SECTION 13.1 INTRODUCTION

The study of *waiting lines*, called **Queuing theory** is one of the oldest and most widely used quantitative analysis techniques. Waiting lines are an everyday occurrence, affecting people shopping for groceries, buying gasoline, making a bank deposit etc. Queues, another term for waiting lines, may also take the form of machines waiting to be repaired, trucks in line to be uploaded, or airplanes lined up on a runway waiting for permission to take off. The three basic components of a queuing process are arrivals, service facilities, and the actual waiting line.

In this chapter, we discuss how analytical models of waiting lines can help managers evaluate the cost and effectiveness of service systems. We begin with a look at waiting line costs and then describe the characteristics of waiting lines and underlying mathematical assumptions used to develop queuing models.

SECTION 13.2 WAITING LINE COSTS

Most **waiting line** problems are focused on the question of finding the ideal level of service a firm should provide. Supermarkets must decide how many cash register checkout positions should be opened. Gasoline stations must decide how many pumps should be opened and how many attendants should be on duty. Banks must decide how many teller windows to keep open to serve customers during various hours of the day. In most cases, this level of service is an option over which management has control.

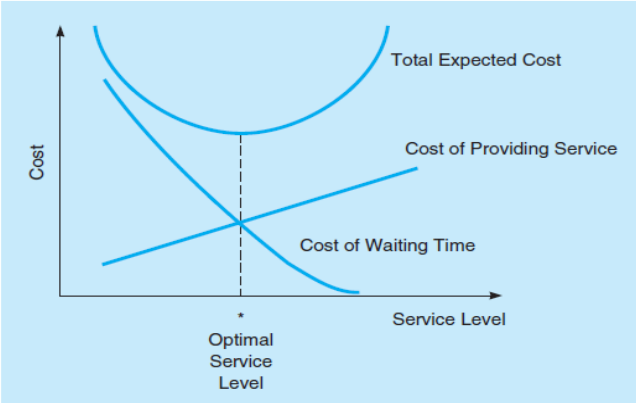
When an organization *does* have control, its objective is usually to find a happy medium between two extremes. On the one hand, a firm can retain a large staff and provide many service facilities. This may result in excellent customer service, with seldom more than one or two customers in a queue. Customers are kept happy with the quick response and appreciate the convenience. This , however can become expensive.

The other extreme is to have the *minimum* possible number of checkout lines, gas pumps, or teller windows open. This keeps the **service cost** down but may result in customer dissatisfaction. As the average length of the queue increases and poor service results, customers and goodwill may be lost.

Most managers recognize the trade-off that must take place between the cost of providing good service and the cost of customer waiting time. They want queues that are short enough so that customers don't become unhappy and either storm out without buying or buy but never return. But they are willing to allow some waiting in line if it is balanced by a significant savings in service costs.

One means of evaluating a service facility is thus to look at a *total expected cost*, a concept illustrated in Figure 13.1.Total expected cost is the sum of expected *service costs* plus expected **waiting costs.**

FIGURE 13.1
Queuing Costs and Service Levels



Service costs are seen to increase as a firm attempts to raise its level of service. For example, if three teams of stevedores, instead of two, are employed to unload a cargo ship, service costs are increased by the additional price of wages. As service improves in speed, however, the cost of time spent waiting in lines decreases. This waiting cost may reflect lost productivity of workers while their tools or machines are awaiting repairs or may simply be an estimate of the costs of customers lost because of poor service and long queues.

Three Rivers Shipping Company Example

As an illustration, let's look at the case of the Three Rivers Shipping Company. Three Rivers runs a huge docking facility located on the Ohio River near Pittsburgh. Approximately five ships arrive to unload their cargoes of steel and ore during every 12-hour work shift. Each hour that a ship sits idle in line waiting to be unloaded costs the firm a great deal of money, about $1,000 per hour. From experience, management estimates that if one team of stevedores is on duty to handle the unloading work, each ship will wait an average of 7 hours to be unloaded. If two teams are working, the average waiting time drops to 4 hours; for three teams, it's 3 hours; and for four teams of stevedores, only 2 hours. But each additional team of stevedores is also an expensive proposition, due to union contracts.

Three Rivers's superintendent would like to determine the optimal number of teams of stevedores to have on duty each shift. The objective is to minimize total expected costs. This analysis is summarized in Table 13.1. To minimize the sum of service costs and waiting costs, the firm makes the decision to employ two teams of stevedores each shift.

TABLE 13.1 Three Rivers Shipping Company Waiting Line Cost Analysis

	NUMBER OF TEAMS OF STEVEDORES WORKING			
	1	2	3	4
(a) Average number of ships arriving per shift	5	5	5	5
(b) Average time each ship waits to be unloaded (hours)	7	4	3	2
(c) Total ship hours lost per shift ($a \times b$)	35	20	15	10
(d) Estimated cost per hour of idle ship time	$1,000	$1,000	$1,000	$1,000
(e) Value of ship's lost time or waiting cost ($c \times d$)	$35,000	$20,000	$15,000	$10,000
(f) Stevedore team salary,* or service cost	$6,000	$12,000	$18,000	$24,000
(g) Total expected cost ($e + f$)	$41,000	$32,000 ← Optimal cost	$33,000	$34,000

*Stevedore team salaries are computed as the number of people in a typical team (assumed to be 50), times the number of hours each person works per day (12 hours), times an hourly salary of $10 per hour. If two teams are employed, the rate is just doubled.

SECTION 13.3 CHARACTERISTICS OF A QUEUING SYSTEM

In this section, we take a look at the three parts of a queuing system:

1) the arrivals or inputs to the system (sometimes referred to as the **calling population**),
2) the queue or the waiting line itself, and
3) the service facility.

These components have their own characteristics that must be examined before mathematical models can be developed.

Arrival Characteristics

The input source that generates arrivals or customers for the service system has three major characteristics. It is important to consider the *size* of the calling population, the *pattern* of arrivals at the queuing system and the *behavior* of the arrivals.

SIZE OF THE CALLING POPULATION: Population sizes are considered to be either **unlimited** (essentially *infinite*) or **limited** (finite). When the number of customers or arrivals on hand at any given moment is just a small portion of potential arrivals, the calling population is considered unlimited. For practical purposes, examples of unlimited populations include cars arriving at a highway tollbooth, shoppers arriving at a supermarket or students arriving to register for classes at a large university. An example of a finite population is a shop with only eight machines that might break down and require service.

PATTERN OF ARRIVALS AT THE SYSTEM: Customers either arrive at a service facility according to some known schedule (for example, one patient every 15 minutes or one student for advising every half hour) or else they arrive *randomly*. Arrivals are considered random when they are independent of one another and their occurrence cannot be predicted exactly. Frequently in queuing problems, the no. of arrivals per unit of time can be estimated by a probability distribution known as **Poisson distribution.**

BEHAVIOR OF THE ARRIVALS: Most queuing models assume that an arriving customer is a patient customer. Patient customers are people or machines that wait in the queue until they are served and don not switch between lines. Unfortunately, life and quantitative analysis are complicated by the fact that people have been known to balk or renege. **Balking** refers to customers who refuse to join the waiting line because it is too long to suit their needs or interests. **Reneging** customers are those who enter the queue but then become impatient and leave without completing their transaction.

Waiting Line Characteristics

The waiting line itself is the second component of a queuing system. The length of a line can be either *limited* or *unlimited*. A queue is limited when it cannot (by law of physical restrictions) increase to an infinite length. This may be the case in a small restaurent that has only 10 tables and can serve no more than 50 diners an evening. Analytic queuing models are treated in this chapter under an assumption of **unlimited queue length**.

A second waiting line characteristic deals with **queue discipline**. This refers to the rule by which customers in the line are to receive service. Most systems use a queue discipline known as the **first-in, first-out (FIFO)** rule. In a hospital emergency room or an express checkout line at a supermarket, however, various assigned priorities may preempt FIFO. Patients who are critically injured will move ahead in traetment priority over patients with broken fingers or noses.Shoppers with fewer than 10 items may be allowed to enter the express checkout queue but are then treated as first come, first served.

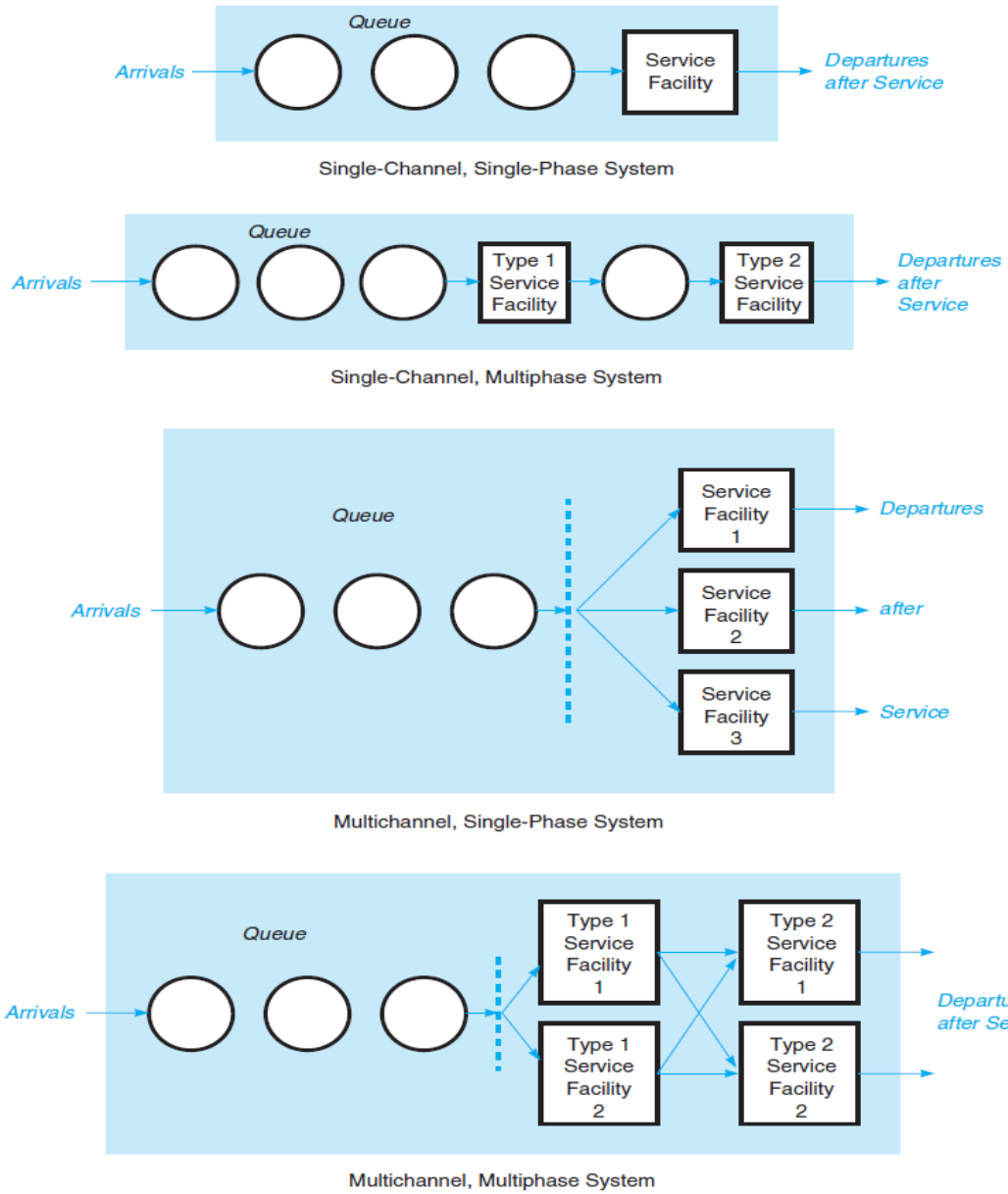
Service Facility Characteristics

The third part of any queuing system is the service facility. It is important to examine two basic properties: (1) the configuration of the service system and (2) the pattern of service times.

BASIC QUEUING SYSTEM CONFIGURATIONS: Service systems are usually classified in terms of their number of channels or number of servers and number of phases or number of service stops, that must be made. A **single-channel system,** with one server, is typified by the drive-in bank that has only one open teller or by the type of drive-through fast-food restaurant that has become so popular in the United States. If, on the other hand, the bank had several tellers on duty and each customer waited in one common line for the first available teller, we should have a **multi-channel system** at work.

A **single-phase system** is one in which the customer receives service from only one station and then exits the system. A fast-food restaurant in which the person who takes your order also brings you the food and takes your money in a single –phase system. But if the restaurant requires you to place your order at one station, pay at a second and pick up the food at a third service stop, it becomes a **multiphase system**. To help you relate the concepts of channels and phases, Figure 13.2 presents four possible configurations.

FIGURE 13.2 Four Basic Queuing System Configurations



SERVICE TIME DISTRIBUTION:Service patterns are like arrival patterns in that they can be either constant or random. If service time is constant, it takes the same amount of time to take care of each customer. This is the case in a machine-performed service operation such as an automatic car wash. More often, service times are randomly distributed. In many cases it can be assumed that random service times are described by the **negative exponential probability distribution**.

Identifying Models Using Kendall Notation:

D.G. Kendall developed a notation that has been widely accepted for specifying the pattern of arrivals, the service time distribution, and the number of channels in a queuing model. This notation is often seen in software for queuing models. The basic three-symbol Kendall notation is in the form

Arrival distribution /Service time distribution/Number of service channels open

Where specific letters are used to represent probability distributions. The following letters are commonly used in Kendall notation:

M = Poisson distribution for number of occurrences (or exponential times)

D = contant (deterministic) rate

G = general distribution with mean and variance known

Thus, a single channel model with Poisson arrivals and exponential service times would be represented by $M / M / 1$

When a second channel is added, we would have $M / M / 2$

If there are m distinct service channels in the queuing system with Poisson arrivals and exponential service times, the Kendall notation would be $M / M / m$. A three-channel system with Poisson arrivals and constant service time would be identified as $M / D / 3$. A four-channel system with Poisson arrivals and service times that are normally distributed would be identified as $M / G / 4$.

There is more detailed notation with additional terms that indicate the maximun number in the system and the population size. When these are omitted, it is assumed there is no limit to the queue length or the population size. Most of the models we study here will have those properties.

SECTION 13.4 SINGLE-CHANNEL QUEUING MODEL WITH POISSON ARRIVALS AND EXPONENTIAL SERVICE TIMES (M/M/1)

In this section, we present an analytical approach to determine important measures of performance in a typical service system. After these numeric measures have been computed, it will be possible to add in cost data and begin to make decisions that balance desirable service levels with waiting line service costs.

Assumptions of the Model

The single-channel, single-phase model considered here is one of the most widely used and simplest queuing models. It involves assuming that seven conditions exist:

1. Arrivals are served on a FIFO basis.
2. Every arrival waits to be served regardless of the length of the line; that is, there is no balking or reneging.
3. Arrivals are independent of preceding arrivals, but the average number of arrivals (the arrival rate) does not change over time.
4. Arrivals are described by a Poisson probability distribution and come from an infinite or very large population.
5. Service times also vary from one customer to the next and are independent of one another, but their average rate is known.
6. Service times occur according to the negative exponential probability distribution.
7. The average service rate is greater than the average arrival rate.

When these seven conditions are met, we can develop a series of equations that define the queue's **operating characteristics**.

Queuing Equations

We let

λ = mean no. of arrivals per time period (for example, per hour)

μ = mean no. of people or items served per time period

When determining the arrival rate (λ) and the service rate (μ), the same time period must be used. For example, if λ is the average no. of arrivals per hour, then μ must indicate the average no. that could be served per hour.

The queuing equations follow:

1. The average no. of customers or units in the system, L, that is, the no. in line plus the no. being served:

$$L = \frac{\lambda}{\mu - \lambda} \qquad (13.1)$$

2. The average time a customer spends in the system, W, that is, the time spent in line plus the time spent being served:

$$W = \frac{1}{\mu - \lambda} \qquad (13.2)$$

3. The average no. of customers in the queue, L_q:

$$L_q = \rho L = \frac{\lambda^2}{\mu(\mu - \lambda)} \qquad (13.3)$$

4. The average time a customer spends waiting in the queue, W_q:

$$W_q = \rho W = \frac{\lambda}{\mu(\mu - \lambda)} \qquad (13.4)$$

5. The **utilization factor** for the system, ρ, that is, the probability that the service facility is being used:

$$\rho = \frac{\lambda}{\mu} \qquad (13.5)$$

6. The percent idle time, P_0, that is, the probability that no one is in the system:

$$P_0 = 1 - \frac{\lambda}{\mu} \qquad (13.6)$$

7. The probability that the no. of customers in the system is greater than k, $P_{n>k}$:

$$P_{n>k} = \left(\frac{\lambda}{\mu}\right)^{k+1} \qquad (13.7)$$

Example: From historical data, Harry's Car Wash estimates that dirty cars arrive at the rate of 10 per hour all day Saturday. With a crew working the wash line, Harry figures that cars can be cleaned at the rate of one every 5 minutes. One car at a time is cleaned in this example of a single-channel waiting line.

Assuming Poisson arrivals and exponential service times, find the
(a) Utilization rate of the car wash.
(b) Average time a car waits before it is washed.
(c) Average time a car spends in the service system.

Solution:

Here Arrival rate, $\lambda = 10$ cars per hour

Service rate, $\mu = \frac{60}{5} = 12$ cars per hour

(a) Utilization rate of car wash, $\rho = \frac{\lambda}{\mu} = \frac{10}{12} = 0.8333$ or 83.33%

(b) Average time a car waits before washed,

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{10}{12(12 - 10)} = 0.4167\ hrs$$

(c) Average time a car spends in the service system,

$$W = \frac{1}{\mu - \lambda} = \frac{1}{12 - 10} = 0.5\ hrs$$

Example: The students patiently form a single line in front of the desk to wait for help at University. Student arrivals are best described by Poisson distribution with mean of 15 students per hour arriving at the help desk. The help desk server can help an average one student in 3 minutes, with the service rate being described by an exponential distribution. Calculate the following characteristics of the service system:

(a) The average number of students in the system.
(b) The average number of students waiting in the line.
(c) Probability that no student is in the system.

Solution:

Here Arrival rate, $\lambda = 15$ students per hour

Service rate, $\mu = \frac{60}{3} = 20$ students per hour

(a) The average number of students in the system,

$$L = \frac{\lambda}{\mu - \lambda} = \frac{15}{20 - 15} = 3\ students$$

(b) The average number of students waiting in the line,

$$L_q = \rho L = \frac{\lambda}{\mu} L = \frac{15}{20}(3) = 2.25\ students$$

(c) Probability that no student is in the system,

$$P(0) = 1 - \frac{\lambda}{\mu} = 1 - \frac{15}{20} = 0.25$$

SECTION 13.5 MULTICHANNEL QUEUING MODEL WITH POISSON ARRIVALS AND EXPONENTIAL SERVICE TIMES (M/M/m)

The next logical step is to look at a *multichannel queuing system*, in which two or more servers or channels are available to handle arriving customers. Let us still assume that customers awaiting service form one single line and then proceed to the first available server. An example of such a multichannel, single-phase waiting line is found in many banks today. A common line is formed and the customer at the head of the line proceeds to the first free teller (Refer to Figure 13.2 for a typical multichannel configuration.)

The multiple-channel system presented here again assumes that arrivals follow a Poisson probability distribution and that service times are distributed exponentially. Service is first come, first served, and all servers are assumed to perform at the same rate. Other assumptions listed earlier for the single-channel model apply as well.

Equations for the Multichannel Queuing Model

If we let

$$\begin{aligned} m &= \text{number of channels open,} \\ \lambda &= \text{average arrival rate, and} \\ \mu &= \text{average service rate at each channel} \end{aligned}$$

the following formulas may be used in the waiting line analysis:

1. The probability that there are zero customers or units in the system:

$$P_0 = \frac{1}{\left[\sum_{n=0}^{n=m-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n\right] + \frac{1}{m!}\left(\frac{\lambda}{\mu}\right)^m \frac{m\mu}{m\mu - \lambda}} \quad \text{for } m\mu > \lambda \tag{13-13}$$

2. The average number of customers or units in the system:

$$L = \frac{\lambda\mu(\lambda/\mu)^m}{(m-1)!(m\mu - \lambda)^2} P_0 + \frac{\lambda}{\mu} \tag{13-14}$$

3. The average time a unit spends in the waiting line or being serviced (namely, in the system):

$$W = \frac{\mu(\lambda/\mu)^m}{(m-1)!(m\mu - \lambda)^2} P_0 + \frac{1}{\mu} = \frac{L}{\lambda} \tag{13-15}$$

4. The average number of customers or units in line waiting for service:

$$L_q = L - \frac{\lambda}{\mu} \tag{13-16}$$

5. The average time a customer or unit spends in the queue waiting for service:

$$W_q = W - \frac{1}{\mu} = \frac{L_q}{\lambda} \tag{13-17}$$

6. Utilization rate:

$$\rho = \frac{\lambda}{m\mu} \tag{13-18}$$

These equations are obviously more complex than the ones used in the single-channel model, yet they are used in exactly the same fashion and provide the same type of information as did the simpler model.

Arnold's Muffler Shop Revisited

For an application of the multichannel queuing model, let's return to the case of Arnold's Muffler Shop. Earlier, Larry Arnold examined two options. He could retain his current mechanic, Reid Blank, at a total expected cost of \$653 per day; or he could fire Blank and hire a slightly more expensive but faster worker named Jimmy Smith. With Smith on board, service system costs could be reduced to \$360 per day.

A third option is now explored. Arnold finds that at minimal after-tax cost he can open a *second* garage bay in which mufflers can be installed. Instead of firing his first mechanic, Blank, he would hire a second worker. The new mechanic would be expected to install mufflers at the same rate as Blank—about $\mu = 3$ per hour. Customers, who would still arrive at the rate of $\lambda = 2$ per hour, would wait in a single line until one of the two mechanics is free. To find out how this option compares with the old single-channel waiting line system, Arnold computes several operating characteristics for the $m = 2$ channel system:

$$P_0 = \frac{1}{\left[\sum_{n=0}^{1}\frac{1}{n!}\left(\frac{2}{3}\right)^n\right] + \frac{1}{2!}\left(\frac{2}{3}\right)^2\left(\frac{2(3)}{2(3)-2}\right)}$$

$$= \frac{1}{1 + \frac{2}{3} + \frac{1}{2}\left(\frac{4}{9}\right)\left(\frac{6}{6-2}\right)} = \frac{1}{1 + \frac{2}{3} + \frac{1}{3}} = \frac{1}{2} = 0.5$$

= probability of 0 cars in the system

$$L = \left(\frac{(2)(3)\left(\frac{2}{3}\right)^2}{1![2(3)-2]^2}\right)\left(\frac{1}{2}\right) + \frac{2}{3} = \frac{\frac{8}{3}}{16}\left(\frac{1}{2}\right) + \frac{2}{3} = \frac{3}{4} = 0.75$$

= average number of cars in the system

$$W = \frac{L}{\lambda} = \frac{\frac{3}{4}}{2} = \frac{3}{8} \text{ hours} = 22\tfrac{1}{2} \text{ minutes}$$

= average time a car spends in the system

$$L_q = L - \frac{\lambda}{\mu} = \frac{3}{4} - \frac{2}{3} = \frac{1}{12} = 0.083$$

= average number of cars in the queue

$$W_q = \frac{L_q}{\lambda} = \frac{0.083}{2} = 0.0415 \text{ hour} = 2\tfrac{1}{2} \text{ minutes}$$

= average time a car spends in the queue

These data are compared with earlier operating characteristics in Table 13.2. The increased service from opening a second channel has a dramatic effect on almost all characteristics. In particular, time spent waiting in line drops from 40 minutes with one mechanic (Blank) or 15 minutes with Smith down to only $2\frac{1}{2}$ minutes! Similarly, the average number of cars in the queue falls to 0.083 (about $\frac{1}{12}$ of a car).* But does this mean that a second bay should be opened?

SECTION 13.6 CONSTANT SERVICE TIME MODEL(M/D/1)

Some service systems have constant service times instead of exponentially distributed times. When customers or equipment are processed according to a fixed cycle, as in the case of an automatic car wash or an amusement park ride, constant service rates are appropriate. Because constant rates are certain, the values for L_q, W_q, L, and W are always less than they would be in the models we have just discussed, which have variable service times. As a matter of fact, both the average queue length and the average waiting time in the queue are *halved* with the constant service rate model.

Equations for the Constant Service Time Model

Constant service model formulas follow:

1. Average length of the queue:

$$L_q = \frac{\lambda^2}{2\mu(\mu - \lambda)} \tag{13-19}$$

2. Average waiting time in the queue:

$$W_q = \frac{\lambda}{2\mu(\mu - \lambda)} \tag{13-20}$$

3. Average number of customers in the system:

$$L = L_q + \frac{\lambda}{\mu} \tag{13-21}$$

4. Average time in the system:

$$W = W_q + \frac{1}{\mu} \tag{13-22}$$

Garcia-Golding Recycling, Inc.

Garcia-Golding Recycling, Inc., collects and compacts aluminum cans and glass bottles in New York City. Its truck drivers, who arrive to unload these materials for recycling, currently wait an average of 15 minutes before emptying their loads. The cost of the driver and truck time wasted while in queue is valued at $60 per hour. A new automated compactor can be purchased that will process truck loads at a constant rate of 12 trucks per hour (i.e., 5 minutes per truck). Trucks arrive according to a Poisson distribution at an average rate of 8 per hour. If the new compactor is put in use, its cost will be amortized at a rate of $3 per truck unloaded. A summer intern from a local college did the following analysis to evaluate the costs versus benefits of the purchase:

$$\begin{aligned}
\textit{Current}\text{ waiting cost/trip} &= \left(\tfrac{1}{4}\text{ hour waiting now}\right)(\$60/\text{hour cost}) \\
&= \$15/\text{trip} \\
\textit{New}\text{ system: } \lambda &= 8\text{ trucks/hour arriving,} \\
\mu &= 12\text{ trucks/hour served} \\
\text{Average waiting time in queue} &= W_q = \frac{\lambda}{2\mu(\mu - \lambda)} = \frac{8}{2(12)(12 - 8)} \\
&= \frac{1}{12}\text{ hour} \\
\text{Waiting cost/trip with new compactor} &= \left(\tfrac{1}{12}\text{ hour wait}\right)(\$60/\text{hour cost}) = \$5/\text{trip} \\
\text{Savings with new equipment} &= \$15\text{ (current system)} - \$5\text{ (new system)} \\
&= \$10/\text{trip} \\
\text{Cost of new equipment amortized} &= \underline{\$3/\text{trip}} \\
\text{Net savings} &= \$7/\text{trip}
\end{aligned}$$

SECTION 13.7 FINITE POPULATION MODEL (M/M/1 WITH FINITE SOURCE)

When there is a limited population of potential customers for a service facility, we need to consider a different queuing model. This model would be used, for example, if you were considering equipment repairs in a factory that has five machines, if you were in charge of maintenance for a fleet of 10 commuter airplanes, or if you ran a hospital ward that has 20 beds. The limited population model permits any number of repair people (servers) to be considered.

The reason this model differs from the three earlier queuing models is that there is now a *dependent* relationship between the length of the queue and the arrival rate. To illustrate the extreme situation, if your factory had five machines and all were broken and awaiting repair, the arrival rate would drop to zero. In general, as the waiting line becomes longer in the limited population model, the arrival rate of customers or machines drops lower.

In this section, we describe a finite calling population model that has the following assumptions:

1. There is only one server.
2. The population of units seeking service is finite.*
3. Arrivals follow a Poisson distribution, and service times are exponentially distributed.
4. Customers are served on a first-come, first-served basis.

Equations for the Finite Population Model

Using

$$\lambda = \text{mean arrival rate}, \mu = \text{mean service rate}, N = \text{size of the population}$$

the operating characteristics for the finite population model with a single channel or server on duty are as follows:

1. Probability that the system is empty:

$$P_0 = \frac{1}{\sum_{n=0}^{N} \frac{N!}{(N-n)!}\left(\frac{\lambda}{\mu}\right)^n} \tag{13-23}$$

2. Average length of the queue:

$$L_q = N - \left(\frac{\lambda + \mu}{\lambda}\right)(1 - P_0) \tag{13-24}$$

3. Average number of customers (units) in the system:

$$L = L_q + (1 - P_0) \tag{13-25}$$

4. Average waiting time in the queue:

$$W_q = \frac{L_q}{(N - L)\lambda} \tag{13-26}$$

5. Average time in the system:

$$W = W_q + \frac{1}{\mu} \tag{13-27}$$

6. Probability of n units in the system:

$$P_n = \frac{N!}{(N-n)!}\left(\frac{\lambda}{\mu}\right)^n P_0 \quad \text{for } n = 0, 1, \ldots, N \tag{13-28}$$

Department of Commerce Example

Past records indicate that each of the five high-speed "page" printers at the U.S. Department of Commerce, in Washington, D.C., needs repair after about 20 hours of use. Breakdowns have been determined to be Poisson distributed. The one technician on duty can service a printer in an average of 2 hours, following an exponential distribution.

To compute the system's operation characteristics we first note that the mean arrival rate is $\lambda = {}^{1}/_{20} = 0.05$ printer/hour. The mean service rate is $\mu = {}^{1}/_{2} = 0.50$ printer/hour. Then

1. $P_0 = \dfrac{1}{\sum_{n=0}^{5} \dfrac{5!}{(5-n)!}\left(\dfrac{0.05}{0.5}\right)^n} = 0.564$ (we leave these calculations for you to confirm)

2. $L_q = 5 - \left(\dfrac{0.05 + 0.5}{0.05}\right)(1 - P_0) = 5 - (11)(1 - 0.564) = 5 - 4.8$
   $= 0.2$ printer

3. $L = 0.2 + (1 - 0.564) = 0.64$ printer

4. $W_q = \dfrac{0.2}{(5 - 0.64)(0.05)} = \dfrac{0.2}{0.22} = 0.91$ hour

5. $W = 0.91 + \dfrac{1}{0.50} = 2.91$ hours

If printer downtime costs \$120 per hour and the technician is paid \$25 per hour, we can also compute the total cost per hour:

$$\begin{aligned}\text{Total hourly cost} &= (\text{Average number of printers down})(\text{Cost per downtime hour}) \\ &\quad + \text{Cost per technician hour} \\ &= (0.64)(\$120) + \$25 = \$76.80 + \$25.00 = \$101.80\end{aligned}$$

SECTION 7.1 INTRODUCTION

Many management decisions involve trying to make the most effective use of an organization's resources. Resources typically include machinery, labor, money, time, warehouse space and row materials. These resources may be used to make products (such as machinery, furniture, food or clothing) or services (such as schedules for airlines or production, advertising policies or investment decisions). Linear programming (LP) is a widely used mathematical modeling technique designed to help managers in planning and decision making relative to resource allocation.

Despite its name, LP and the more general category of techniques called "mathematical" programming have very little to do with computer programming. In the world of management science, programming refers to modeling and solving a problem mathematically. Computer programming has, of course, played an important role in the advancement and use of L.P. Real life LP problems are too cumbersome to solve by hand or with a calculator.

SECTION 7.2 REQUIREMENTS OF A LINEAR PROGRAMMING PROBLEM

In the past 60 years, LP has been applied extensively to military, industrial, financial, marketing, accounting and agricultural problems. Even though these applications are diverse, all LP problems have several properties and assumptions in common.

All problems seek to maximize or minimize some quantity, usually profit or cost. We refer to this property as the **Objective Function** of an LP problem.

The second property that LP problems have in common is the presence of restrictions, or **constraints**, that limit the degree to which we can pursue our objective. For example, deciding how many units of each product in a firm's product line to manufacture is restricted by available personnel and machinery. Selection of an advertising policy or a financial portfolio is limited by the amount of money available to be spent or invested. We want, therefore, to maximize or minimize a quantity (the objective function) subject to limited resources (the constraints).

There must be alternative courses of action to choose from. For example, if a company produces 3 different products, management may use LP to decide how to allocate among them its limited production resources (of personnel, machinery and so on). Should it devote all manufacturing capacity to make only the first product, should it produce equal amounts of each product, or should it allocate the resources in some other ratio? If there were no alternatives to select from, we would not need L.P.

The objective and constraints in LP problems must be expressed in terms of linear equations or inequalities. Linear mathematical relationships just mean that all terms used in the objective function and constraints are of the first degree (i.e., not squared, or to the third or higher power or appearing more than once).

The term **linear** implies both proportionality and additivity. Proportionality means that if production of 1 unit of a product uses 3 hours, production of 10 units would use 30 hours. Additivity means that the total of all activities equals the sum of the individual activities. If the production of one product generated $3 profit and the production of another product generated $8 profit, the total profit would be the sum of these two, which would be $11.

We assume that conditions of **certainty** exit: that is, number in the objective and constraints are known with certainty and do not change during the period being studied.

We make the **divisibility** assumption that solutions need not to be in whole numbers (integers). Instead, they are divisible and may take any fractional value. In production problems, we often define variables as the number of units produced per week or per month, and a fractional value (i.e., 0.3 chairs) would simply mean that there is work in process. Something that was started in one week can be finished in the next. However, in other types of problems, fractional values do not make sense. If a fraction of a product cannot be purchased (for example, one-third of a submarine), an integer programming problem exists.

Finally, we assume that all answers or variables are **nonnegative.** Negative values of physical quantities are impossible; you simply cannot produce a negative number of chairs, shirts, lamps or computers. Table 7.1 summarizes these properties and assumptions.

TABLE 7.1
LP Properties and Assumptions

PROPERTIES OF LINEAR PROGRAMS
1. One objective function
2. One or more constraints
3. Alternative courses of action
4. Objective function and constraints are linear—proportionality and divisibility
5. Certainty
6. Divisibility
7. Nonnegative variables

SECTION 7.3 FORMULATING LP PROBLEMS

Formulating a linear program involves developing a mathematical model to represent the managerial problem. Thus, in order to formulate a linear program, it is necessary to completely understand the managerial problem being faced. The steps in formulating a linear program follow:

(1) Completely understand the managerial problem being faced.
(2) Identify the objective and the constraints.
(3) Define the decision variables.
(4) Use the decision variables to write mathematical expressions for the objective function and the constraints.

One of the most common LP applications is the Product Mix Problem. Two or more products are usually produced using limited resources such as personnel, machines, raw materials, and so on. The profit that the firm seeks to maximize is based on the profit contribution per unit of each product. The company would like to determine how many units of each product it should produce so as to maximize overall profit given its limited resources.

Example : Flair Furniture Company

The Flair Furniture Company produces inexpensive tables and chairs. The production process for each is similar in that both require a certain number of hours of carpentry work and a certain number of labor hours in the painting and varnishing department. Each table takes 4 hours of carpentry and 2 hours in the painting and varnishing shop. Each chair requires 3 hours in carpentry and 1 hour in painting and varnishing. During the current production period, 240 hours of carpentry time are available and 100 hours in painting and varnishing time are available. Each table sold yields a profit of $70; each chair produced is sold for $50 profit.

Flair Furniture's problem is to determine the best possible combination of tables and chairs to manufacture in order to reach the maximum profit. The firm would like this production mix situation formulated as an LP problem.

We begin by summarizing the information needed to formulate and solve this problem (see Table 7.2)

TABLE 7.2 Flair Furniture Company Data

	HOURS REQUIRED TO PRODUCE 1 UNIT		
DEPARTMENT	**TABLES (*T*)**	**CHAIRS (*C*)**	**AVAILABLE HOURS THIS WEEK**
Carpentry	**4**	**3**	**240**
Painting & varnishing	**2**	**1**	**100**
Profit per unit	**$70**	**$50**	

Formulation:

The decision variables that represent the actual decisions we will make are defined as:

T = number of tables to be produced per week.

C = number of chairs to be produced per week.

Now we can create the LP objective function in terms of T and C:

Maximize profit = $\$70T + \$50C$

Our next step is to develop mathematical relationships for the two constraints:

For carpentry, total time used is:

(4hours per table)(Number of tables produced)

+ (3 hours per chair)(Number of chairs produced).

So the first constraint may be stated as follows:

Carpentry time used ≤ Carpentry time available.

$4T + 3C \leq 240$ (hours of carpentry time)

Similarly, the second constraint is as follows:

Painting and varnishing time used ≤ Painting and varnishing time available.

$2T + 1C \leq 100$ (hours of painting and varnishing time)

Both of these constraints represent production capacity restrictions and, of course, affect the total profit.

To obtain meaningful solutions, the values for T and C must be nonnegative numbers. That is, all potential solutions must represent real tables and real chairs. Mathematically, it means that

$T \geq 0$ (number of tables produced is greater than or equal to 0)

$C \geq 0$ (number of chairs produced is greater than or equal to 0)

The complete problem may now be restated mathematically as

Maximize profit $= \$70T + \$50C$

Subjects to the constraints

$4T + 3C \leq 240$ (carpentry constraint)

$2T + 1C \leq 100$ (painting and varnishing constraint)

$T, C \geq 0$ (nonnegativity constraints)

SECTION 7.4 GRAPHICAL SOLUTION TO AN LP PROBLEM

The easiest way to solve a small LP problem such as that of the Flair Furniture Company is with the graphical solution approach. The graphical procedure is useful only when there are two decision variables (such as no. of tables, T and no. of Chairs, C) in the problem. When there are more than two variables, it is not possible to plot the solution on a two-dimensional graph and we must turn to more complex approaches.

GENERAL LINEAR PROGRAMMING PROBLEM IN TWO VARIABLES:

Find the values of x_1 & x_2 that optimize (either maximize or minimize)

$$z = c_1x_1 + c_2x_2 \quad \text{[Linear Objective Function]}$$

Subject to Linear Constraints $a_{11}x_1 + a_{12}x_2\ (\leq, \geq \ or\ =) b_1$

$$a_{21}x_1 + a_{22}x_2\ (\leq, \geq \ or\ =) b_2$$

....

$$a_{m1}x_1 + a_{m2}x_2\ (\leq, \geq \ or\ =) b_m$$

And $x_1 \geq 0,\ x_2 \geq 0$ [Non-Negative Constraints]

NOTE (1) A pair of values (x_1, x_2) that satisfy all the constraints is called a **Feasible Solution**. The set of all feasible solutions determines a subset of x_1x_2-plane called the feasible region. A feasible solution that optimizes the objective function is called an **Optimal Solution**.

NOTE (2) The feasible region of an LPP has a boundary consisting of a finite number of straight line segments. If the feasible region can be enclosed in a sufficiently large circle, it is called **Bounded**; otherwise it is called Unbounded.

If the feasible region is empty (contains no points), then the constraints are **Inconsistent** and the LPP has no solution.

Those boundary points of a feasible region that are intersections of two of the straight line boundary segments are called **Extreme points (or Corner points)**.

THEOREM: If the feasible region of an LPP is non-empty and bounded, then the objective function attains both a maximum and a minimum value and these occur at extreme points of the feasible region. If the feasible region is Unbounded, then the objective function may or may not attain a maximum or minimum value; however, if it attains a maximum or minimum value, it does so at an extreme point.

Example: Solve the following LPP by Graphical method-

Maximize profit $= \$70T + \$50C$

Subjects to the constraints

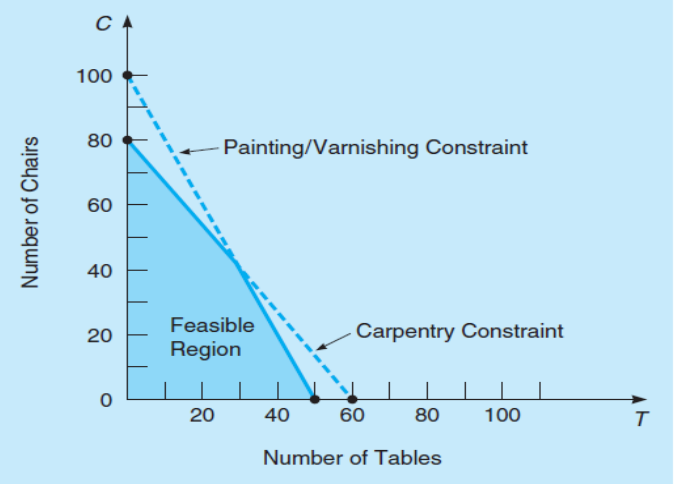
$4T + 3C \leq 240$ (carpentry constraint)

$2T + 1C \leq 100$ (painting and varnishing constraint)

$T, C \geq 0$ (nonnegativity constraints)

Solution: In Fig, we have drawn the feasible region of this problem.

FIGURE 7.5
Feasible Solution Region for the Flair Furniture Company Problem



Since the feasible region is bounded, the maximum value of z is attained at one of the extreme points. For this example, the coordinates of three of the corner points are obvious from observing the graph. These are (0, 0), (50, 0) and (0, 80). The fourth corner point is where the two constraint lines intersect and the coordinates must be found algebraically by solving the two equations simultaneously for two variables.

Therefore solving the equations $4T + 3C = 240$

$2T + C = 100$

We get $T = 30$ and $C = 40$ so the intersection point is (30, 40).

The values of objective function at four extreme points are given in the following table:

Extreme Points (T, C)	(0, 0)	(50, 0)	(30, 40)	(0,80)
$z = 70T + 50C$	0	3500	4100	4000

From the Table, the maximum value of z is 4100 which is attained at $T = 30$ & $C = 40$.