Chapter 7
Queueing Processes

Many phenomena for which mathematical descriptions are desired involve waiting lines either of people or material. A queue is a waiting line, and queueing processes are those stochastic processes arising from waiting line phenomena. For example, the modeling of the arrival process of grain trucks to an elevator, the utilization of data processing services at a computer center, and the flow of jobs at a job shop facility all involve waiting lines. Although queues are ubiquitous, they are usually ignored when deterministic models are developed to describe systems. Furthermore, the random fluctuations inherent in queueing processes often cause systems to act in a counter intuitive fashion. Therefore, the study of queues is extremely important for the development of system models and an understanding of system behavior.

In this chapter we present modeling techniques employed for queueing systems governed by the exponential process. The final section of the chapter deals with some approximation techniques useful for implementing these models within complex systems when the exponential assumptions are not satisfied. The next chapter deals with processes that arise when several queueing system operate within one system; that is, Chap. 8 deals with queueing networks. The final chapter (Chap. 13) in the textbook presents some advanced analytical techniques useful for modeling non-exponential queueing systems.

7.1 Basic Definitions and Notation

A queueing process involves the arrival of customers to a service facility and the servicing of those customers. All customers that have arrived but are not yet being served are said to be in the *queue*. The queueing *system* includes all customers in the queue and all customers in service (see Fig. 7.1).

Several useful conventions have evolved over the last 20-40 years that help in specifying the assumptions used in a particular analysis. D.G. Kendall [3] is usually given credit for initiating the basic notation of today, and it was standardized in 1971 (*Queueing Standardization Conference Report*, May 11, 1971). Kendall's notation is

R.M. Feldman, C. Valdez-Flores, *Applied Probability and Stochastic Processes*, 2nd ed.,
DOI 10.1007/978-3-642-05158-6_7,

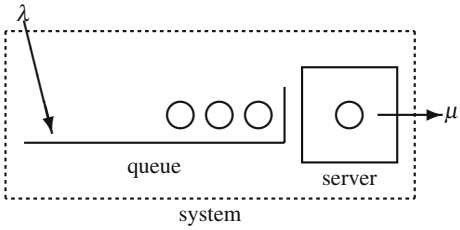

Fig. 7.1 Representation of a queueing system with a mean arrival rate of λ, and mean service rate of μ, four customers in the system, and three in the queue

a shorthand to indicate quickly the assumptions used in a particular queueing model. For example, a formula developed for a G/D/1/∞/FIFO queue would be a formula that could be used for any general arrival process, only a deterministic service time, one server, an unlimited system capacity, and a discipline that serves on a "first-in first-out" basis. The general notation has the following form.

$$\left(\begin{array}{c}\text{arrival}\\ \text{process}\end{array}\Big/\begin{array}{c}\text{service}\\ \text{process}\end{array}\Big/\begin{array}{c}\text{number}\\ \text{of servers}\end{array}\Big/\begin{array}{c}\text{maximum}\\ \text{possible}\\ \text{in system}\end{array}\Big/\begin{array}{c}\text{queue}\\ \text{discipline}\end{array}\right)$$

Table (7.1) gives the common abbreviations used with this notation. Whenever an infinite capacity and FIFO discipline are used, the last two descriptors can be left off; thus, an M/M/1 queue would refer to exponential inter-arrival and service times, one server, unlimited capacity, and a FIFO discipline.

Table 7.1 Queueing symbols used with Kendall's notation

Symbols	Explanation
M	Exponential (Markovian) inter-arrival or service time
D	Deterministic inter-arrival or service time
E_k	Erlang type k inter-arrival or service time
G	General inter-arrival or service time
$1, 2, \cdots, \infty$	Number of parallel servers or capacity
FIFO	First in, first out queue discipline
LIFO	Last in, first out queue discipline
SIRO	Service in random order
PRI	Priority queue discipline
GD	General queue discipline

Our purpose in this chapter is to give an introduction to queueing processes and introduce the types of problems commonly encountered while studying queues. To maintain the introductory level of this material, arrival processes to the queueing systems will be assumed to be Poisson processes (see Chap. 4), and service times will be exponential. Thus, this chapter is mainly concerned with investigating M/M/c/K systems for various values of c and K. The last section of the chapter will present a simple approximation that can be used for simple systems.

7.2 Single Server Systems

The simplest queueing systems to analyze are those involving a Poisson arrival process and a single exponential server. Such systems are not only relatively easy to study, but they will serve to demonstrate some general queueing analysis techniques that can be extended to more complicated system. We shall start by considering a system that has unlimited space for arriving customers and then move to systems with limited space.

7.2.1 Infinite Capacity Single-Server Systems

We begin with an M/M/1 system, or equivalently, an M/M/1∞/FIFO system. The M/M/1 system assumes customers arrive according to a Poisson process with mean rate λ and are served by a single server whose time for service is random with an exponential distribution of mean $1/\mu$. If the server is idle and a customer arrives, then that customer enters the server immediately. If the server is busy and a customer arrives, then the arriving customer enters the queue which has infinite capacity. When service for a customer is completed, the customer leaves and the customer that had been in the queue the longest instantaneously enters the service facility and service begins again. Thus, the flow of customers through the system is a Markov process with state space $\{0,1,\cdots\}$. The Markov process is denoted by $\{N_t; t \geq 0\}$ where N_t denotes the number of customers in the system at time t. The steady-state probabilities are

$$p_n = \lim_{t\to\infty} Pr\{N_t = n\} .$$

We let N be a random variable with probability mass function $\{p_0, p_1, \cdots\}$. The random variable N thus represents the number of customers in the system at steady-state, and p_n represents the long-run probability that there are n customers in the system. (You might also note that another way to view p_n is as the long-run fraction of time that the system contains n customers.) Sometimes we will be interested in the number of customers that are in the queue and thus waiting for service; therefore, let the random variable N_q denote the steady-state number in the queue. In other words, if the system is idle, $N_q = N$; when the system is busy, $N_q = N - 1$.

Our immediate goal is to derive an expression for $p_n, n = 0,1,\cdots$, in terms of the mean arrival and service rates. This derivation usually involves two steps: (1) obtain a system of equations defining the probabilities and (2) solve the system of equations. After some experience, you should find Step (1) relatively easy; it is usually Step (2) that is difficult. In other words, the system of equations defining p_n is not hard to obtain, but it is sometimes hard to solve.

An intuitive approach for obtaining the system of equations is to draw a state diagram and then use a rate balance approach, which is a system of equations formed by setting "rate in" equal to "rate out" for each node or state of the queueing system. Figure 7.2 shows the state diagram for the M/M/1 system. Referring to this figure,

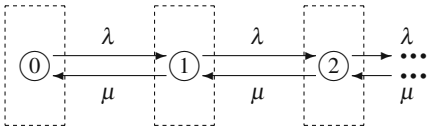

Fig. 7.2 State diagram for an M/M/1 queueing system illustrating the rate balance approach

it is seen that the rate into the box around Node 0 is μp_1; the rate out of the box around Node 0 is λp_0; thus, "rate in" = "rate out" yields

$$\mu p_1 = \lambda p_0 \, .$$

The rate into the box around Node 1 is $\lambda p_0 + \mu p_2$; the rate out of the box around Node 1 is $(\mu + \lambda) p_1$; thus

$$\lambda p_0 + \mu p_2 = (\mu + \lambda) p_1 \, .$$

Continuing in a similar fashion and rearranging, we obtain the system

$$\begin{aligned} p_1 &= \frac{\lambda}{\mu} p_0 \text{ and} \\ p_{n+1} &= \frac{\lambda + \mu}{\mu} p_n - \frac{\lambda}{\mu} p_{n-1} \text{ for } n = 1, 2, \cdots \, . \end{aligned} \tag{7.1}$$

A more rigorous approach for obtaining the system of Eqs. (7.1) is to first give the generator matrix (from Chap. 6) for the M/M/1 system. Since the inter-arrival and service times are exponential, the queueing system is a Markov process and thus Property 6.1 can be used. The rate at which the process goes from State n to State $n+1$ is λ, and the rate of going from State n to State $n-1$ is μ; therefore, the generator is the infinite dimensioned matrix given as

$$\mathbf{G} = \begin{bmatrix} -\lambda & \lambda & & & \\ \mu & -(\mu+\lambda) & \lambda & & \\ & \mu & -(\mu+\lambda) & \lambda & \\ & & \ddots & \ddots & \ddots \\ & & & & \end{bmatrix} \, . \tag{7.2}$$

The system of equations formed by $\mathbf{pG} = \mathbf{0}$ then yields Eqs. (7.1) again.

The system of Eqs. (7.1) can be solved by successively forward substituting solutions and expressing all variables in terms of p_0. Since we already have $p_1 = (\lambda/\mu)\, p_0$, we look at p_2 and then p_3:

$$\begin{aligned} p_2 &= \frac{\lambda + \mu}{\mu} p_1 - \frac{\lambda}{\mu} p_0 \\ &= \frac{\lambda + \mu}{\mu} \left(\frac{\lambda}{\mu} p_0\right) - \frac{\lambda}{\mu} \frac{\mu}{\mu} p_0 \; = \; \frac{\lambda^2}{\mu^2} p_0 \, , \end{aligned} \tag{7.3}$$

$$p_3 = \frac{\lambda+\mu}{\mu}p_2 - \frac{\lambda}{\mu}p_1$$
$$= \frac{\lambda+\mu}{\mu}(\frac{\lambda^2}{\mu^2}p_0) - \frac{\lambda}{\mu}\frac{\mu}{\mu}(\frac{\lambda}{\mu}p_0) = \frac{\lambda^3}{\mu^3}p_0 .$$

At this point, a pattern begins to emerge and we can assert that

$$p_n = \frac{\lambda^n}{\mu^n}p_0 \text{ for } n \geq 0 . \tag{7.4}$$

The assertion is proven by mathematical induction; that is, using the induction hypothesis together with the general equation in Eq. (7.1) yields

$$p_{n+1} = \frac{\lambda+\mu}{\mu}(\frac{\lambda^n}{\mu^n}p_0) - \frac{\lambda}{\mu}\frac{\mu}{\mu}(\frac{\lambda^{n-1}}{\mu^{n-1}}p_0)$$
$$= \frac{\lambda^{n+1}}{\mu^{n+1}}p_0 ,$$

and thus Eq. (7.4) is shown to hold. The ratio λ/μ is called the *traffic intensity* for the queueing system and is denoted by ρ for the M/M/1 system. (More generally, ρ is usually defined as the arrival rate divided by the maximum system service rate.)

We now have p_n for all n in terms of p_0 so the long-run probabilities become known as soon as p_0 can be obtained. If you review the material on Markov processes, you should see that we have taken advantage of Property 6.1, except we have not yet used the second equation given in the property. Thus, an expression for p_0 can be determined by using the norming equation, namely

$$1 = \sum_{n=0}^{\infty} p_n = p_0 \sum_{n=0}^{\infty} \frac{\lambda^n}{\mu^n} = p_0 \sum_{n=0}^{\infty} \rho^n \tag{7.5}$$
$$= \frac{p_0}{1-\rho} .$$

The equality in the above expression made use of the geometric progression[1] so it is only valid for $\rho < 1$. If $\rho \geq 1$, the average number of customers and time spent in the system increase without bound and the system becomes unstable. In some respects this is a surprising result. Based on deterministic intuitive, a person might be tempted to design a system such that the service rate is equal to the arrival rate, thus creating a "balanced" system. This is false logic for random interarrival or service times since in that case the system will never reach steady-state.

The above value for p_0 can be combined with Eq. (7.4) to obtain for the M/M/1 system

$$p_n = (1-\rho)\rho^n \text{ for } n = 0, 1, \cdots , \tag{7.6}$$

where $\rho = \lambda/\mu$ and $\rho < 1$.

[1] The geometric progression is $\sum_{n=0}^{\infty} r^n = 1/(1-r)$ for $|r| < 1$.

The steps followed in deriving Eq. (7.6) are the pattern for many other Markov queueing systems. Once the derivation of the M/M/1 system is known, all other queueing system derivations in this text will be easy. It is, therefore, good to review these steps so that they become familiar.

1. Form the Markov generator matrix, **G** (Eq. 7.2).
2. Obtain a system of equations by solving $\mathbf{pG} = \mathbf{0}$ (Eq. 7.1).
3. Solve the system of equations in terms of p_0 by successive forward substitution and induction if possible (Eq. 7.4).
4. Use the norming equation to find p_0 (Eq. 7.5).

Once the procedure becomes familiar, the only difficult step will be the third step. It is not always possible to find a closed-form solution to the system of equations, and often techniques other than successive forward substitution must be used. However, these techniques are beyond the scope of this text and will not be presented.

Example 7.1. An operator of a small grain elevator has a single unloading dock. Arrivals of trucks during the busy season form a Poisson process with a mean arrival rate of four per hour. Because of varying loads (and desire of the drivers to talk) the length of time each truck spends in front of the unloading dock is approximated by an exponential random variable with a mean time of 14 minutes. Assuming that the parking spaces are unlimited, the M/M/1 queueing system describes the waiting lines that form. Accordingly, we have

$$\lambda = 4/\text{hr}, \;\; \mu = \frac{60}{14}/\text{hr}, \;\; \rho = 0.9333 \, .$$

The probability of the unloading dock being idle is

$$p_0 = 1 - \rho = 0.0667 \, .$$

The probability that there are exactly three trucks waiting is

$$Pr\{N_q = 3\} = Pr\{N = 4\} = p_4 = 0.9333^4 \times 0.0667 = 0.05 \, .$$

Finally, the probability that four or more trucks are in the system is

$$Pr\{N \geq 4\} = \sum_{n=4}^{\infty} p_n = (1-\rho) \sum_{n=4}^{\infty} \rho^n = \rho^4 = 0.759 \, .$$

(In the above expression, the second equality is obtained by using Eq. (7.6) to substitute out p_n. The third equality comes by observing that ρ^4 is a multiplicative factor in each term of the series so that it can be "moved" outside the summation sign, making the resulting summation a geometric progression.) □

Several measures of effectiveness are useful as descriptors of queueing systems. The most common measures are the expected number of customers in the system, denoted by L, and the expected number in the queue, denoted by L_q. These expected

values are obtained by utilizing Eq. (7.6) and the derivative of the geometric progression[2], yielding an expression for the expected number in an M/M/1 system as:

$$\begin{aligned} L = E[N] &= \sum_{n=0}^{\infty} np_n \\ &= \sum_{n=1}^{\infty} np_n = \sum_{n=1}^{\infty} n\rho^n(1-\rho) \\ &= (1-\rho)\rho \sum_{n=1}^{\infty} n\rho^{n-1} = \frac{\rho}{1-\rho} \,. \end{aligned} \tag{7.7}$$

The expected number of customers waiting within an M/M/1 queueing system is obtained similarly:

$$\begin{aligned} L_q &= 0 \times (p_0 + p_1) + \sum_{n=1}^{\infty} np_{n+1} \\ &= \sum_{n=1}^{\infty} n\rho^{n+1}(1-\rho) \\ &= (1-\rho)\rho^2 \sum_{n=1}^{\infty} n\rho^{n-1} = \frac{\rho^2}{1-\rho} \,. \end{aligned} \tag{7.8}$$

When describing a random variable, it is always dangerous to simply use its mean value as the descriptor. For this reason, we also give the variance of the number in the system and queue:

$$V[N] = \frac{\rho}{(1-\rho)^2} \,, \tag{7.9}$$

$$V[N_q] = \frac{\rho^2(1+\rho-\rho^2)}{(1-\rho)^2} \,. \tag{7.10}$$

Waiting times are another important measure of a queueing system. Fortunately for our computational effort, there is an easy relationship between the mean number waiting and average length of time a customer waits. Little [4] showed in 1961 that for almost *all steady-state queueing systems* there is a simple relationship between the mean number in the system, the mean waiting times, and the arrival rates.

Property 7.1. ***Little's Law***. *Consider a queueing system for which steady-state occurs. Let $L = E[N]$ denote the mean long-run number in the system, $W = E[T]$ denote the mean long-run waiting time within the system, and λ_e the mean arrival rate of jobs* into *the system. Also let $L_q = E[N_q]$ and $W_q = E[T_q]$ denote the analogous quantities restricted to the queue. Then*

[2] Taking the derivative of both sides of the geometric progression yields $\sum_{n=1}^{\infty} nr^{n-1} = 1/(1-r)^2$ for $|r| < 1$.

$$L = \lambda_e W$$
$$L_q = \lambda_e W_q \, .$$

Notice that λ_e refers to the *effective* mean arrival rate into the system; whereas, λ refers to the mean arrival rate to the system. In other words, λ includes those customers who come to the system but for some reason, like a finite capacity system that is full, they do not enter; λ_e only counts those customers who make it to the server. For the M/M/1 system, the effective arrival rate is the same as the arrival rate (i.e., $\lambda_e = \lambda$); thus

$$W = E[T] = \frac{1}{\mu - \lambda} \, , \tag{7.11}$$
$$W_q = E[T_q] = \frac{\rho}{\mu - \lambda} \, ,$$

where T is the random variable denoting the time a customer (in steady-state) spends in the system and T_q is the random variable for the time spent in the queue.

When the arrival process is Poisson, there is a generalization of Little's formula that holds for variances.

Property 7.2. *Consider a queueing system for which steady-state occurs and with a Poisson arrival stream of customers entering the system. Let N denote the number in the system, T denote the customer waiting time within the system, and* λ_e *the mean arrival rate of jobs* into *the system. Also let* N_q *and* T_q *denote the analogous quantities restricted to the queue. Then the following hold:*

$$V[N] - E[N] = \lambda_e^2 V[T]$$
$$V[N_q] - E[N_q] = \lambda_e^2 V[T_q] \, .$$

Little's Law (Property 7.1) is a very powerful result because of its generality. The version applied to variances (Property 7.2) is not quite as powerful since it is restricted to Poisson arrivals. Applying Property 7.2 to the M/M/1 system we obtain

$$V[T] = \frac{1}{(\mu - \lambda)^2} = \frac{1}{\mu^2 (1-\rho)^2} \tag{7.12}$$
$$V[T_q] = \frac{2\rho - \rho^2}{(\mu - \lambda)^2} = \frac{\rho(2-\rho)}{\mu^2 (1-\rho)^2} \, .$$

In Example 7.1, the mean number of trucks in the system is 14 (Eq. 7.7) with a standard deviation of 14.5 (Eq. 7.9), the mean number of trucks in the queue is

approximately 13.1 (Eq. 7.8) with standard deviation of 14.4 (Eq. 7.10), the mean time each truck spends in the system is 3.5 hours (Eq. 7.11) with a standard deviation of 3.5 hours (Eq. 7.12), and the mean time each truck waits in line until his turn at the dock is 3 hours and 16 minutes (Eq. 7.11) with a standard deviation of 3 hours and 29 minutes (Eq. 7.12).

Another easy and obvious relationship for queueing systems due to the definition of the system and the queue is that the mean waiting time in the system must equal the mean time in the queue plus the mean service time; thus, we have the following.

Property 7.3. *Consider a queueing system for which steady-state occurs. Let W denote the mean long-run waiting time within the system, W_q the mean long-run waiting time in the queue, and μ the mean service rate; then*

$$W = W_q + \frac{1}{\mu}\,.$$

Notice that this property is general like Little's Law; namely, it holds for non-exponential and multi-server systems as well as for finite and infinite capacity systems.

Example 7.2. A large car dealer has a policy of providing cars for its customers that have car problems. When a customer brings the car in for repair, that customer has use of a dealer's car. The dealer estimates that the dealer cost for providing the service is \$10 per day for as long as the customer's car is in the shop. (Thus, if the customer's car was in the shop for 1.5 days, the dealer's cost would be \$15.) Arrivals to the shop of customers with car problems form a Poisson process with a mean rate of one every other day. There is one mechanic dedicated to those customer's cars. The time that the mechanic spends on a car can be described by an exponential random variable with a mean of 1.6 days. We would like to know the expected cost per day of this policy to the car dealer. Assuming infinite capacity, we have the assumptions of the M/M/1 queueing system satisfied, with $\lambda = 0.5$/day and $\mu = 0.625$/day, yielding a $\rho = 0.8$. (Note the mean rate is the *reciprocal* of the mean time.) Using the M/M/1 equations, we have $L = 4$ and $W = 8$ days. Thus, whenever a customer comes in with car problems, it will cost the dealer \$80. Since a customer comes in every other day (on the average) the total cost to the dealer for this policy is \$40 per day. In other words, cost is equal to $\$10 \times W \times \lambda$. But by Little's formula, this is equivalent to $\$10 \times L$. The cost structure illustrated with this example is a very common occurrence for queueing systems. In other words if c is the cost per item per time unit that the item spends in the system, the expected system cost per time unit is cL. □

- *Suggestion: Do Problems 7.1–7.4.*

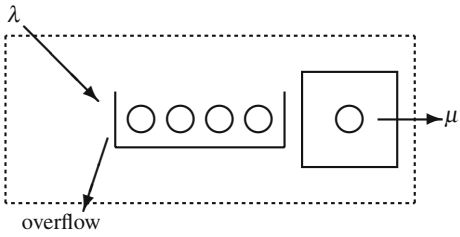

Fig. 7.3 Representation of a full M/M/1/5 queueing system

7.2.2 Finite Capacity Single Server Systems

The assumption of infinite capacity is often not suitable. When a finite system capacity is necessary, the state probabilities and measures of effectiveness presented in Eqs. (7.6–7.12) are inappropriate to use; thus, new probabilities and measures of effectiveness must be developed for the M/M/1/K system (see Fig. 7.3).

As will be seen, the equations for the state probabilities are identical except for the norming equation. If K is the maximum number of customers possible in the system, the generator matrix is of dimension $(K+1)\times(K+1)$ and has the form

$$\mathbf{G} = \begin{bmatrix} -\lambda & \lambda & & & \\ \mu & -(\lambda+\mu) & \lambda & & \\ & \ddots & \ddots & \ddots & \\ & & \mu & -(\lambda+\mu) & \lambda \\ & & & \mu & -\mu \end{bmatrix} .$$

The system $\mathbf{pG} = \mathbf{0}$ yields

$$\begin{aligned} \mu p_1 &= \lambda p_0 \\ \mu p_{n+1} &= (\lambda+\mu)p_n - \lambda p_{n-1} \quad \text{for } n = 1,\cdots,K-1 \\ \mu p_K &= \lambda p_{K-1} . \end{aligned} \tag{7.13}$$

Using successive substitution, we again have

$$p_n = \rho^n p_0 \text{ for } n = 0,1,\cdots,K ,$$

where $\rho = \lambda/\mu$. The last equation in (7.13) is ignored since there is always a redundant equation in a finite irreducible Markov system. The norming equation is now used to obtain p_0 as follows:

$$1 = \sum_{n=0}^{K} p_n = p_0 \sum_{n=0}^{K} \rho^n = \begin{cases} p_0 \frac{1-\rho^{K+1}}{1-\rho} & \text{for } \rho \neq 1, \\ p_0 (K+1) & \text{for } \rho = 1 . \end{cases}$$

Since the above sum is finite, we used the finite geometric progression[3] so that ρ may be larger than one. Therefore, for an M/M/1/K system,

$$p_n = \begin{cases} \rho^n \frac{1-\rho}{1-\rho^{K+1}} & \text{for } \rho \neq 1, \\ \frac{1}{K+1} & \text{for } \rho = 1\,, \end{cases} \tag{7.14}$$

for $n = 0, 1, \cdots, K$.

The mean for the number in the system and queue are

$$L = \sum_{n=0}^{K} np_n = \sum_{n=1}^{K} np_n = p_0\rho \sum_{n=1}^{K} n\rho^{n-1} \tag{7.15}$$

$$= \begin{cases} \rho \frac{1+K\rho^{K+1}-(K+1)\rho^K}{(1-\rho)(1-\rho^{K+1})} & \text{for } \rho \neq 1 \\ \frac{K}{2} & \text{for } \rho = 1 \end{cases}$$

$$L_q = \sum_{n=1}^{K-1} np_{n+1} = p_0\rho^2 \sum_{n=1}^{K-1} n\rho^{n-1} \tag{7.16}$$

$$= \begin{cases} L - \frac{\rho(1-\rho^K)}{1-\rho^{K+1}} & \text{for } \rho \neq 1 \\ \frac{K(K-1)}{2(K+1)} & \text{for } \rho = 1\,. \end{cases}$$

The variances for these quantities for the M/M/1/K system are

$$V[N] = \begin{cases} [\rho/(1-\rho^{K+1})(1-\rho)^2] \\ \quad \times [1+\rho-(K+1)^2\rho^K \\ \quad + (2K^2+2K-1)\rho^{K+1} - K^2\rho^{K+2}] - L^2 & \text{for } \rho \neq 1\,, \\ K(K+2)/12 & \text{for } \rho = 1\,. \end{cases}$$

$$V[N_q] = V[N] - p_0(L + L_q) \tag{7.17}$$

The probability that an arriving customer enters the system is the probability that the system is not full. Therefore, to utilize Little's formula, we set $\lambda_e = \lambda(1-p_K)$ for the effective arrival rate to obtain the waiting time equations as follows:

$$W = \frac{L}{\lambda(1-p_K)}\,,$$

[3] The finite geometric progression is $\sum_{n=0}^{k-1} r^n = (1-r^k)/(1-r)$ if $r \neq 1$.

$$W_q = W - \frac{1}{\mu} .$$

Notice that the expression for W_q is simply a restatement of Property 7.3.

The temptation is to use the formulas given by Property 7.2 to obtain expressions for the variances of the waiting times. However, the "Little-like" relationships for variances are based on the assumption that the system sees Poisson arrivals. The finite capacity limitation prohibits Poisson arrivals *into* the system so the variance generalization of Little's Law cannot be used.

Example 7.3. A corporation must maintain a large fleet of tractors. They have one repairman that works on the tractors as they break down on a first-come first-serve basis. The arrival of tractors to the shop needing repair work is approximated by a Poisson distribution with a mean rate of three per week. The length of time needed for repair varies according to an exponential distribution with a mean repair time of $1/2$ week per tractor. The current corporate policy is to utilize an outside repair shop whenever more than two tractors are in the company shop so that, at most, one tractor is allowed to wait. Each week that a tractor spends in the shop costs the company \$100. To utilize the outside shop costs \$500 per tractor. (The \$500 includes lost time.) We wish to review corporate policy and determine the optimum cutoff point for the outside shop; that is, we shall determine the maximum number allowed in the company shop before sending tractors to the outside repair facility. The total operating costs per week are

$$\text{Cost } = 100L + 500\lambda p_K .$$

For the current policy, which is an M/M/1/2 system, we have $L = 1.26$ and $p_2 = 0.474$; thus, the cost is

$$\text{Cost}_{K=2} = 100 \times 1.26 + 500 \times 3 \times 0.474 = \$837/wk .$$

If the company allows three in the system, then $L = 1.98$ and $p_3 = 0.415$ which yields

$$\text{Cost}_{K=3} = 100 \times 1.98 + 500 \times 3 \times 0.415 = \$820/wk .$$

If a maximum of four are allowed in the shop, then

$$\text{Cost}_{K=4} = 100 \times 2.76 + 500 \times 3 \times 0.384 = \$852/wk .$$

Therefore, the recommendation is to send a tractor to an outside shop only when more than three are in the system. □

- *Suggestion: Do Problems 7.6 and 7.7.*

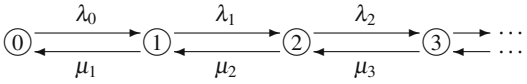

Fig. 7.4 State diagram for a birth-death process

7.3 Multiple Server Queues

A birth-death process is a special type of Markov process which is applicable to many types of Markov queueing systems. The birth-death process is a process in which changes of state are only to adjacent states (Fig. 7.4). The generator matrix for a general birth-death process is given by

$$\mathbf{G} = \begin{bmatrix} -\lambda_0 & \lambda_0 & & & \\ \mu_1 & -(\mu_1+\lambda_1) & \lambda_1 & & \\ & \mu_2 & -(\mu_2+\lambda_2) & \lambda_2 & \\ & & \ddots & \ddots & \ddots \\ & & & & \end{bmatrix},$$

where λ_n and μ_n are the birth rate (arrival rate) and death rate (service rate), respectively, when the process is in state n.

As before, the long-run probabilities are obtained by first forming the system of equations defined by $\mathbf{pG} = \mathbf{0}$. (Or, equivalently, the system of equations may be obtained using the rate-balance approach applied to the state diagram of Fig. 7.4.) The resulting system is

$$\begin{aligned} \lambda_0 p_0 &= \mu_1 p_1 \\ (\lambda_1+\mu_1)p_1 &= \lambda_0 p_0 + \mu_2 p_2 \\ (\lambda_2+\mu_2)p_2 &= \lambda_1 p_1 + \mu_3 p_3 \,. \\ &\vdots \end{aligned}$$

This system is solved in terms of p_0 by successive substitution, and we have

$$\begin{aligned} p_1 &= p_0 \frac{\lambda_0}{\mu_1} \\ p_2 &= p_0 \frac{\lambda_0 \lambda 1}{\mu_1 \mu 2} \\ &\vdots \\ p_n &= p_0 \frac{\lambda_0 \times \cdots \times \lambda_{n-1}}{\mu_1 \times \cdots \times \mu_n} \\ &\vdots \end{aligned} \tag{7.18}$$

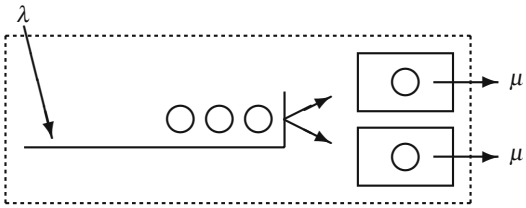

Fig. 7.5 Representation of an M/M/2 queueing system where the transfer from queue to server is instantaneous

where

$$p_0 = \left[1 + \sum_{n=1}^{\infty} \prod_{k=0}^{n-1} \frac{\lambda_k}{\mu_{k+1}}\right]^{-1} .$$

The birth-death process can be used for many types of queueing systems. Equation (7.6), for the M/M/1 system, arises from Eq. (7.18) by letting the birth rates be the constant λ and the death rates be the constant μ. The M/M/1/K system equations can be obtained by letting $\lambda_n = 0$ for all $n > K$.

The M/M/c queueing system is a birth-death process (see Fig. 7.5) with the following values for the birth rates and death rates:

$$\begin{aligned} \lambda_n &= \lambda \quad \text{for } n = 0, 1, \cdots , \\ \mu_n &= \begin{cases} n\mu & \text{for } n = 1, \cdots , c-1 \\ c\mu & \text{for } n = c, c+1, \cdots . \end{cases} \end{aligned} \tag{7.19}$$

The reason that $\mu_n = n\mu$ for $n = 1, \cdots , c-1$ is that when there are less than c customers in the system, each customer in the system is being served and thus the service rate would be equal to the number of customers since unoccupied servers remain idle (i.e., free servers do not help busy servers). If there are more than c customers in the system, then exactly c servers are busy and thus the service rate must be $c\mu$. Substituting Eq. (7.19) into Eqs. (7.18) for the M/M/c system, yields

$$p_n = \begin{cases} p_0 r^n / n! & \text{for } n = 0, 1, \cdots , c-1 \\ p_0 r^n / (c^{n-c} c!) & \text{for } n = c, c+1, \cdots , \end{cases} \tag{7.20}$$

$$p_0 = \left[\frac{c r^c}{c!(c-r)} + \sum_{n=0}^{c-1} \frac{r^n}{n!}\right]^{-1} ,$$

where $r = \lambda / \mu$ and $\rho = r/c < 1$.

The measures of effectiveness for the M/M/c queueing system involve slightly more manipulations, but it can be shown that

$$\begin{aligned} L_q &= \sum_{n=c}^{\infty} (n-c) p_n \\ &= \frac{p_0 r^c \rho}{c!(1-\rho)^2} . \end{aligned}$$

Little's formula is then applied to obtain

$$W_q = \frac{L_q}{\lambda} ,$$

and by Property 7.3

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

and finally applying Little's formula again

$$L = L_q + r .$$

The reason that we let $r = \lambda/\mu$ is because most textbooks and technical papers usually reserve ρ to be the arrival rate divided by the maximum service rate, namely, $\rho = \lambda/(c\mu)$. It can then be shown that ρ gives the server utilization; that is, ρ is the fraction of time that an arbitrarily chosen server is busy.

For completeness, we also give the formula needed to obtain the variances for the M/M/c system as

$$E[(N_q(N_q-1)] = \frac{2p_0 r^c \rho^2}{c!(1-\rho)^3} \tag{7.21}$$

$$E[T_q^2] = \frac{2p_0 r^c}{\mu^2 c^2 c!(1-\rho)^3} \tag{7.22}$$

$$V[T] = V[T_q] + \frac{1}{\mu^2} \tag{7.23}$$

$$V[N] = \lambda^2 V[T] + L . \tag{7.24}$$

Example 7.4. The corporation from the previous example has implemented the policy of never allowing more than three tractors in their repair shop. For \$600 per week, they can hire a second repairman. Is it worthwhile to do so if the expected cost is used as the criterion? To answer this question, the old cost for the M/M/1/3 system (refer back to page 212) is compared to the proposed cost for an M/M/2/3 system. The birth-death equations (7.18) are used with

$$\lambda_n = \begin{cases} \lambda \text{ for } n = 0,1,2 \\ 0 \text{ for } n = 3,4,\cdots \end{cases}$$

$$\mu_n = \begin{cases} \mu \text{ for } n = 1 \\ 2\mu \text{ for } n = 2 \text{ and } 3 , \end{cases}$$

where $\lambda = 3/\text{week}$ and $\mu = 2/\text{week}$. This gives

$$
\begin{aligned}
p_1 &= 1.5p_0 \\
p_2 &= 1.125p_0 \\
p_3 &= 0.84375p_0 \\
p_0 &= \frac{1}{1+1.5+1.125+0.84375} = 0.224 .
\end{aligned}
$$

The expected number in the system is

$$
\begin{aligned}
L &= 0.224 \times (1 \times 1.5 + 2 \times 1.125 + 3 \times 0.84375) \\
&= 1.407 .
\end{aligned}
$$

The cost of the proposed system is

$$
\begin{aligned}
\text{Cost}_{c=2} &= 100 \times 1.407 + 500 \times 3 \times 0.189 + 600 \\
&= \$824.20/\text{week} .
\end{aligned}
$$

Therefore, it is not worthwhile to hire a second man since this cost is greater than the \$820 calculated in the previous example. □

This section is closed with a common example that illustrates the versatility of the birth-death equation. Arrivals to a system usually come from an infinite (or very large) population so that an individual arrival does not affect the overall arrival rate. However, in some circumstances the arrivals come from a finite population, and thus, the arrival rate cannot be assumed constant.

Example 7.5. A small corporation has three old machines that continually breakdown. Each machine breaks down on the average of once a week. The corporation has one repairman that takes, on the average, one half of a week to repair a machine. (See Fig. 7.6 for a schematic of this "machine-repair" queueing system.) Assuming breakdowns and repairs are exponential random variables, the birth-death equations can be used as follows:

$$
\begin{aligned}
\lambda_n &= \begin{cases} (3-n)\lambda \text{ for } n = 0,1,2 \\ 0 \text{ for } n = 3,\cdots \end{cases} \\
\mu_n &= \mu \quad \text{for } n = 1,2,3
\end{aligned}
$$

with $\lambda = 1$ per week and $\mu = 2$ per week. This yields

$$
\begin{aligned}
p_1 &= 1.5p_0 \\
p_2 &= 1.5p_0 \\
p_3 &= 0.75p_0 \\
p_0 &= \frac{1}{1+1.5+1.5+0.75} = 0.21 \text{ and} \\
L &= 0.21 \times (1 \times 1.5 + 2 \times 1.5 + 3 \times 0.75) = 1.42 .
\end{aligned}
$$

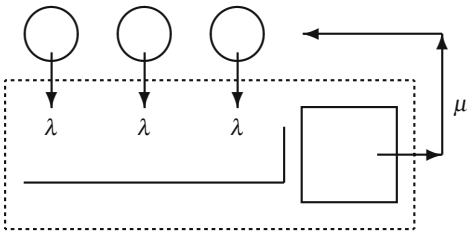

Fig. 7.6 Representation of the machine-repair queueing system of Example 7.5

Let us further assume that for every hour that a machine is tied up in the repair shop, the corporation losses \$25. The cost of this system per hour due to the unavailability of the machines is calculated as

$$\begin{aligned} \text{Cost} &= 25 \times L = 25 \times 1.42 \\ &= \$35.50/\text{ hr}\,. \end{aligned}$$

□

- *Suggestion: Do Problems 7.5 and 7.8–7.14.*

7.4 Approximations

The models developed in the previous sections depended strongly on the use of the exponential distribution. Unfortunately, the exponential assumption is not appropriate for many practical systems for which queueing models are desired. However, without the exponential assumption, exact results are much more difficult. In response to the need for numerical results for nonexponential queueing systems, there has been a significant amount of research dealing with approximations in queueing theory. In this section we report on some of these approximations. We recommend that the interested student read the survey paper by Ward Whitt [6] for a thorough review.

Assume that we are interested in modeling a G/G/c queueing system. An approximation obtained from a diffusion approximation developed by Kingman (see [1, Chap. 3]) uses the squared coefficient of variation; therefore, we define c_a^2 to be the variance of the interarrival times divided by the square of the mean of the interarrival times and c_s^2 to be the variance of the service times divided by the square of the mean of the service times.

To obtain an approximation for the mean waiting time, W, spent in a G/G/c system, we first determine the mean time spent in the queue and then add the mean service time to the queue time; namely, we use the fact that $W = W_q + 1/\mu$. The value for W_q can be obtained according to the following property.

Property 7.4. *Let λ and μ be the mean arrival rate and mean service rate, respectively, for a G/G/c queueing system. In addition, let c_a^2 and c_s^2 be the squared coefficient of variation for the inter-arrival times and service times, respectively. Let $W_{q,M/M/c}$ denote the mean waiting time for a M/M/c queue with the same mean arrival and service rates. When $\lambda < c\mu$, the time spent in the queue for the G/G/c system is approximated by*

$$W_q \approx \left(\frac{c_a^2 + c_s^2}{2} \right) W_{q,M/M/c} \,. \tag{7.25}$$

Notice that the equation holds as an equality if the interarrival and service times are indeed exponential, since the exponential has a squared coefficient of variation equal to one. Equation (7.25) is also exact for an M/G/1 queueing system, and it is known to be an excellent approximation for the M/G/c system. In general, the approximation works best for $c_a \geq 1$ with c_a and c_s being close to the same value.

Example 7.6. Suppose we wish to model the unloading dock at a manufacturing facility. At the dock, there is only one crew who does the unloading. The questions of interest are the average waiting time for arriving trucks and the average number of trucks at the dock at any point in time. Unfortunately, we do not know the underlying probability laws governing the arrivals of trucks or the service times; therefore, data are collected over a representative time period to obtain the necessary statistical estimates. The data yield a mean of 31.3 minutes and a standard deviation of 35.7 minutes for the interarrival times. For the service times, the results of the data give 20.4 minutes for the mean and 10.2 minutes for the standard deviation. Thus, $\lambda = 1.917$ per hour, $c_a^2 = 1.30$, $\mu = 2.941$ per hour, and $c_s^2 = 0.24$ with a traffic intensity of $\rho = 0.652$. (It would be appropriate to collect more data and perform a statistical "goodness-of-fit" test to help determine whether or not these data may come from an exponential distribution. Assuming a reasonable number of data points, it becomes intuitively clear that at least the service times are not exponentially distributed since an exponential distribution has the property that its mean is equal to its standard deviation.)

In order to obtain W, we first determine W_q. From Eq. (7.11), we get that $W_{q,M/M/1} = 0.652/(2.941 - 1.917)$ hr $= 40$ min. Now using Eq. (7.25), the queue time for the trucks is $W_q \approx 40 \times (1.30 + 0.24)/2 = 30.8$ min, yielding a waiting time in the system of

$$W \approx 30.8 + 60/2.941 = 51.2 \text{ min} \,.$$

(Notice that since the mean service rate is 2.941 per hour, the reciprocal is the mean service time which must be multiplied by 60 to convert it to minutes.) It now follows from Little's Law (Property 7.1) that the mean number in the system is $L \approx 1.636$. (Notice, before using Little's formula, care must be taken to insure that the units are consistent, i.e., the quantity W must be expressed in terms of hours, because λ is in terms of hours.) □

A simple adjustment to Eq. 7.25 to account for multiple servers has been proposed by Sakasegawa [5] and used by others (e.g., [2] and [1]) as an extension to Property 7.4.

Property 7.5. *Let λ and μ be the mean arrival rate and mean service rate, respectively, for a G/G/c queueing system. In addition, let c_a^2 and c_s^2 be the squared coefficient of variation for the inter-arrival times and service times, respectively. Then the time spent in the queue for the G/G/c system is approximated by*

$$W_q \approx \left(\frac{c_a^2 + c_s^2}{2} \right) \frac{\rho^{\sqrt{2c+2}-1}}{c(1-\rho)} \frac{1}{\mu} ,$$

for $\rho < 1$ where $\rho = \lambda/(c\mu)$.

- *Suggestion: Do Problem 7.15.*

Appendix

The simulation of a single-server queueing system is relatively easy; however, it is not easily extended to multiple-server systems. For multiple-servers, it is best to use a event driven simulation as discussed in Chap. 9. The following material is taken from Curry and Feldman [1] to illustrate single-server simulations.

Consider a G/G/1 queueing system in which each job is numbered sequentially as it arrives. Let the service time of the n^{th} job be denoted by the random variable S_n, the delay time (time spent in the queue) by the random variable D_n, and the inter-arrival time between the n-1^{st} and n^{th} job by the random variable A_n. The delay time of the n^{th} job must equal the delay time of the previous job, plus the previous job's service time, minus the inter-arrival time; however, if inter-arrival time is larger than the previous job's delay time plus service time, then the queueing delay will be zero. In other words, the following must hold

$$D_n = \max\{ 0,\ D_{n-1} + S_{n-1} - A_n \} . \tag{7.26}$$

Thus, to simulate the G/G/1 system, we need only to generate random variates for A_n and S_n for $n = 1, \cdots, n_{\max}$.

We shall simulate a G/G/1 queueing system with a mean arrival rate of 4 customers per hour and a mean service rate of 5 per hour. The inter-arrival times has a large variation having a squared coefficient of variation equal to $c_a^2 = 4$. The service times are less variable having service times distributed according to an Erlang Type-4 distribution. We shall use the gamma distribution (Eq. 1.18) to model the interarrival times. It has been the authors experience that the gamma distributed random variates with shape parameters less than one are not always reliable; however, they

are fairly easy to simulate. (Can you give a good reason why normal random variates are not used for the inter-arrival times?) Since the Erlang is a special case of the gamma distribution (see p. 19), we shall also use the gamma for services. To begin the simulation, type the following in the first three rows of an Excel spreadsheet. If you do not remember how to generate the gamma random variates, see Table 2.13.)

	A	B	C
1	InterArrive	Service	Delay
2	0	`=GAMMAINV(RAND(),4,3)`	0
3	`=GAMMAINV(RAND(),0.25,60)`	`=GAMMAINV(RAND(),4,3)`	`=MAX(0,C2+B2-A3)`

Notice that the references in the C3 cell are relative references and that two of the references are to the previous row, but the third reference is to the same row. Now copy the third row down for 30,000 rows and obtain an average of the values in the C column. This average is an estimate for the mean cycle time. Hitting the F9 key will give some idea of the variability of this estimate.

- *Suggestion: Do Problem 7.16.*

Problems

7.1. Cars arrive to a toll booth 24 hours per day according to a Poisson process with a mean rate of 15 per hour.
(a) What is the expected number of cars that will arrive to the booth between 1:00 p.m. and 1:30 p.m.?
(b) What is the expected length of time between two consecutively arriving cars?
(c) It is now 1:12 p.m. and a car has just arrived. What is the expected number of cars that will arrive between now and 1:30 p.m.?
(d) It is now 1:12 p.m. and a car has just arrived. What is the probability that two more cars will arrive between now and 1:30 p.m.?
(e) It is now 1:12 p.m. and the last car to arrive came at 1:05 p.m. What is the probability that no additional cars will arrive before 1:30 p.m.?
(f) It is now 1:12 p.m. and the last car to arrive came at 1:05 p.m. What is the expected length of time between the last car to arrive and the next car to arrive?

7.2. A large hotel has placed a single fax machine in an office for customer services. The arrival of customers needing to use the fax follows a Poisson process with a mean rate of eight per hour. The time each person spends using the fax is highly variable and is approximated by an exponential distribution with a mean time of 5 minutes.
(a) What is the probability that the fax office will be empty?
(b) What is the probability that nobody will be waiting to use the fax?
(c) What is the average time that a customer must wait in line to use the fax?
(d) What is the probability that an arriving customer will see two people waiting in line?

7.3. A drill press in a job shop has parts arriving to be drilled according to a Poisson process with mean rate 15 per hour. The average length of time it takes to complete each part is a random variable with an exponential distribution function whose mean is 3 minutes.
(a) What is the probability that the drill press is busy?
(b) What is the average number of parts waiting to be drilled?
(c) What is the probability that at least one part is waiting to be drilled?
(d) What is the average length of time that a part spends in the drill press room?
(e) It costs the company 8 cents for each minute that each part spends in the drilling room? For an additional expenditure of $10 per hour, the company can decrease the average length of time for the drilling operation to 2 minutes. Is the additional expenditure worthwhile?

7.4. Derive results for the M/M/2 system using the methodology developed in Sect. 7.2 (i.e., ignore the general birth-death derivations of Sect. 7.3). Denote the mean arrival rate by λ and the mean service rate for each server by μ.
(a) Give the generator matrix for the system.
(b) Solve the system of equations given by $\mathbf{pG} = \mathbf{0}$ by using successive substitution to obtain $p_n = 2\rho^n p_0$ for $n = 1, 2, \cdots$, and $p_0 = (1-\rho)/(1+\rho)$, where $\rho = \lambda/(2\mu)$.
(c) Show that $L = 2\rho/(1-\rho^2)$ and $L_q = 2\rho^3/(1-\rho^2)$.

7.5. Derive results for the M/M/3 system using the birth-death equations in Sect. 7.3. Denote the mean arrival rate by λ, the mean service rate for each server by μ, and the traffic intensity by $\rho = \lambda/(3\mu)$
(a) Show that $p_1 = 3\rho p_0$, $p_n = 4.5\rho^n p_0$ for $n = 2, 3, \cdots$, and $p_0 = (1-\rho)/(1+2\rho+1.5\rho^2)$.
(b) Show that $L_q = 9\rho^4/((1-\rho)(2+4\rho+3\rho^2))$.

7.6. A small gasoline service station next to an interstate highway is open 24 hours per day and has one pump and room for two other cars. Furthermore, we assume that the conditions for an M/M/1/3 queueing system are satisfied. The mean arrival rate of cars is 8 per hour and the mean service time at the pump is 6 minutes. The expected profit received from each car is $5.00. For an extra $60 per day, the owner of the station can increase the capacity for waiting cars by one (thus, becoming an M/M/1/4 system). Is the extra $60 worthwhile?

7.7. A repair center within a manufacturing plant is open 24 hours a day and there is always one person present. The arrival of items needing to be fixed at the repair center is according to a Poisson process with a mean rate of 6 per day. The length of time it takes for the items to be repaired is highly variable and follows an exponential distribution with a mean time of 5 hours. The current management policy is to allow a maximum of three jobs in the repair center. If three jobs are in the center and a fourth job arrives, then the job is sent to an outside contractor who will return the job 24 hours later. For each day that an item is in the repair center, it costs the company $30. When an item is sent to the outside contractor, it costs the company $30 for the lost time, plus $75 for the repair.
(a) It has been suggested that management change the policy to allow four jobs in

the center; thus jobs would be sent to the outside contractor only when four are present. Is this a better policy?
(b) What would be the optimum cut-off policy? In other words, at what level would it be best to send the overflow jobs to the outside contractor?
(c) In order to staff and maintain the repair center 24-hours per day, it costs $400 per day. Is that a wise economic policy or would it be better to shut down the repair center and use only the outside contractor?
(d) We assume the above questions were answered using a minimum long-run expected cost criterion. Discuss the appropriateness of other considerations besides the long-run expected cost.

7.8. A small computer store has two clerks to help customers (but infinite capacity to hold customers). Customers arrive to the store according to a Poisson process with a mean rate of 5 per hour. Fifty percent of the arrivals want to buy hardware and 50% want to buy software. The current policy of the store is that one clerk is designated to handle only software customers, and one clerk is designated to handle only hardware customers; thus, the store actually acts as two independent M/M/1 systems. Whether the customer wants hardware or software, the time spent with one of the store's clerks is exponentially distributed with a mean of 20 minutes. The owner of the store is considering changing the operating policy of the store and having the clerks help with both software and hardware; thus, there would never be a clerk idle when two or more customers are in the store. The disadvantage is that the clerks would be less efficient since they would have to deal with some things they were unfamiliar with. It is estimated that the change would increase the mean service time to 21 minutes.
(a) If the goal is to minimize the expected waiting time of a customer, which policy is best?
(b) If the goal is to minimize the expected number of customers in the store, which policy is best?

7.9. In a certain manufacturing plant, the final operation is a painting operation. The painting center is always staffed by two workers operating in parallel, although because of the physical setup they cannot help each other. Thus the painting center acts as an M/M/2 system where arrivals occur according to a Poisson process with a mean arrival rate of 100 per day. Each worker takes an average of 27 minutes to paint each item. There has been recent concern about excess work in process so management is considering two alternatives to reduce the average inventory in the painting center. The first alternative is to expand the painting center and hire a third worker. (The assumption is that the third worker, after a training period, will also average 27 minutes per part.) The second alternative is to install a robot that can paint automatically. However, because of the variability of the parts to be painted, the painting time would still be exponentially distributed, but with the robot the mean time would be 10 minutes per part.
(a) Which alternative reduces the inventory the most?
(b) The cost of inventory (including the part that is being worked on) is estimated to be $0.50 per part per hour. The cost per worker (salary and overhead) is estimated to

be $40,000 per year, and the cost of installing and maintaining a robot is estimated to be $100,000 per year. Which alternative, if any, is justifiable using a long-term expected cost criterion?

7.10. In the gasoline service station of Problem 6.6, consider the alternative of adding an extra pump for $90 per day. In other words, is it worthwhile to convert the M/M/1/3 system to an M/M/2/3 system?

7.11. A parking facility for a shopping center is large enough so that we can consider its capacity infinite. Cars arrive to the parking facility according to a Poisson process with a mean arrival rate of λ. Each car stays in the facility an exponentially distributed length of time, independent from all other cars, with a mean time of $1/\mu$. Thus, the parking facility can be viewed as an M/M/∞ queueing system.
(a) What is the probability that the facility contains n cars?
(b) What is the long-run expected number of cars in the facility?
(c) What is the long-run variance of the number of cars in the facility?
(d) What is the long-run expected queue length for the M/M/∞ system?
(e) What is the mean expected time spent in the system?

7.12. A company offers a correspondence course for students not passing high school algebra. People sign up to take the course according to a Poisson process with a mean of two per week. Students taking the course progress at their own rate, independent of how many other students are also taking the correspondence course. The actual length of time that a student remains in the course is an exponential random variable with a mean of 15 weeks. What is the long-run expected number of students in the course at an arbitrary point in time?

7.13. A company has assigned one worker to be responsible for the repair of a group of five machines. The machines break down independent of each other according to an exponential random variable. The mean length of working time for a machine is 4 days. The time it takes the worker to repair a machine is exponentially distributed with a mean of two days.
(a) What is the probability that none of the machines are working?
(b) What is the expected number of machines working?
(c) When a machine fails, what is the expected length of time until it will be working again?

7.14. A power plant operating 24 hours each day has four turbine-generators it uses to generate power. All turbines are identical and are capable of generating 3 megawatts of power. The company needs 6 megawatts of power, so that when all turbines are in a working condition, it keeps one turbine on "warm-standby", one turbine on "cold-standby", and two turbines operating. If one turbine is down, then two are operating and one is on "warm-standby". If two turbines are down, both working turbines are operating. If only one turbine is working, then the company must purchase 3 megawatts of power from another source. And, if all turbines are down, the company must purchase 6 megawatts. If a turbine is in the operating mode, its time until failure is 3 weeks. If a turbine is in "warm-standby", its time

until failure is 9 weeks. And, if a turbine is in "cold-standby", it cannot fail. (We assume all switch-overs from warm standby to working or cold standby to warm standby are instantaneous.) The company has two workers that can serve to repair a failed turbine and it takes a worker one half a week, on the average, to repair a failed turbine. Assuming all times are exponentially distributed, determine the expected megawatt hours that must be purchased each year.

7.15. A store manager with training in queueing theory wants to take quick action on the first day at work. One of the biggest complaints that have been heard is the length of the waiting time and the length of the line. The manager asked one of the employees to record arrival times of customers to the cashier arriving roughly between 8:00 AM and noon. The following arrival times were collected: 8:05, 8:07, 8:17, 8:18, 8:19, 8:25, 8:27, 8:32, 8:35, 8:40, 8:45, 8:47, 8:48, 8:48, 9:00, 9:02, 9:14, 9:15, 9:17, 9:23, 9:27, 9:29, 9:35, 9:37, 9:45, 9:55, 10:01, 10:12, 10:15, 10:30, 10:32, 10:39, 10:47, 10:50, 11:05, 10:07, 11:25, 11:27, 11:31, 11:33, 11:43, 11:49, 12:05.

Another employee measured the service times of the cashier. The service times had a mean of 3.5 minutes and a standard deviation of 5.0 minutes. There is only one cashier that services the customers.

(a) Estimate the expected length of time that a customer has to wait before service and the expected number of customers in front of the cashier using Property 7.4 or 7.5.

(b) Suppose that the manager's knowledge of queueing theory was marginal and that an M/M/1 queueing model was wrongly used assuming that the arrival rate was Poisson with the mean estimated from the data and that the service times were exponential with mean 3.5 minutes. Determine the approximate difference between the estimates obtained in part (a) and the estimates that would be obtained using the (incorrect) Markovian assumptions.

(c) The store manager has two alternatives to reduce the waiting time at the cashier. One alternative is to buy a bar reader machine that would reduce the standard deviation of the service time to 1.7 minutes and pay a monthly service fee of $200. The other alternative is to hire a second cashier to work with a currently available cashier machine. The new cashier would work at the same rate as the other and will cost the store $350 per month. Assuming that it costs $0.25 per minute per customer waiting to pay, the manager wants to know what is the best strategy. (Assume this is a problem only four hours per day, five days a week. Furthermore, assume that the steady-state solution is a good approximation to the queueing behavior that occurs during those four hours.)

7.16. Continue the simulation of the appendix for several years. For each year, calculate the average time spent in the queue per customer and the number of customers who arrive during the year. Give 95% confidence intervals for mean waiting time per customer and the annual arrivals. Each data point in the random sample used for the estimates is composed of the annual data; thus, if the simulation is run for 25 years, there will be 25 data points for each estimate. Compare the simulated results with analytical estimates for W_q and the number of arrival per year.

References

1. Curry, G.L., and Feldman, R.M. (2009). *Manufacturing Systems Modeling and Analysis*, Springer-Verlag, Berlin.
2. Hopp, W.J., and Spearman, M.L. (1996). *Factory Physics: Foundations of Manufacturing Management*, Irwin, Chicago.
3. Kendall, D.G. (1953). Stochastic Processes Occuring in the Theory of Queues and their Analysis by the Method of Imbedded Markov Chains. *Annals of Mathematical Statistics*, **24**:338–354.
4. Little, J.D.C. (1961). A Proof for the Queuing Formula $L = \lambda W$. *Operations Research*, **9**:383–387.
5. Sakasegawa, H. (1977). An Approximation Formula $L_q = \alpha\beta^{\rho}/(1-\rho)$, *Annals of the Institute for Statistical Mathematics*, **29**:67–75.
6. Whitt, W. (1993). Approximations for the GI/G/m Queue, *Production and Operations Management*, **2**:114–161.