Waiting Line Models

SUPPLEMENT C

Before studying this supplement you should know or, if necessary, review

1. Competitive advantages, Chapter 2.
2. Priority rules, Chapter 16.

LEARNING OBJECTIVES

After completing this supplement you should be able to

1. Describe the elements of a waiting line problem.
2. Use waiting line models to estimate system performance.
3. Use waiting line models to make managerial decisions.

SUPPLEMENT OUTLINE

WHAT'S IN OM FOR ME?

ACC FIN MKT OM HRM MIS

Waiting in lines is part of everyday life. Some estimates state that Americans spend 37 billion hours per year waiting in lines. Whether it is waiting in line at a grocery store to buy deli items (by taking a number) or checking out at the cash registers (finding the quickest line), waiting in line at the bank for a teller, or waiting at an amusement park to go on the newest ride, we spend a lot of time waiting. We wait in lines at the movies, campus dining rooms, the registrar's office for class registration, at the Division of Motor Vehicles, and even at the end of the school term to sell books back. Think about the lines you have waited in just during the past week. How long you wait in line depends on a number of factors. Your wait is a result of the number of people served before you, the number of servers working, and the amount of time it takes to serve each individual customer.

Wait time is affected by the design of the waiting line system. A **waiting line system** (or **queuing system**) is defined by two elements: the population source of its customers and the process or service system itself. In this supplement we examine the elements of waiting line systems and appropriate performance measures. Performance characteristics are calculated for different waiting line systems. We conclude with descriptions of managerial decisions related to waiting line system design and performance.

▶ **Waiting line system** Includes the customer population source as well as the process or service system.

▶ **Queuing system** Another name to define a waiting line.

ELEMENTS OF WAITING LINES

Any time there is more customer demand for a service than can be provided, a waiting line occurs. Customers can be either humans or inanimate objects. Examples of objects that must wait in lines include a machine waiting for repair, a customer order waiting to be processed, subassemblies in a manufacturing plant (that is, work-in-process inventory), electronic messages on the Internet, and ships or railcars waiting for unloading.

In a waiting line system, managers must decide what level of service to offer. A low level of service may be inexpensive, at least in the short run, but may incur high costs of customer dissatisfaction, such as lost future business and actual processing costs of complaints. A high level of service will cost more to provide and will result in lower dissatisfaction costs. Because of this trade-off, management must consider what is the optimal level of service to provide. This is illustrated in Figure C-1.

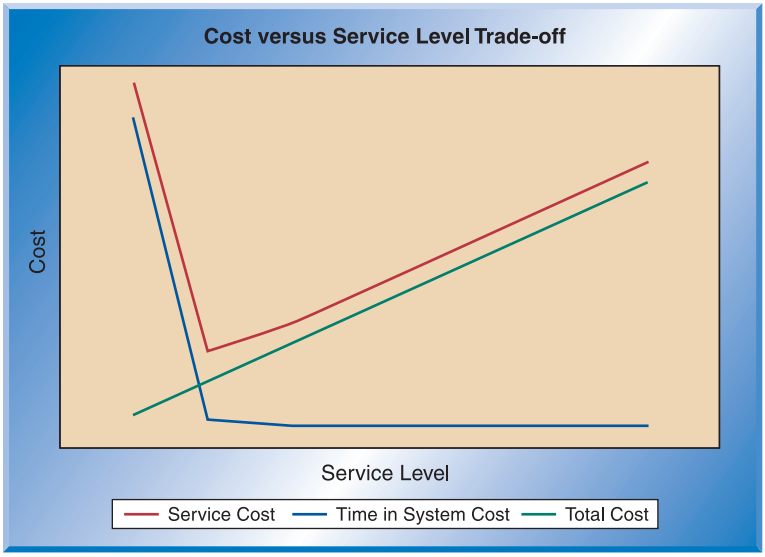


FIGURE C-1

Waiting cost and service level trade-off

LINKS TO PRACTICE

Waiting for Fast Food

Kristin Sladen/The Image Works

Fast-food restaurants illustrate the transient nature of waiting line systems. Waiting lines occur at a fast-food restaurant drive-through during peak meal times each day. There is a temporary surge in demand that cannot be quickly handled with the available capacity. In an effort to speed up delivery, some restaurants use an extra window—the first window for paying and the second window for picking up the food. At other times of the day, the restaurant uses a single window and may have no waiting line at the drive-through window.

The challenge is designing service systems with adequate but not excessive amounts of capacity. A fast-food restaurant experiences variable demand and variable service times. The restaurant cannot be sure how much customer demand there will be, and it does not know exactly what each customer will order—each order can be unique and require a different service time. It is important to understand the different elements of a waiting line system. These elements include the customer population source, the service system, the arrival and service patterns, and the priorities used for controlling the line. Let's first look at the primary input into the waiting line system: the customers.

The Customer Population

The customer population can be considered to be finite or infinite. When potential new customers for the waiting line system are affected by the number of customers already in the system, the customer population is **finite**. For example, if you are in a

▶ **Finite customer population**
The number of potential new customers is affected by the number of customers already in the system.

class with nine other students, the total customer population for meeting with the professor during office hours is ten students. As the number of students waiting to meet with the professor increases, the population of possible new customers decreases. There is a finite limit as to how large the waiting line can ever be.

When the number of customers waiting in line does not significantly affect the rate at which the population generates new customers, the customer population is considered **infinite**. For example, if you are taking a class with 500 other students (a relatively large population) and the probability of all the students trying to meet with the professor at the same time is very low, then the number of students in line does not significantly affect the population's ability to generate new customers.

▶ **Infinite customer population**
The number of potential new customers is not affected by the number of customers already in the system.

In addition to waiting, a customer has other possible actions. For example, a customer may balk, renege, or jockey. **Balking** occurs when the customer decides not to enter the waiting line. For example, you see that there are already 12 students waiting to meet with your professor, so you choose to come back later. **Reneging** occurs when the customer enters the waiting line but leaves before being serviced. For example, you enter the line waiting to meet with your professor, but after waiting 15 minutes and seeing little progress, you decide to leave. **Jockeying** occurs when a customer changes from one line to another, hoping to reduce the waiting time. A good example of this is picking a line at the grocery store and changing to another line in the hope of being served quicker.

▶ **Balking**
The customer decides not to enter the waiting line.

▶ **Reneging**
The customer enters the line but decides to exit before being served.

▶ **Jockeying**
The customer enters one line and then switches to a different line in an effort to reduce the waiting time.

The models used in this supplement assume that customers are patient; they do not balk, renege, or jockey; and the customers come from an infinite population. The mathematical formulas become more complex for systems in which customer population must be considered finite and when customers balk, renege, or jockey.

The Service System

The service system is characterized by the number of waiting lines, the number of servers, the arrangement of the servers, the arrival and service patterns, and the service priority rules.

The Number of Waiting Lines Waiting line systems can have single or multiple lines. Banks often have a single line for customers. Customers wait in line until a teller is free and then proceed to that teller's position. Other examples of single-line systems include airline counters, rental car counters, restaurants, amusement park attractions, and call centers. The advantage of using a single line when multiple servers are available is the customer's perception of fairness in terms of equitable waits. That is, the customer is not penalized by picking the slow line but is served in a true first-come, first-served fashion. The single-line approach eliminates jockeying behavior. Finally, a single-line, multiple-server system has better performance in terms of waiting times than the same system with a line for each server.

The multiple-line configuration is appropriate when specialized servers are used or when space considerations make a single line inconvenient. For example, in a grocery store some registers are express lanes for customers with a small number of items. Using express lines reduces the waiting time for customers making smaller purchases. Examples of single- and multiple-line systems are shown in Figure C-2.

The Number of Servers System serving capacity is a function of the number of service facilities and server proficiency. In waiting line systems, the terms *server* and *channel* are used interchangeably. It is assumed that a server or channel can serve one

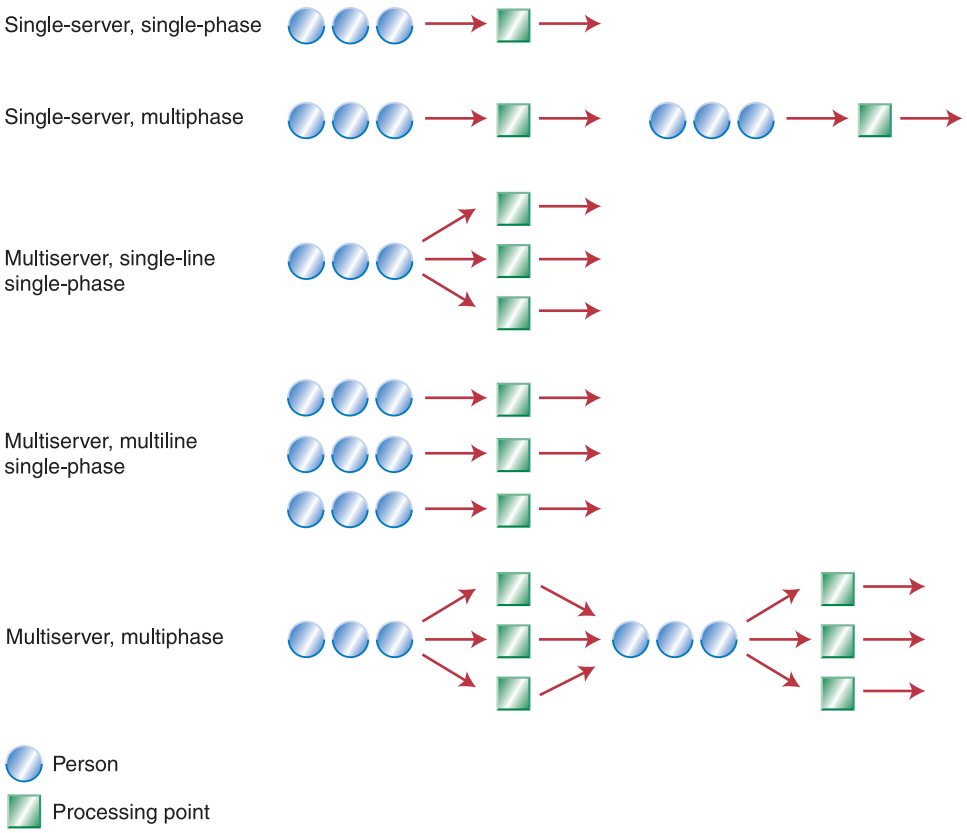


FIGURE C-2

Examples of waiting line systems

customer at a time. Waiting line systems are either single server (single channel) or multiserver (multichannel). Single-server examples include small retail stores with a single checkout counter, a theater with a single person selling tickets and controlling admission into the show, or a ballroom with a single person controlling admission. Multiserver systems have parallel service providers offering the same service. Multiserver examples include grocery stores (multiple cashiers), drive-through banks (multiple drive-through windows), and gas stations (multiple gas pumps).

The Arrangement of the Servers Services require a single activity or a series of activities and are identified by the term *phase*. Refer to Figure C-2. In a single-phase system, the service is completed all at once, such as with a bank transaction or a grocery store checkout. In a multiphase system, the service is completed in a series of steps, such as at a fast-food restaurant with ordering, pay, and pick-up windows or in many manufacturing processes.

In addition, some waiting line systems have a finite size of the waiting line. Sometimes this happens in multiphase systems. For example, perhaps only two cars can physically fit between the ordering and pay window of a fast-food drive-through. Finite size limitations can also occur in single-phase systems and can be associated either with the physical system (for example, a call center has only a finite number of incoming phone lines) or with customer behavior (if a customer arrives when a certain number of people are already waiting, the customer chooses to not join the line).

Arrival and Service Patterns

▶ **Arrival rate**
The average number of customers arriving per time period.

▶ **Service rate**
The average number of customers that can be served per time period.

Waiting line models require an **arrival rate** and a **service rate**. The arrival rate specifies the average number of customers per time period. For example, a system may have ten customers arrive on average each hour. The service rate specifies the average number of customers that can be serviced during a time period. The service rate is the capacity of the service system. If the number of customers you can serve per time period is less than the average number of customers arriving, the waiting line grows infinitely. You never catch up with the demand!

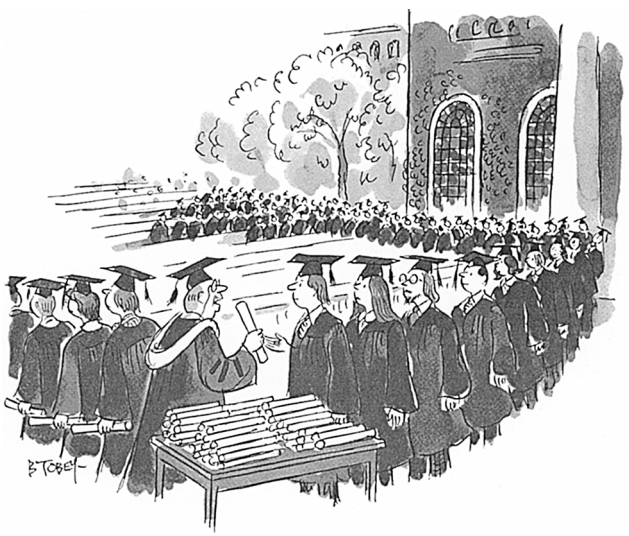
It is the variability in arrival and service patterns that causes waiting lines. Lines form when several customers request service at approximately the same time. This surge of customers temporarily overloads the service system and a line develops. Waiting line models that assess the performance of service systems usually assume that customers arrive according to a Poisson probability distribution, and service times are described by an exponential distribution. The Poisson distribution specifies the probability that a certain number of customers will arrive in a given time period (such as per hour). The exponential distribution describes the service times as the probability that a particular service time will be less than or equal to a given amount of time.

Problem-Solving Tip: Make sure the arrival rate and service rate are for the same time period, that is, the number of customers per hour, or per day, or per week.

Waiting Line Priority Rules

A waiting line priority rule determines which customer is served next. A frequently used priority rule is *first-come, first-served*. This priority rule selects customers based on who has been waiting the longest in line. Generally, customers consider *first-come, first-served* to be the fairest method for determining priority.

However, it is not the only priority rule used. Other rules include *best customers first, highest-profit customer first, quickest-service requirement first, largest-service requirement first, emergencies first*, and so on. Although each priority rule has merit, it is important to use the priority rule that best supports the overall organizational strategy. For example, a first-come, first-served rule doesn't make sense in a hospital emergency room and in fact could cause unnecessary deaths.



Barney Tobey/The Cartoon Bank, Inc.

"Congratulations, keep moving, please. Congratulations, keep moving, please. Congratulations . . ."

The priority rule used affects the performance of the waiting line system. As an example, first-come, first served is generally considered fair, yet it is biased against customers requiring short service times. When checking out at a store that is using first-come, first-served as a priority rule, a customer waiting behind another customer with a large number of items waits longer than a customer waiting behind a second customer with only a few items. Although processing is sequential, the wait times vary because of the preceding customer. Also, priority rules besides first-come, first-served may imply that some customers wait extremely long periods of time. For example, in a busy emergency room, someone not critically sick or injured could wait a significant period of time.

The models in this chapter assume a service system with a single waiting line, single or multiple servers, a single phase, and a first-come, first-served priority rule.

WAITING LINE PERFORMANCE MEASURES

Performance measures are used to gain useful information about waiting line systems. These measures include:

1. *The average number of customers waiting in line and in the system.* The number of customers waiting in line can be interpreted in several ways. Short waiting lines can result from relatively constant customer arrivals (no major surges in demand) or from the organization's having excess capacity (many cashiers open). On the other hand, long waiting lines can result from poor server efficiency, inadequate system capacity, and/or significant surges in demand.
2. *The average time customers spend waiting, and the average time a customer spends in the system.* Customers often link long waits to poor-quality service. When long waiting times occur, one option may be to change the demand pattern. That is, the company can offer discounts or better service at less busy times of the day or week. For example, a restaurant offers early-bird diners a discount so that demand is more level. The discount moves some demand from prime-time dining hours to the less desired dining hours.

   If too much time is spent in the system, customers might perceive the competency of the service provider as poor. For example, the amount of time customers spend in line and in the system at a retail checkout counter can be a result of a new employee not yet proficient at handling the transactions.
3. *The system utilization rate.* Measuring capacity utilization shows the percentage of time the servers are busy. Management's goal is to have enough servers to assure that waiting is within allowable limits but not so many servers as to be cost-inefficient.

We calculate these measures for two different waiting line models: the single-server model and the multiserver model.

SINGLE-SERVER WAITING LINE MODEL

The easiest waiting line model involves a single-server, single-line, single-phase system. The following assumptions are made when we model this environment:

1. The customers are patient (no balking, reneging, or jockeying) and come from a population that can be considered infinite.
2. Customer arrivals are described by a Poisson distribution with a mean arrival rate of λ (lambda). This means that the time between successive customer arrivals follows an exponential distribution with an average of $1/\lambda$.
3. The customer service rate is described by a Poisson distribution with a mean service rate of μ (mu). This means that the service time for one customer follows an exponential distribution with an average of $1/\mu$.
4. The waiting line priority rule used is first-come, first-served.

Using these assumptions, we can calculate the operating characteristics of a waiting line system using the following formulas:

λ = mean arrival rate of customers (average number of customers arriving per unit of time)

μ = mean service rate (average number of customers that can be served per unit of time)

$p = \frac{\lambda}{\mu}$ = the average utilization of the system

$L = \frac{\lambda}{\mu - \lambda}$ = the average number of customers in the service system

$L_Q = pL =$ the average number of customers waiting in line

$W = \frac{1}{\mu - \lambda} =$ the average time spent waiting in the system, including service

$W_Q = pW =$ the average time spent waiting in line

$P_n = (1 - p)p^n =$ the probability that n customers are in the service system at a given time

Note: The service rate must be greater than the arrival rate, that is, $\mu > \lambda$. If $\mu \leq \lambda$, the waiting line would eventually grow infinitely large. Before using the formulas, check to be sure that $\mu > \lambda$.

EXAMPLE C.1

Single-Server Operating Characteristics at the Help Desk

The computer lab at State University has a help desk to assist students working on computer spreadsheet assignments. The students patiently form a single line in front of the desk to wait for help. Students are served based on a first-come, first-served priority rule. On average, 15 students per hour arrive at the help desk. Student arrivals are best described using a Poisson distribution. The help desk server can help an average of 20 students per hour, with the service rate being described by an exponential distribution. Calculate the following operating characteristics of the service system.

(a) The average utilization of the help desk server
(b) The average number of students in the system
(c) The average number of students waiting in line
(d) The average time a student spends in the system
(e) The average time a student spends waiting in line
(f) The probability of having more than 4 students in the system

- **Before You Begin:** The key to solving queuing problems is to identify the mean arrival rate of customers and the mean service rate. In this case, on average, 15 customers arrive each hour. On average, the consultant can serve 20 customers per hour. Once you have established these values, you merely plug them into the appropriate formula.

- **Solution:**

(a) Average utilization: $p = \frac{\lambda}{\mu} = \frac{15}{20} = 0.75$, or 75%.

(b) Average number of students in the system: $L = \frac{\lambda}{\mu - \lambda} = \frac{15}{20 - 15} = 3$ students

(c) Average number of students waiting in line: $L_Q = pL = 0.75 \times 3 = 2.25$ students

(d) Average time a student spent in the system: $W = \frac{1}{\mu - \lambda} = \frac{1}{20 - 15} = 0.2$ hours, or 12 minutes

(e) Average time a student spent waiting in line: $W_Q = pW = 0.75 \times (0.2) = 0.15$ hours, or 9 minutes

(f) The probability that there are more than four students in the system equals one minus the probability that there are four or fewer students in the system. We use the following formula.

Problem-Solving Tip: Any term raised to the zero power is equal to 1.

$$P = 1 - \sum_{n=0}^{4} P_n = 1 - \sum_{n=0}^{4} (1-p)p^n$$
$$= 1 - 0.25(1 + 0.75 + 0.75^2 + 0.75^3 + 0.75^4)$$
$$= 1 - 0.7626 = 0.2374$$

or a 0.2374 (23.74 percent) chance of having more than four students in the system.

Figure C-3 shows a spreadsheet solution of this problem. The spreadsheet formulas shown are a direct implementation of the single-server formulas for performance measures. Figure C-4 is a graph of the probabilities of certain numbers of customers in the system.

	A	B	C
1	**Queuing Analysis: Single Server**		
2			
3	**Inputs**		
4	Time unit	hour	
5	Arrival Rate (lambda)	15	customers/hour
6	Service Rate (mu)	20	customers/hour
7			
8	**Intermediate Calculations**		
9	Average time between arrivals	0.066667	hour
10	Average service time	0.05	hour
11			
12	**Performance Measures**		
13	Rho (average server utilization)	**0.75**	
14	P0 (probability the system is empty)	**0.25**	
15	L (average number in the system)	**3**	customers
16	Lq (average number waiting in the queue)	**2.25**	customers
17	W (average time in the system)	**0.2**	hour
18	Wq (average time in the queue)	**0.15**	hour
19			
20	**Probability of a specific number of customers in the system**		
21	Number	2	
22	Probability	**0.140625**	

FIGURE C-3

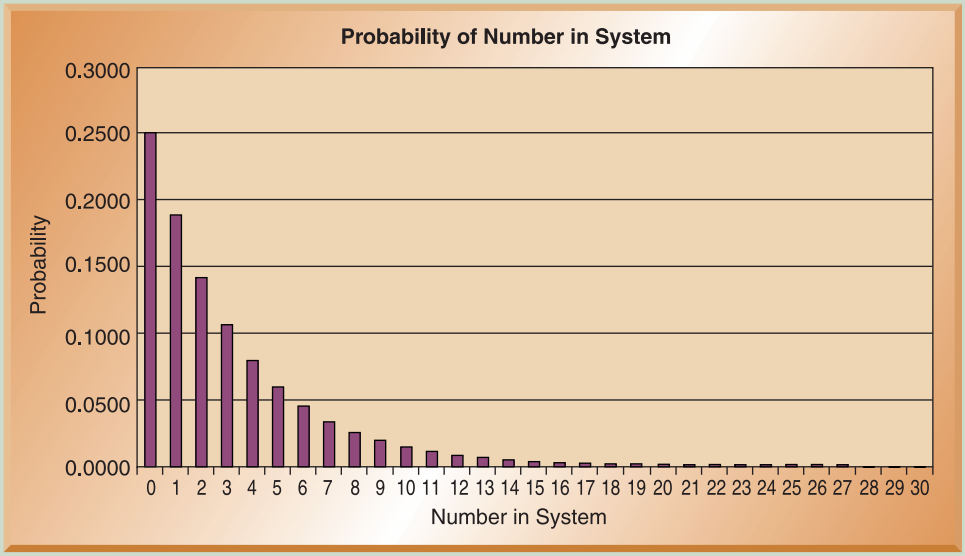
Spreadsheet for single-server operating characteristics

Key Formulas

B9: = 1/B5
B10: = 1/B6
B13: = B5/B6
B14: = 1−B13
B15: = B5/(B6−B5)
B16: = B13*B15
B17: = 1/(B6−B5)
B18: = B13*B17
B22: = (1−B$13)*(B13^B21)

FIGURE C-4

Single-server probabilities of customers in the system



MULTISERVER WAITING LINE MODEL

In the single-line, multiserver, single-phase model, customers form a single line and are served by the first server available. The model assumes that there are s identical servers, the service time distribution for *each server* is exponential, and the mean service time is $1/\mu$. Using these assumptions, we can describe the operating characteristics with the following formulas:

s = the number of servers in the system

$$p = \frac{\lambda}{s\mu} = \text{the average utilization of the system}$$

$$P_0 = \left[\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!}\left(\frac{1}{1-p}\right)\right]^{-1} = \text{the probability that no customers are in the system}$$

$$L_Q = \frac{P_o(\lambda/\mu)^s p}{s!(1-p)^2} = \text{the average number of customers waiting in line}$$

$$W_Q = \frac{L_Q}{\lambda} = \text{the average time spent waiting in line}$$

$$W = W_Q + \frac{1}{\mu} = \text{the average time spent in the system, including service}$$

$$L = \lambda W = \text{the average number of customers in the service system}$$

$$P_n = \begin{cases} \dfrac{(\lambda/\mu)^n}{n!} P_0 & \text{for } n \le s \\ \dfrac{(\lambda/\mu)^n}{s!s^{n-s}} P_0 & \text{for } n > s \end{cases} = \text{the probability that } n \text{ customers are in the system at a given time}$$

Note: The total service rate must be greater than the arrival rate, that is, $s\mu > \lambda$. If $s\mu \leq \lambda$, the waiting line would eventually grow infinitely large. Before using the formulas, check to be sure that $s\mu > \lambda$.

Multiserver Operating Characteristics at the Help Desk — EXAMPLE C.2

State University has decided to increase the number of computer assignments in its curriculum and is concerned about the impact on the help desk. Instead of a single person working at the help desk, the university is considering a plan to have three identical service providers. It expects that students will arrive at a rate of 45 per hour, according to a Poisson distribution. The service rate for each of the three servers is 18 students per hour, with exponential service times. Calculate the following operating characteristics of the service system:

(a) The average utilization of the help desk
(b) The probability that there are no students in the system
(c) The average number of students waiting in line
(d) The average time a student spends waiting in line
(e) The average time a student spends in the system
(f) The average number of students in the system

Problem-Solving Tip: By definition, zero factorial, 0!, equals 1.

• Solution:

(a) Average utilization: $p = \dfrac{\lambda}{s\mu} = \dfrac{45}{(3 \times 18)} = 0.833$, or 83.3%

(b) The probability that there are no students in the system:

$$P_0 = \left[\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!}\left(\frac{1}{1-p}\right)\right]^{-1}$$

$$= \left[\frac{(45/18)^0}{0!} + \frac{(45/18)^1}{1!} + \frac{(45/18)^2}{2!} + \left(\frac{(45/18)^3}{3!}\left(\frac{1}{1-0.833}\right)\right)\right]^{-1}$$

$$= \frac{1}{22.215} = 0.045, \text{ or } 4.5\% \text{ of having no students in the system}$$

(c) The average number of students waiting in line:

$$L_Q = \frac{P_0(\lambda/\mu)^s p}{s!(1-p)^2} = \frac{0.045(45/18)^3 \times 0.833}{3! \times (1-0.833)^2} = \frac{0.5857}{0.1673} = 3.5 \text{ students}$$

(d) The average time a student spends waiting in line: $W_Q = \dfrac{L_Q}{\lambda} = \dfrac{3.5}{45} = 0.078$ hour, or 4.68 minutes

(e) The average time a student spends in the system:

$$W = W_Q + \frac{1}{\mu} = 0.078 + \frac{1}{18} = 0.134 \text{ hour, or } 8.04 \text{ minutes}$$

(f) The average number of students in the system:

$$L = \lambda W = 45(0.134) = 6.03 \text{ students}$$

Figure C-5 shows a spreadsheet solution of this problem. The spreadsheet formulas are a direct implementation of the multiple-server formulas for performance measures. Because of the complexity of the P_0 calculation, the columns E–H break this computation down.

Then the formula in cell B16 looks up the value from column H corresponding to the number of servers. The spreadsheet shown here will work for up to a 100-server system. Key formulas are listed here:

Key Formulas

F10: = F$5^E10 (copied down)
G10: = E10*G9 (copied down)
H10: = H9+(F10/G10) (copied down)
F5: = B5/B6
F6: = INDEX(G9:G109, B7+1)
B10: = 1/B5
B11: = 1/B6
B12: = B7*B6
B15: = B5/B12
B16: = (INDEX(H9:H109, B7)+(((F5^B7)/F6)*((1)/(1−B15))))^(−1)
B17: = B5*B19
B18: = (B16*(F5^B7)*B15)/(INDEX(G9:G109, B7+1)*(1−B15)^2)
B19: = B20+(1/B6)
B20: = B18/B5
B24: = IF(B23 <=B7, ((F5^B23)*B16)/INDEX(G9:G109, B23+1),
((F5^B23)*B16)/(INDEX(G9:G109, B7+2)*(B7^(B23−B7))))

FIGURE C-5

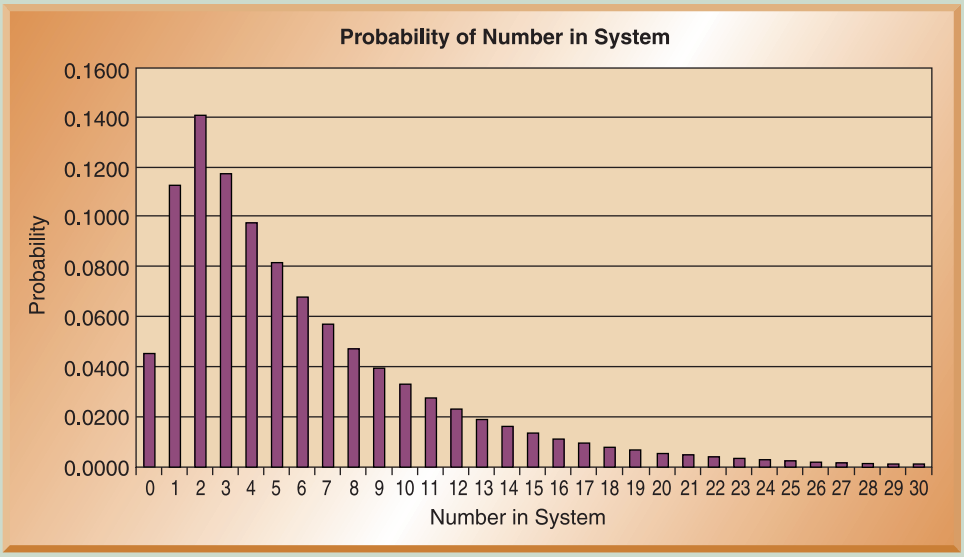
Spreadsheet for multiple-server operating characteristics

	A	B	C	D	E	F	G	H
1	**Queuing Analysis: Multiple Servers**							
2								
3	**Inputs**				**Working Calculations, mainly for P0 Calculation**			
4	Time unit	hour						
5	Arrival Rate (lambda)	45	customers/hour		lambda/mu	2.5		
6	Service Rate per Server (mu)	18	customers/hour		s!	6		
7	Number of Servers (s)	3	servers					
8					**n**	**(λ/μ)^n**	**n!**	**Sum**
9	**Intermediate Calculations**				0	1	1	1
10	Average time between arrivals	0.022222	hour		1	2.5	1	3.5
11	Average service time per server	0.055556	hour		2	6.25	2	6.625
12	Combined service rate (s*mu)	54	customers/hour		3	15.625	6	9.22916667
13					4	39.0625	24	10.8567708
14	**Performance Measures**				5	97.65625	120	11.6705729
15	Rho (average server utilization)	**0.833333**			6	244.14063	720	12.0096571
16	P0 (probability the system is empty)	**0.044944**			7	610.35156	5040	12.1307586
17	L (average number in the system)	**6.011236**	customers		8	1525.8789	40320	12.1686028
18	Lq (average number waiting in the queue)	**3.511236**	customers		9	3814.6973	362880	12.1791151
19	W (average time in the system)	**0.133583**	hour		10	9536.7432	3628800	12.1817432
20	Wq (average time in the queue)	**0.078027**	hour		11	23841.858	39916800	12.1823405
21					12	59604.645	479001600	12.1824649
22	**Probability of a specific number of customers in the system**				13	149011.61	6.227E+09	12.1824888
23	Number	5			14	372529.03	8.718E+10	12.1824931
24	Probability	**0.081279**			15	931322.57	1.308E+12	12.1824938
25					16	2328306.4	2.092E+13	12.1824939
26					17	5820766.1	3.557E+14	12.182494
27					18	14551915	6.402E+15	12.182494
108					99	2.489E+39	9.33E+155	12.182494
109					100	6.223E+39	9.33E+157	12.182494

Figure C-6 is a graph of the probabilities of certain numbers of customers in the system.

FIGURE C-6

Multiple-server probabilities of customers in the system



CHANGING OPERATIONAL CHARACTERISTICS

After calculating the operating characteristics for a waiting line system, sometimes you need to change the system to alter its performance. Let's look at the type of changes you can make to the different elements of the waiting line system.

Customer arrival rates. You can try to change arrival rates in a number of ways. For example, you can provide discounts or run special promotions during the nonpeak hours to attract customers.

Number and type of service facilities. You can either increase or decrease the number of server facilities. For example, a grocery store can easily change the number of cashiers open for business (up to the number of registers available). The store increases the number of cashiers open when lines are too long.

Another approach is to dedicate specific servers for specific transactions. One example would be to limit the number of items that can be processed at a particular cashier (ten items or less) or to limit a cashier to cash-only transactions. Still another possibility is to install self-service checkout systems.

Changing the number of phases. You can use a multiphase system where servers specialize in a portion of the total service rather than needing to know the entire service provided. Since a server has fewer tasks to learn, the individual server proficiency should improve. This goes back to the concept of division of labor.

Server efficiency. You can improve server efficiency through process improvements or dedication of additional resources. For example, cashier accuracy and speed are improved through the use of scanners. Service speed can also be increased by dedicating additional resources. For example, if a grocery bagger is added at each cashier station, service speed will be improved and customers will flow through the system more quickly.

Changing the priority rule. The priority rule determines who should be served next. There are priority rules other than first-come, first-served. If you want to change priority rules, consider the impact on those customers who will wait longer.

Changing the number of lines. Changing to a single-line model from a multiline model is most appropriate when the company is concerned about fairness for its customers. A single line ensures that customers do not jockey in an attempt to gain an advantage over another customer. Multiline models easily accommodate specialty servers (express lanes).

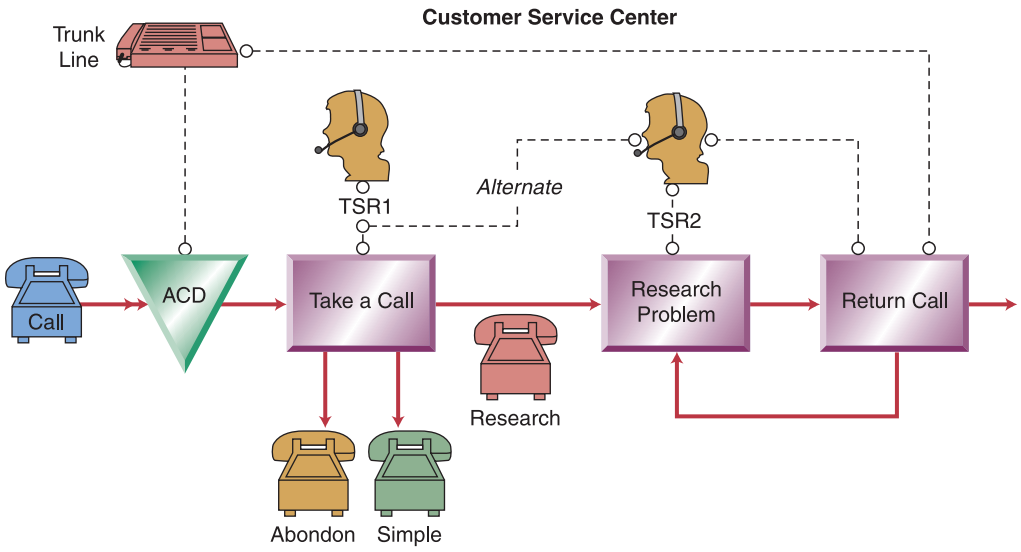
Once changes are suggested, evaluate their impact on the performance characteristics of the waiting line system. Changes in one area can require changes in other areas. For example, if you achieve a more constant customer arrival rate, you may be able to reduce the number of service facilities.

LARGER-SCALE WAITING LINE SYSTEMS

This supplement provides an introduction to waiting line system analysis. There are many variations of the mathematical models that are not discussed here, such as finite queue length, finite customer population, and nonexponential arrival and service times. As waiting line systems become more complex, especially as they involve more than one phase, concise mathematical formulas in general do not exist for system performance measures. Therefore, for most larger-scale waiting line systems, discrete-event simulation is often used to analyze these systems. Discrete-event simulation products allow the user to define the system, arrival and service patterns, and other aspects of the system. Then the simulation is run to mimic the behavior of the system in reality, and the results are statistically analyzed to determine system performance. Figure C-7 shows a

FIGURE C-7

Customer service center simulation layout



screen shot from a model completed in the ProcessModel™ simulation software. In this system, calls arrive at a customer service center and are answered by one operator (with a second operator as a backup server). Some calls are resolved quickly, others need further research, and still others don't get answered in a timely fashion and the callers renege (hang up) before being served. Once the system is completely specified, including arrival and service rates, the simulation is run and system performance measures are automatically calculated.

WAITING LINE MODELS WITHIN OM: HOW IT ALL FITS TOGETHER

Although it is unlikely that you calculate performance measures for the lines you wait in on a day-to-day basis, you should now be aware of the potential for mathematical analysis of these systems. More importantly, management has a tool by which it can evaluate system performance and make decisions as to how to improve the performance while weighing performance against the costs to achieve that performance.

Waiting line models are important to a company because they directly affect customer service perception and the costs of providing a service. Several functional areas are affected by waiting line decisions. Accounting is concerned with the cost of the waiting line system used. If system average utilization is low, that suggests the waiting line design is inefficient and too expensive. Poor system design can result in overstaffing or unnecessary capital acquisitions in an effort to improve customer service. Marketing is concerned about response time for customers—how long customers must wait in line before being served and how long it takes to be served. Quick service or response can be a competitive advantage. Long waits suggest a lack of concern by the organization or can be linked to a perception of poor service quality. Purchasing must be sure to buy capital equipment capable of achieving the proposed service rate. Operations uses waiting line theory to estimate queues or waiting times at different processing points, to allow for a better estimate of lead time and improve due-date delivery promising. Operations is also affected by the system design. When single-phase systems are used, operators must have greater skills. The organization needs to hire employees with higher skill levels or provide training to upgrade the workforce.

Supplement Highlights

1. The elements of a waiting line system include the customer population source, the patience of the customer, the service system, arrival and service distributions, waiting line priority rules, and system performance measures. Understanding these elements is critical when analyzing waiting line systems.

2. Waiting line models allow us to estimate system performance by predicting average system utilization, average number of customers in the service system, average number of customers waiting in line, average time a customer spends in the system, average time a customer waits in line, and the probability of n customers in the service system.

3. The benefit of calculating operational characteristics is to provide management with information as to whether system changes are needed. Management can change the operational performance of the waiting line system by altering any or all of the following: the customer arrival rates, the number of service facilities, the number of phases, server efficiency, the priority rule, and the number of lines in the system. Based on proposed changes, management can then evaluate the expected performance of the system.

Key Terms

waiting line system C2
queuing system C2
finite customer population C3
infinite customer population C4
balking C4
reneging C4
jockeying C4
arrival rate C6
service rate C6

Formula Review

For Single-Server Waiting Line Models

$p = \frac{\lambda}{\mu}$ is the average utilization of the system

$W = \frac{\lambda}{\mu - \lambda}$ is the average number of customers in the service system

$L_Q = pL$ is the average number of customers waiting in line

$W = \frac{1}{\mu - \lambda}$ is the average time spent waiting in the system, including service

$W_Q = pW$ is the average time spent waiting in line

$P_n = (1 - p)p^n$ is the probability that n customers are in the service system at a given time

For the Multiserver Waiting Line Model

$p = \frac{\lambda}{s\mu}$ is the average utilization of the system

$P_0 = \left[\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!}\left(\frac{1}{1-p}\right)\right]^{-1}$ is the probability that no customers are in the system

$L_Q = \frac{P_o(\lambda/\mu)^s p}{s!(1-p)^2}$ is the average number of customers waiting in line

$W_Q = \frac{L_Q}{\lambda}$ is the average time spent waiting in line

$W = W_Q + \frac{1}{\mu}$ is the average time spent in the system, including service

$L = \lambda W$ is the average number of customers in the service system

$$P_n = \begin{cases} \frac{(\lambda/\mu)^n}{n!} P_0 & \text{for } n \le s \\ \frac{(\lambda/\mu)^n}{s!s^{n-s}} P_0 & \text{for } n > s \end{cases} = \text{the probability that } n \text{ customers are in the system at a given time}$$

Solved Problems

• Problem 1

The local Division of Motor Vehicles (DMV) is concerned with its waiting line system. Currently, the DMV uses a single-server, single-line, single-phase system when processing license renewals. Based on historical evidence, the average number of customers arriving per hour is 9 and is described by a Poisson distribution. The service rate is 12 customers per hour with the service times following an exponential distribution. The customers are patient and come from an infinite population. The manager of the DMV would like you to calculate the operational characteristics of the waiting line system.

(a) What is the average system utilization?
(b) What is the average number of customers in the system?
(c) What is the average number of customers waiting in line?
(d) What is the average time a customer spends in the system?
(e) What is the average time a customer spends waiting in line?

• *Before You Begin*

The key here is identifying the mean arrival rate and the mean service rate. The average number of customers per hour is given as 9. The server is able to service 12 customers per hour on average. Plug these values into the appropriate formula.

• *Solution*

(a) Average utilization is 0.75, or 75 percent.

$$p = \frac{\lambda}{\mu} = \frac{9}{12} = 0.75$$

(b) Average number of customers in the system is 3.

$$L = \frac{\lambda}{\mu - \lambda} = \frac{9}{12 - 9} = 3 \text{ customers}$$

(c) Average number of customers waiting in line is 2.25.

$$L_Q = pL = 0.75 \times 2.25 \text{ customers}$$

(d) Average time a customer spends in the system is 0.33 hours, or 20 minutes.

$$W = \frac{1}{\mu - \lambda} = \frac{1}{12 - 9} = 0.33 \text{ hours}$$

(e) Average time a customer spends waiting in line is 0.25 hours, or 15 minutes.

$$W_Q = pW = 0.75 \times 0.33 = 0.25 \text{ hours}$$

These operational characteristics can be calculated as shown in the spreadsheet. Using a spreadsheet allows the modeler to vary parameters quickly and see the resulting operational characteristics.

Note: This is the same spreadsheet model as introduced with Example C.1; only the input values have been changed.

	A	B	C
1	**Queuing Analysis: Single Server**		
2	**Solved Problem C.1**		
3	**Inputs**		
4	Time unit	hour	
5	Arrival Rate (lambda)	9	customers/hour
6	Service Rate (mu)	12	customers/hour
7			
8	**Intermediate Calculations**		
9	Average time between arrivals	0.111111	hour
10	Average service time	0.083333	hour
11			
12	**Performance Measures**		
13	Rho (average server utilization)	**0.75**	
14	P0 (probability the system is empty)	**0.25**	
15	L (average number in the system)	**3**	customers
16	Lq (average number waiting in the queue)	**2.25**	customers
17	W (average time in the system)	**0.333333**	hour
18	Wq (average time in the queue)	**0.25**	hour
19			
20	**Probability of a specific number of customers in the system**		
21	Number	2	
22	Probability	**0.140625**	

• Problem 2

The county has decided to consolidate several of its DMV facilities into a larger, centrally located facility. The DMV manager wants you to calculate the operational characteristics of a multiserver, single-phase waiting line system. The arrival rate is expected to be 72 customers per hour and follows a Poisson distribution. The number of identical servers is seven. Each server will be able to serve an average of 12 customers per hour. The service times are described by an exponential distribution. Your job is to calculate the following:

(a) The average system utilization
(b) The probability of no customers in the system
(c) The average number of customers waiting in line
(d) The average time a customer waits in line
(e) The average time a customer spends in the system

• *Before You Begin*

Once again, the key is identifying the mean arrival rate and the mean service rate. Here the arrival rate is given as 72 customers per hour, while each of the seven servers is able to serve 12 customers per hour. Therefore, 84 customers can be served per hour.

• *Solution*

(a) Average system utilization is 0.857, or 85.7 percent.

$$p = \frac{\mu}{s\lambda} = \frac{72}{7 \times 12} = 0.8571$$

(b) The probability that no customers are in the system is 0.0016, or 0.2 percent.

$$P_0 = \left[\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!}\left(\frac{1}{1-p}\right)\right]^{-1}$$

$$= \left[\left[\frac{(72/12)^0}{0!} + \frac{(72/12)^1}{1!} + \frac{(72/12)^2}{2!} + \frac{(72/12)^3}{3!} + \frac{(72/12)^4}{4!} + \frac{(72/12)^5}{5!} + \frac{(72/12)^6}{6!}\right] + \left[\frac{(72/12)^7}{7!}\left(\frac{1}{1-p}\right)\right]\right]^{-1}$$

$$= \frac{1}{623.8} = 0.001579$$

(c) The average number of customers waiting in line is 3.68.

$$L_Q = \frac{P_0(\lambda/\mu)^s p}{s!(1-p)^2}$$
$$= \frac{0.001579(72/12)^7 \times 0.857}{7!(1-0.857)^2} = 3.683 \text{ customers}$$

(d) The average time spent waiting in line is 0.05115 hours, or 3.07 minutes.

$$W_Q = \frac{L_Q}{\lambda} = \frac{3.683}{72} = 0.05115 \text{ hours}$$

(e) The average time a customer spends in the system is 0.1345 hours, or 8.07 minutes.

$$W = W_Q + \frac{1}{\mu} = 0.05115 + \frac{1}{12} = 0.1345 \text{ hours}$$

The spreadsheet solution is shown. This model is the same as that introduced with Example C.2, with only the input values changed.

	A	B	C
1	**Queuing Analysis: Multiple Servers**		
2	**Solved Problem C.2**		
3	**Inputs**		
4	Time unit	hour	
5	Arrival Rate (lambda)	72	customers/hour
6	Service Rate per Server (mu)	12	customers/hour
7	Number of Servers (s)	7	servers
8			
9	**Intermediate Calculations**		
10	Average time between arrivals	0.013889	hour
11	Average service time per server	0.083333	hour
12	Combined service rate (s*mu)	84	customers/hour
13			
14	**Performance Measures**		
15	Rho (average server utilization)	**0.857143**	
16	P0 (probability the system is empty)	**0.001579**	
17	L (average number in the system)	**9.682981**	customers
18	Lq (average number waiting in the queue)	**3.682981**	customers
19	W (average time in the system)	**0.134486**	hour
20	Wq (average time in the queue)	**0.051153**	hour
21			
22	**Probability of a specific number of customers in the system**		
23	Number	5	
24	Probability	**0.102305**	

Discussion Questions

1. Describe the elements of a waiting line system.
2. Provide examples of when a single-line, single-server, single-phase waiting line system is appropriate.
3. Describe the operating performance characteristics calculated for evaluating waiting line systems.
4. Describe the implications for customer service and server skills when using a single-line, single-server, single-phase waiting line system.
5. Describe the implications for customer service and server skills when using a single-line, multiserver, single-phase waiting line system.
6. Describe the implications for customer service and server skills when a multiserver, multistage waiting line system is used.
7. Describe a situation in your daily life that could be improved by waiting line analysis.
8. Explain how the design of a waiting system can negatively affect customers.
9. Visit your local bank and observe the waiting line system. Describe the system in terms of number of lines, number of facilities, and number of phases.
10. On your next trip to the Division of Motor Vehicles, evaluate its waiting line system.
11. Describe any disadvantages of using waiting line models.

Problems

1. Melanie is the manager of the Clean Machine car wash and has gathered the following information. Customers arrive at a rate of eight per hour according to a Poisson distribution. The car washer can service an average of ten cars per hour with service times described by an exponential distribution. Melanie is concerned about the number of customers waiting in line. She has asked you to calculate the following system characteristics:
   (a) Average system utilization
   (b) Average number of customers in the system
   (c) Average number of customers waiting in line

2. Melanie realizes that how long the customer must wait is also very important. She is also concerned about customers balking when the waiting line is too long. Using the arrival and service rates in Problem 1, she wants you to calculate the following system characteristics:
   (a) The average time a customer spends in the system
   (b) The average time a customer spends waiting in line
   (c) The probability of having more than three customers in the system
   (d) The probability of having more than four customers in the system

3. If Melanie adds an additional server at Clean Machine car wash, the service rate changes to an average of 16 cars per hour. The customer arrival rate is 10 cars per hour. Melanie has asked you to calculate the following system characteristics:
   (a) Average system utilization
   (b) Average number of customers in the system
   (c) Average number of customers waiting in line

4. Melanie is curious to see the difference in waiting times for customers caused by the additional server added in Problem 3. Calculate the following system characteristics for her:
   (a) The average time a customer spends in the system
   (b) The average time a customer spends waiting in line
   (c) The probability of having more than three customers in the system
   (d) The probability of having more than four customers in the system

5. After Melanie added another car washer at Clean Machine (service rate is an average of 16 customers per hour), business improved. Melanie now estimates that the arrival rate is 12 customers per hour. Given this new information, she wants you to calculate the following system characteristics:
   (a) Average system utilization
   (b) Average number of customers in the system
   (c) Average number of customers waiting in line

6. As usual, Melanie then requested you to calculate system characteristics concerning customer time spent in the system.
   (a) Calculate the average time a customer spends in the system.
   (b) Calculate the average time a customer spends waiting in line.
   (c) Calculate the probability of having more than four customers in the system.

7. Business continues to grow at Clean Machine. Melanie has decided to use a second car washing bay, staffed with another identical two-person team. Clean Machine will now use a single-line, multiserver, single-phase waiting line system. The arrival rate is estimated to average 24 customers per hour according to a Poisson distribution. Each of the car wash teams can service an average of 16 customers per hour according to an exponential distribution. Calculate the following operational characteristics:
   (a) Average system utilization
   (b) Average number of customers in the system
   (c) Average number of customers waiting in line
   (d) Average time a customer spends in the system
   (e) Average time a customer spends waiting in line
   (f) Probability of having more than four customers in the system.

8. Melanie is very concerned about the number of customers waiting in line. Given the information in Problem 7, calculate how high the customer arrival rate can increase without the average number of customers waiting in line exceeding four.

CASE: The Copy Center Holdup

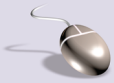
Catherine Blake, the office manager for the College of Business Administration, has received numerous complaints lately from several department chairpersons. In the past few months, the chairpersons have insisted that something be done about the amount of time their administrative assistants waste waiting in line to make copies. Currently, the college has two photocopy centers dedicated to small copying jobs: copy center A on the third floor and copy center B on the fourth floor. Both centers are self-serve and have identical processing capabilities. The copying machines are not visible to the administrative assistants from their offices. When copying is required, the administrative assistant goes to the copy room and waits in line to make the necessary copies. Catherine's assistant, Brian, was assigned to investigate the problem.

Brian reported that, on average, administrative assistants arrive at copy center A at the rate of 10 per hour and at copy center B at the rate of 14 per hour. Each of the copy centers can service 15 jobs per hour. The administrative assistants' arrivals essentially follow a Poisson distribution, and the service times are approximated by a negative exponential distribution. Brian has proposed that the two copy centers be combined into a single copy center with either two or three identical copy machines. He estimates that the arrival rate would be 24 per hour. Each

machine would still service 15 jobs per hour. Currently, administrative assistants earn an average of $15 per hour.

(a) Determine the utilization of each of the copy centers.
(b) Determine the average waiting time at each of the copy centers.
(c) What is the annual cost of the administrative assistants' average waiting time using the current system?
(d) Determine the utilization of the combined copy center with two copiers.
(e) Determine the average waiting time at the combined copy center.
(f) What would be the annual cost of the administrative assistants' average waiting time using the combined two-copier setup?
(g) What would be the utilization of the combined copy center with three copiers?
(h) What would be the annual cost of the administrative assistants' average waiting time using the combined three-copier setup?
(i) What would you recommend to Catherine?

On-line Resources



Visit our dynamic Web site, www.wiley.com/college/reid, for more cases, Web links, and additional information.

1. **Spreadsheets for Examples C.1 and C.2 and Solved Problems 1 and 2 are available on the CD.**
2. **Additional Web Resources**

- Real Queuing Examples: *http://www2.uwindsor.ca/~hlynka/qreal.html* This site contains excerpts from news articles that deal with aspects of waiting lines.
- ClearQ: *http://clearq.com/* This company produces "take-a-number" systems for service facilities (e.g., delis), but also provides performance information about the waiting line.
- Qmatic: *http://us.q-matic.com/index.html* This company produces informational displays and other products to keep customers informed about waiting times.
- "Queuing Presentation" by Richard Larson, given at the Institute for Operations Research and the Management Sciences: *http://caes.mit.edu/people/larson/MontrealINFORMS1/sld001.htm*.
- The Queuing Theory Tutor: *http://www.dcs.ed.ac.uk/home/jeh/Simjava/queueing/mm1_q/mm1_q.html*. This site has two animated displays of waiting lines. The user can change arrival and service rates to see how performance is affected.
- Myron Hlynka's Queuing Page: *http:www2.uwindsor.ca/~hlynka/queue.html*. This Web site contains information about waiting lines as well as links to other interesting sites.
- Queuing ToolPak: *http://www.bus.ualberta.ca/aingolfsson/qtp/*. The Queuing ToolPak is an Excel add-in that allows you to easily compute performance measures for a number of different waiting line models.

Internet Challenge

Visit the Web site for the United Network for Organ Sharing (UNOS) ⟨http://www.unos.org⟩. This is the United States organization coordinating the donation, assignment, and transplantation of human organs. Research the organ donation situation, and prepare a report summarizing your findings from a waiting line standpoint. What are some of the important performance measures in this waiting line system? Describe the elements of the waiting line system. What is the priority rule used for selecting the next customer? Propose at least two other priority rules, and discuss what changes you think would occur in the operating characteristics. Besides the quantitative performance characteristics, what other considerations would need to be made in order to change the priority rule? Support your discussion with data and information obtainable at the UNOS Web site.

Selected Bibliography

Albright, S. Christian, and Wayne Winston. *Essentials of Practical Management Science.* Mason, OH: Cengage, 2005.

Hall, Randolph W. *Queueing Methods for Services and Manufacturing.* Englewood Cliffs, N.J.: Prentice-Hall, 1991.

Moore, P.M. *Queues, Inventories and Maintenance.* New York: Wiley, 1958.

Ragsdale, Cliff T. *Spreadsheet Modeling & Decision Analysis*, Fifth Edition. Stamford, Conn.: Thomson, 2008.