*Waiting lines*

Waiting in queues (or waiting lines) is one of the most common occurrences in everyone's life. Anyone who has gone shopping or to a movie has had the inconvenience of waiting in line to make purchases or buy a ticket. Not only do people spend a significant portion of their time waiting in lines but product queues also form in production plants, machinery waits in line to be serviced, etc. Since time is a valuable resource, the reduction of waiting time is an important topic of analysis.

*A probabilistic form of analysis*

Like decision analysis and Markov analysis, queuing analysis is a probabilistic form of analysis, and not a deterministic technique. As such, the results of queuing analysis, referred to as "operating characteristics," are probabilistic. These operating statistics (such as the average time a person in line must wait to be served) are subsequently used by the manager of the operation containing the queue to make decisions.

A number of different queuing models exist to deal with different queuing systems. While we will eventually discuss many of these queuing variations, we will concentrate on two of the most common types of systems—the single-server system and the multiple-server system.

## The Single-Server Waiting Line System

The single server with a single waiting line is the simplest form of queuing system. As such, we will use it to demonstrate the basic fundamentals of a queuing system. As an example of this kind of system, we will consider the Fast Shop Drive-In Market.

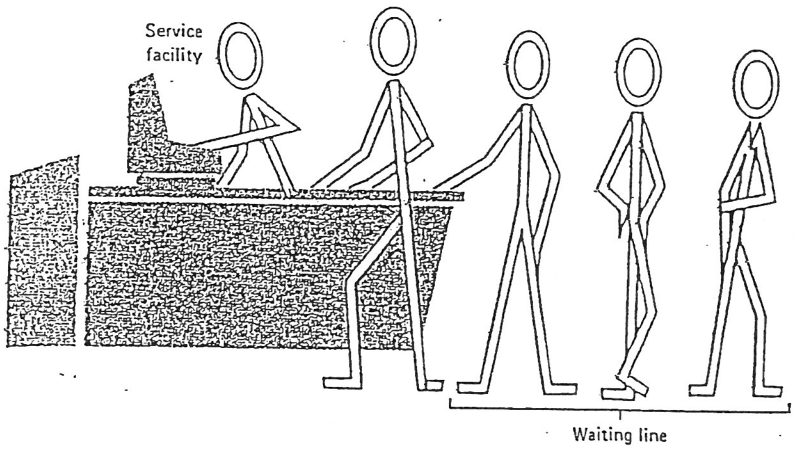
*A queuing example*

The Fast Shop Market consists of one checkout counter and one employee who operates the cash register at the checkout counter. The combination of the cash register and operator is the *server* (or service facility) in this queuing system, while the customers who line up at the counter to pay for their selections form the *waiting line*, or *queue*. The configuration of this example queuing system is shown in figure 14.1.

*Server*

*Queue*

Figure 14.1 The Fast Shop Market queuing system.




The most important factors in analyzing a queuing system such as the one in figure 14.1 are

1. The queue discipline (in what order customers are served)
2. The nature of the calling population (where customers come from)
3. The arrival rate (how often customers arrive at the queue)
4. The service rate (how fast customers are served)

We will discuss each of these items as it relates to our example.

## The Queue Discipline

*The order in which customers are served*

The *queue discipline* is the order in which waiting customers are served. Customers at the Fast Shop Market are served on a "first-come, first-served" basis. That is, the first person in line at the checkout counter is served first. This is the most common type of queue discipline. However, other disciplines are possible. For example, a machine operator might stack in-process parts beside a machine so that the last part is on top of the stack and will be selected first. This queue discipline would be referred to as "last-in, first-out." Alternatively, the machine operator might simply reach into a box full of parts and select one at *random*. In this case the queue discipline is random. Often customers are scheduled for service according to a predetermined appointment, such as patients at a doctor's or dentist's office, or diners at a restaurant where reservations are required. In this case the customers are taken according to a prearranged schedule regardless of when they arrive at the facility. One final example of the many types of queue disciplines that can occur is when a number of customers are processed alphabetically according to their last name, such as occurs at school registration or at job interviews.

## The Calling Population

*Source of arrivals to the queuing system*

The *calling population* is the source of the customers to the market, which is assumed in this case to be infinite. In other words, there is such a large number of possible customers in the area where the store is located, which could come into the market, that the potential customers are assumed to be infinite. Alternatively, on occasion queuing systems have "finite" calling populations. For example, the repair garage of a trucking firm, which has 20 trucks, has a finite calling population. The queue is the number of trucks waiting to be repaired and the finite calling population is 20 trucks. However, queuing systems that have an assumed infinite calling population are more frequent.

## The Arrival Rate

*The number of arrivals during a time period*

The *arrival rate* is the average number of arrivals at the service facility during a specified period of time. For example, if 100 customers arrive at the store checkout counter during a 10-hour day, we could say the arrival rate averages 10 customers per hour. However, although we may be able to determine a rate for arrivals by counting the number of paying customers at the market during a 10-hour day, we would not know exactly when these customers would arrive on the premises. In other words, no customers might arrive during one hour, yet 20 customers might arrive during another hour. In general, these arrivals are assumed to be independent of each other and vary randomly over time.

*Arrivals conform to a Poisson probability distribution*

Given these assumptions, it is further assumed that arrivals at a service facility conform to some probability distribution. Although arrivals could be described by any distribution, it has been determined (through years of research and the practical experience of people in the field of queuing) that the number of arrivals per unit of time at a service facility can frequently be defined by a *Poisson distribution*. (Appendix C at the end of this text contains a more detailed presentation of the Poisson distribution.)

## The Service Rate

*The average number served during a time period*

The *service rate* is the average number of customers that can be served during a specified period of time. For our Fast Shop Market example, 30 customers can be checked out (served) in one hour. A service rate is similar to an arrival rate in that it is a random variable. In other words, such factors as different sizes of customer purchases, the amount of change the cashier must determine, and different forms of payment alter the number of persons that can be served over time. While only 10 customers might be checked out during one hour, 40 customers might be checked out during the following hour.

*Service defined by an exponential probability distribution*

The description of arrivals in terms of a *rate* and service in terms of *time* is a convention that has developed in the literature of queuing theory. As in the case of an arrival rate, the service time is also assumed to be defined by a probability distribution. It has been determined by researchers in the field of queuing that service times can frequently be defined by an *exponential probability* distribution. (Appendix C at the end of this text contains a more detailed presentation of the exponential distribution.) However, in order to analyze a queuing system, both arrivals and service must be in compatible units of measure. Thus, service time must be expressed as a service rate to correspond with an arrival rate.

## The Single-Server Model

*Characteristics of the single-server queuing system*

The Fast Shop Market checkout counter is an example of a single-server queuing system, which consists of the following characteristics.

1. An infinite calling population
2. A "first-come, first-served" queue discipline
3. Poisson arrival rate
4. Exponential service times

These assumptions have been used to develop a model of a single-server queuing system. However, the analytical derivation of even this simplest queuing model is relatively complex and lengthy. As such, we will refrain from deriving the model in detail and will present only the resulting queuing formulas. The reader must keep in mind, however, that these formulas are applicable only to queuing systems having the above conditions.

Given that

$\lambda$ = the arrival rate (average number of arrivals per time period)
$\mu$ = the service rate (average number served per time period)

*The service rate exceeds the arrival rate*

and that $\lambda < \mu$ (customers are served at a faster rate than they arrive), the following formulas for the operating characteristics of a single-server model exist.

*The single-server model queuing formulas*

The probability that no customers are in the queuing system (either in the queue or being served),

$$P_0 = \left(1 - \frac{\lambda}{\mu}\right)$$

The probability of $n$ customers in the queuing system,

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

The average number of customers in the queuing system (i.e., the customers being serviced and in the waiting line),

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

The average number of customers in the waiting line,

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

The average time a customer spends in the total queuing system (i.e., waiting and being served),

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

The average time a customer spends waiting in the queue to be served,

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)}$$

The probability that the server is busy (i.e., the probability that a customer has to wait), known as the *utilization factor,*

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

The probability that the server is idle (i.e., the probability that a customer can be served),

$$I = 1 - U$$
$$= 1 - \frac{\lambda}{\mu}$$

This last term, $1 - (\lambda/\mu)$, is also equal to $P_0$. The probability of no customers in the queuing system is the same as the probability that the server is idle.

*Computing the operating statistics for the queuing example*

These various operating characteristics can be computed for the Fast Shop Market by simply substituting the average arrival and service rates into the above formulas. For example, if

$\lambda = 24$ customers per hour arrive at checkout counter
$\mu = 30$ customers per hour can be checked out

then

$$P_0 = \left(1 - \frac{\lambda}{\mu}\right)$$

$$= (1 - 24/30)$$

$= .20$ probability of 0 customers in the system

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

$$= \frac{24}{30 - 24}$$

$= 4$ customers on the average in the queuing system

$$L_q = \frac{\lambda^2}{\mu\,(\mu - \lambda)}$$

$$= \frac{(24)^2}{30\,(30 - 24)}$$

$= 3.2$ customers on the average in the waiting line

$$W = \frac{1}{\mu - \lambda}$$

$$= \frac{1}{30 - 24}$$

$= .167$ hours (10 minutes) average time in the system per customer

$$W_q = \frac{\lambda}{\mu\,(\mu - \lambda)}$$

$$= \frac{24}{30\,(30 - 24)}$$

$= .133$ hours (8 minutes) average time in the waiting line per customer

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

$$= \frac{24}{30}$$

$= .80$ probability that the server will be busy and the customer must wait

$$I = 1 - U$$

$$= 1 - .80$$

$= .20$ probability that the server will be idle and the customer can be served

Several important items concerning both the general model and this particular example will now be discussed in greater detail.

*Steady-state results*

First, recall that the operating characteristics are "averages." Also, they are assumed to be *steady-state* averages. In our discussion of Markov analysis (see chap. 13), it was indicated that the steady state was a constant average level that a system realized after a period of time. For a queuing system, the steady state is represented by the average operating statistics, also determined over a period of time.

*The utilization factor must be less than one*

Related to this condition is the fact that the utilization factor, $U$, must be less than 1.0

$$U < 1$$

or

$$\frac{\lambda}{\mu} < 1.0$$

and

$$\lambda < \mu$$

In other words, the ratio of the arrival rate to the service rate must be less than one, which also means *the service rate must be greater than the arrival rate* if this model is to be used. The server must be able to serve customers faster than they come into the store, or the waiting line will grow to an infinite size and the system will never reach a steady state.

## The Effect of Operating Characteristics on Managerial Decisions

We must now reflect on the operating characteristics as they relate to management decisions. The arrival rate of 24 customers per hour means that, on the average, a customer arrives every 2.5 minutes (i.e., $1/24 \times 60$ minutes). This indicates that the store is very busy. This is because the customer purchases few items and expects quick service, since that is the nature of the store. Customers expect to spend a relatively large amount of time in a supermarket, since typically they make larger purchases. Alternatively, customers who trade at a drive-in market do so, at least partially, because it is quicker than a supermarket.

*Alternatives for reducing customer waiting time*

Given this condition, the store's manager feels that if a customer waits 8 minutes and spends a total of 10 minutes in the queuing system (not including the actual shopping time) this amount of time is excessive. As such, the manager wants to test several alternatives for reducing customer waiting time: (1) the addition of another employee to "sack" the purchases and (2) the addition of an additional checkout counter.

### Alternative I: The Addition of an Employee

The addition of an extra employee will cost the store manager \$150 per week. With the help of the national office's marketing research group, the manager has determined that for each minute of customer waiting time that can be eliminated, a loss in sales of \$75 per week will be avoided (i.e., customers who leave prior to shopping because of the long line, or customers who do not return).

*Increasing the service rate*

If a new employee is hired, customers can be served within a shorter period of time. In other words, the *service rate,* which is the number of customers served per time period, will *increase.* The previous service rate was

$$\mu = 30 \text{ customers served/hour}$$

The addition of a new employee will increase the service rate to

$$\mu = 40 \text{ customers served/hour}$$

It will be assumed that the arrival rate will remain the same ($\lambda = 24$ per hour), since the effect of the increased service rate will not increase arrivals but instead will minimize the loss of customers. (However, it is not illogical to also assume that an increase in service might increase arrivals.)

*Recomputing the operating statistics*

Given the "$\lambda$" and "$\mu$" values, the operating characteristics can be recomputed as follows.

$$P_0 = \left(1 - \frac{\lambda}{\mu}\right)$$

$$= \left(1 - \frac{24}{40}\right)$$

$$= .40 \text{ probability of 0 customers in the system}$$

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

$$= \frac{24}{40 - 24}$$

$$= 1.5 \text{ customers in the queuing system on the average}$$

$$L_q = \frac{\lambda^2}{\mu\,(\mu - \lambda)}$$

$$= \frac{(24)^2}{40\,(16)}$$

$$= .90 \text{ customers in the waiting line on the average}$$

$$W = \frac{1}{\mu - \lambda}$$

$$= \frac{1}{40 - 24}$$

$$= .063 \text{ hours (3.75 minutes) average time per customer in the system}$$

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)}$$

$$= \frac{24}{40\,(16)}$$

$= .038$ hours (2.25 minutes) average time in the waiting line per customer

$$U = \frac{\lambda}{\mu} = \frac{24}{40}$$

$= .60$ probability that the customer must wait

$$I = 1 - U$$

$= 1 - .60 = .40$ probability that the server will be idle

It must be remembered that these operating characteristics are *averages* that result over a period of time; they are not absolutes. In other words, customers who arrive at the Fast Shop Market checkout counter will not find .90 customers in line. There could be no customers or 1, 2, or 3 customers, for example. The value .90 is simply an average that occurs over time, as do the other operating characteristics.

*Deriving the savings resulting from the increased service rate*

The average waiting time per customer has been reduced from 8 minutes to 2.25 minutes, a significant amount. The savings (due to the aversion of lost sales) is computed as follows.

8.00 minutes − 2.25 minutes = 5.75 minutes
5.75 minutes × \$75/minute = \$431.25

Since the extra employee costs management \$150 per week, the total savings will be

\$431.25 − \$150 = \$281.25 per week

The store manager would probably consider this savings and the above operating statistics to be preferable to the previous condition where only one employee existed.

### Alternative II: The Addition of a New Checkout Counter

Next we will consider the manager's alternative of constructing a new checkout counter. The total cost of this project will be \$6,000 plus an extra \$200 per week for an additional cashier.

The new checkout counter will be opposite the present counter (such that the servers will have their backs to each other in an enclosed counter area). There will be several display cases and racks between the two lines so that customers waiting in line will not move back and forth between the lines. (Such movement, called *jockeying*, would not allow us to use the queuing formulas we already developed.) We will assume that the customers will divide themselves equally between both lines so that the arrival rate for each line will be half of the prior arrival rate for a single checkout counter. Thus, the new arrival rate for each checkout counter is

*Reducing the arrival rate*

$$\lambda = 12 \text{ customers per hour}$$

and, the service rate will remain the same for each of the counters,

$$\mu = 30 \text{ customers served per hour}$$

Substituting this new arrival rate and the service rate into our queuing formulas results in the following operating characteristics.

*The recomputed operating statistics*

$P_0$ = .60 probability of "0" customers in the system
$L$ = .67 customers in the queuing system
$L_q$ = .27 customers in the waiting line
$W$ = .055 hours (3.33 minutes) per customer in the system
$W_q$ = .022 hours (1.33 minutes) in the waiting line per customer
$U$ = .40 probability that a customer must wait
$I$ = .60 probability that the server will be idle

Using the same sales savings of \$75 per week per minute reduction in waiting time, the store will save

*Deriving the savings resulting from the decreased arrival rate*

8.00 minutes − 1.33 minutes = 6.67 minutes
6.67 minutes × \$75/minute = \$500.00 per week

First we must subtract the \$200 per week cost of the new cashier from this amount saved.

\$500 − 200 = \$300

Since the capital outlay of this project is \$6,000, it will take 20 weeks (\$6,000/\$300 = 20 weeks) to recoup the initial cost (ignoring the possibility of interest on the \$6,000). Once the cost has been recovered, the store would make \$18.75 (\$300.00 − 281.25) more by adding a new checkout counter rather than simply hiring an extra employee. However, we must not disregard the fact that during the 20-week cost recovery period, the \$281.25 savings incurred by simply hiring a new employee would be lost.

Table 14.1 presents a summary of the operating characteristics for each alternative.

*Summarizing the operating characteristics for each alternative*

Table 14.1 Operating Characteristics for Each Alternative System

| Operating Characteristics | Present System | Alternative I | Alternative II |
|---|---|---|---|
| $L$ | 4.00 customers | 1.50 customers | .67 customers |
| $L_q$ | 3.20 customers | .90 customers | .27 customers |
| $W$ | 10.00 minutes | 3.75 minutes | 3.33 minutes |
| $W_q$ | 8.00 minutes | 2.25 minutes | 1.33 minutes |
| $U$ | .80 | .60 | .40 |

*Decision making using the operating statistics*

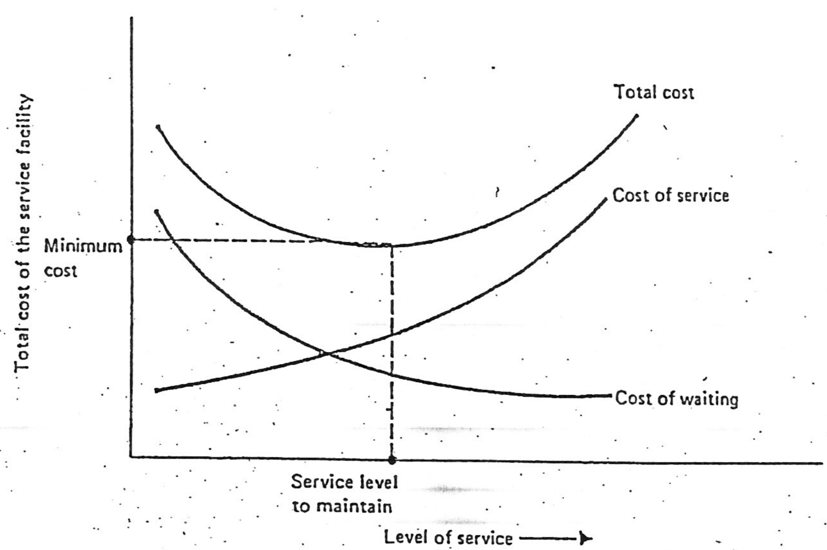
For the store manager both of these alternatives seem preferable to the original conditions, which resulted in a lengthy waiting time of 8 minutes per customer. However, the manager might have a more difficult time selecting between the two alternatives. It might be appropriate to consider other factors besides waiting time. For example, employee idle time is .40 with the first alternative and .60 with the second, which seems to be a significant difference. An additional factor is the "loss of space" resulting from a new checkout counter.

However, the final decision must be based on the manager's own experience and perceived needs. As we have noted previously, the results of queuing analysis provide information for decision making, but do not result in an actual recommended decision as an optimization model would.

*The cost trade-off related to improved service*

Our two example alternatives illustrate the cost trade-off situation that is related to improved service. As the level of service increases, the corresponding cost of this service also increases. For example, when we added an extra employee in alternative I, the service was improved but the cost of providing service also increased. Alternatively, when the level of service was increased, the costs associated with customer waiting decreased. Maintaining an appropriate level of service should minimize the sum of these two costs as much as possible. This cost trade-off relationship is summarized in figure 14.2. As the level of service increases, the cost of service goes up while the waiting cost goes down. The sum of these costs results in a total cost curve, and the level of service that should be maintained is where this total cost curve is at a minimum. (However, this does not mean we can determine an exact optimal minimum cost solution, since the service and waiting characteristics we can determine are averages, and thus uncertain.)

Figure 14.2 Cost trade-offs for service levels.



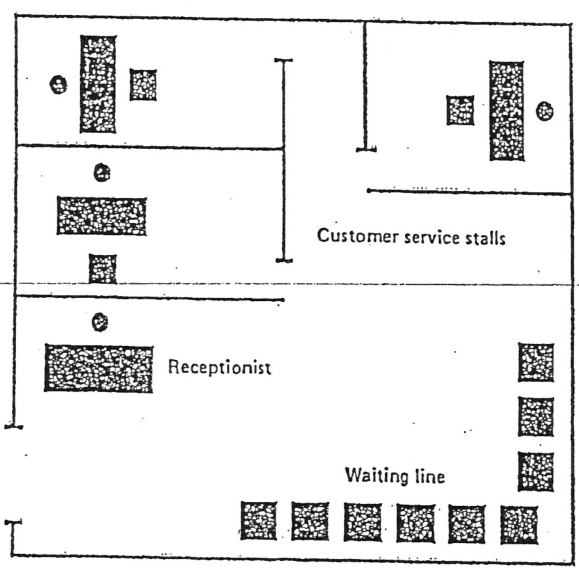
## The Multiple-Server Waiting Line

A slightly more complex queuing system than the single-server system is the case of a single waiting line being serviced by more than one server (i.e., multiple servers). As an example of this type of system, we will consider the customer service department of the Biggs Department Store.

*A multiple-server system example*

The customer service department of the store has a waiting room consisting of chairs placed along the wall of the area which, in effect, forms a single waiting line. Customers come to this area with questions or complaints or to clarify matters regarding credit card bills. The customers are served by three store representatives, each located in a partitioned stall. Customers are treated on a "first-come, first-served" basis. Figure 14.3 presents a schematic of this queuing system.

Figure 14.3 Customer service queuing system.




The store management desires to analyze this queuing system because excessive waiting times can result in irate customers, who will then often trade at other stores. Typically, customers who come to this area have some problem and thus are impatient anyway. Waiting a long time only serves to increase their impatience.

*The characteristics of the multiple-server system*

First we will present the queuing formulas for a multiple-server queuing system. These formulas, like single-server model formulas, have been developed on the assumption of a *first-come, first-served queue discipline, Poisson arrivals, exponential service times,* and *an infinite calling*

Saaty, T. L. *Elements of Queueing Theory*. New York: McGraw-Hill, 1961.

Shamblin, J. E., and Stevens, G. T., Jr. *Operations Research: A Fundamental Approach*. New York: McGraw-Hill, 1974.

Taha, H. A. *Operations Research: An Introduction*. New York: Macmillan Co., 1971.

Vogel, M. "Queueing Theory Applied to Machine Manning." *Interfaces* 9, no. 4 (August 1979): 1–7.


## Problems

1. Identify ten examples of queuing systems that you are familiar with in real life.

2. Define the following components of a queuing system:
   a. Queue discipline
   b. Calling population
   c. Arrival rate
   d. Service rate

3. A single-server queuing system with an infinite calling population and a first-come, first-served queue discipline has an arrival rate of 50 customers per hour and a service rate of 70 customers per hour. Determine the following.
   a. The probability of no customers in the queuing system.
   b. The average number of customers in the queuing system.
   c. The average number of customers in the waiting line.
   d. The average time a customer is in the queuing system.
   e. The probability that the server is busy.

4. A single-server queuing system with an infinite calling population and a first-come, first-served queue discipline has the following arrival and service rates:

   $\lambda = 16$ customers per hour
   $\mu = 24$ customers per hour

   Determine $P_0$, $P_3$, $L$, $L_q$, $W$, $W_q$, and $U$.

5. The ticket booth on the Tech campus is operated by one person, who is selling tickets for the annual Tech vs. State football game on Saturday. Each hour, the ticket seller can serve an average of 12 customers and an average of 10 customers arrive to purchase tickets. Determine the average time a ticket buyer must wait and the portion of time the ticket seller is busy.

6. The Petroco Service Station has one pump for unleaded gas, which (with an attendant) can service 10 customers per hour. Cars arrive at the unleaded pump at a rate of 6 per hour. Determine the average queue length, the average time in the system for a car, and the average time a car must wait. If during a gasoline shortage the arrival rate increases to 12 cars per hour, what effect will it have on the average queue length?

7. The Dynaco Manufacturing Company produces a particular product in an assembly line operation. One of the machines on the line is a drill press that has a single assembly line feeding into it. A partially completed unit arrives at the press to be worked on every 7.5 minutes, on the average. The machine operator averages processing 10 parts per hour. Determine the average number of parts waiting to be worked on, the percentage of time the operator is working, and the percentage of time the machine is idle.

8. The management of Dynaco Manufacturing Company (problem 7) likes to have their operators working 90% of the time. What must the assembly line arrival rate be in order for the operator to be as busy as management would like?

9. The Peachtree Airport in Atlanta serves light aircraft. It has a single runway and one air traffic controller to land planes. It takes an airplane 12 minutes to land and clear the runway. Planes arrive at the airport at the rate of 4 per hour. Determine the following:
   a. The number of planes that will stack up (on the average) waiting to land.
   b. The average time a plane must wait in line before it can land.
   c. The average time it takes a plane to clear the runway once it notifies the airport it is in the vicinity and wants to land.
   d. The FAA has a rule that an air traffic controller can on the average land planes a maximum of 45 minutes out of every hour. (There must be 15 minutes idle time available to relieve the tension.) Will this airport have to hire an extra air traffic controller?

10. The First American Bank of Rapid City presently has one outside drive-up teller. It takes the teller an average of 4 minutes to serve a bank customer. Customers arrive at the drive-up window at the rate of 12 per hour. The bank operations officer is currently analyzing the possibility of adding a second drive-up window at an annual cost of $20,000. It is assumed that arriving cars would be equally divided between both windows. The operations officer estimates that each minute reduction in customer waiting time will increase the bank's revenue by $2,000 annually. Should the second drive-up window be installed?