally

$$\begin{aligned} E[B_1] &= \int_0^{\infty} E[B_1|Y_1 = \gamma] dG(\gamma) \\ &= \int_0^{\infty} \{\gamma + \lambda\gamma E[B_1]\} dG(\gamma) \\ &= \nu\{1 + \lambda E[B_1]\}. \end{aligned} \quad (9.21)$$

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

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

To compute the long run fraction of idle time, we use (9.20) and

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

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

### The Embedded Markov Chain

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

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

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

where  $A_n$  is the number of customers that arrive during the service of the  $n$ th customer and where  $x^+ \doteq \max\{x, 0\}$ . Since the arrival process is Poisson, the number of customers  $A_n$  that arrive during the service of the  $n$ th

customer is independent of earlier arrivals, and the Markov property follows instantly. We calculate

$$\begin{aligned}\alpha_k &= \Pr\{A_n = k\} = \int_0^{\infty} \Pr\{A_n = k | Y_n = \gamma\} dG(\gamma) \\ &= \int_0^{\infty} \frac{(\lambda\gamma)^k e^{-\lambda\gamma}}{k!} dG(\gamma)\end{aligned}\quad (9.25)$$

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

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

### The Mean Queue Length in Equilibrium $L$

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

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

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

$$X' = X - \delta + N \quad (9.27)$$

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

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

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

$$L = E[X] = E[X'], \quad (9.28)$$

and taking expectation in (9.27) gives

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

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

$$E[N] = E[\delta] = 1 - \pi_0 = \lambda\nu. \quad (9.29)$$

Squaring (9.27) gives

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

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

$$(X')^2 = X^2 + \delta + N^2 - 2X + 2N(X - \delta). \quad (9.30)$$

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

$$E[N(X - \delta)] = E[N]E[X - \delta] \quad (9.31)$$

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

$$E[(X')^2] = E[X^2]. \quad (9.32)$$

Taking expectations in (9.30) we deduce that

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

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

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

or

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

It remains to evaluate  $E[N^2]$  where  $N$  is the number of arrivals during a service time  $Y$ . Conditioned on  $Y = \gamma$ , the random variable  $N$  has a Poisson distribution with a mean (and variance) equal to  $\lambda\gamma$  [see (9.25)], whence  $E[N^2 | Y = \gamma] = \lambda\gamma + (\lambda\gamma)^2$ . Using the law of total probability then gives

$$\begin{aligned}E[N^2] &= \int_0^{\infty} E[N^2 | Y = \gamma] dG(\gamma) \\ &= \lambda \int_0^{\infty} \gamma dG(\gamma) + \lambda^2 \int_0^{\infty} \gamma^2 dG(\gamma) \\ &= \lambda\nu + \lambda^2(\tau^2 + \nu^2)\end{aligned}\quad (9.34)$$

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

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

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

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

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

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

### The M/G/∞ System

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

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

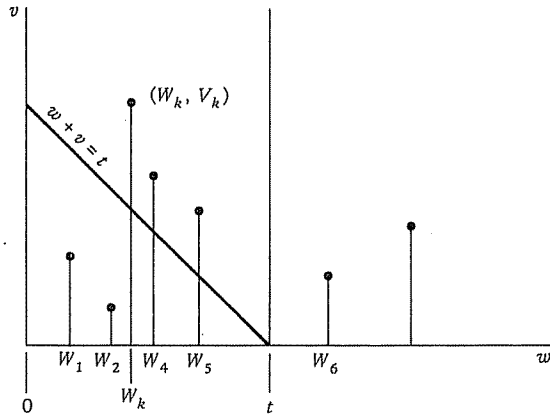


Figure 9.5 For the M/G/∞ queue the number of customers in the system at time  $t$  corresponds to the number of pairs  $(W_k, V_k)$  for which  $W_k \leq t$  and  $W_k + V_k > t$ . In the sample illustrated here, the number of customers in the system at time  $t$  is 3.

That is, it is the number of points  $(W_k, V_k)$  in the unbounded trapezoid described by

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

According to Theorem 5.6, the number of points in  $A_t$  follows a Poisson distribution with mean

$$\begin{aligned} \mu(A_t) &= \iint_{A_t} \lambda(dw)dG(v) \\ &= \lambda \int_0^t \left\{ \int_{t-w}^{\infty} dG(v) \right\} dw \\ &= \lambda \int_0^t [1 - G(t - w)] dw \\ &= \lambda \int_0^t [1 - G(x)] dx. \end{aligned} \tag{9.37}$$

In summary,

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

where  $\mu(A_t)$  is given by (9.37). As  $t \rightarrow \infty$  then

$$\lim_{t \rightarrow \infty} \mu(A_t) = \lambda \int_0^{\infty} [1 - G(x)] dx = \lambda \nu$$

where  $\nu$  is the mean service time. Thus we obtain the limiting distribution

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

### Appendix

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