18

The Application of Queueing Theory

Queueing theory has enjoyed a prominent place among the modern analytical techniques of OR. However, the emphasis thus far has been on developing a descriptive mathematical theory. Thus, queueing theory is not directly concerned with achieving the goal of OR: optimal decision making. Rather, it develops information on the behavior of queueing systems. This theory provides part of the information needed to conduct an OR study attempting to find the best design for a queueing system.

This chapter discusses the *application* of queueing theory in the broader context of an overall OR study. It begins by introducing three examples that will be used for illustration throughout the chapter. Section 18.2 discusses the basic considerations for decision making in this context. The following two sections then develop decision models for the *optimal* design of queueing systems. The chapter concludes with a survey of some award-winning applications of queueing theory.

18.1 EXAMPLES

Example 1—How Many Repairers?

SIMULATION, INC., a small company that makes gidgets for analog computers, has 10 gidget-making machines. However, because these machines break down and require repair frequently, the company has only enough operators to operate eight machines at a time, so two machines are available on a standby basis for use while other machines are down. Thus, eight machines are always operating whenever no more than two machines are waiting to be repaired, but the number of operating machines is reduced by 1 for each additional machine waiting to be repaired.

The time until any given operating machine breaks down has an exponential distribution, with a mean of 20 days. (A machine that is idle on a standby basis cannot break down.) The time required to repair a machine also has an exponential distribution, with a mean of 2 days. Until now the company has had just one repairer to repair these machines, which has frequently resulted in reduced productivity because fewer than eight machines are operating. Therefore, the company is considering hiring a second repairer, so that two machines can be repaired simultaneously.

Thus, the queueing system to be studied has the repairers as its servers and the machines requiring repair as its customers, where the problem is to choose between having one or two servers. (Notice the analogy between this problem and the County Hospital emergency room problem described in Sec. 17.1.) With one slight exception, this system fits the *finite calling population variation* of the *M*/*M*/*s* model presented in Sec. 17.6, where $N = 10$ machines, $\lambda = \frac{1}{20}$ customer per day (for each operating machine), and $\mu = \frac{1}{2}$ customer per day. The exception is that the λ_0 and λ_1 parameters of the birth-and-death process are changed from $\lambda_0 = 10\lambda$ and $\lambda_1 = 9\lambda$ to $\lambda_0 = 8\lambda$ and $\lambda_1 = 8\lambda$. (All the other parameters are the same as those given in Sec. 17.6.) Therefore, the C_n factors for calculating the P_n probabilities change accordingly (see Sec. 17.5).

Each repairer costs the company approximately $280 per day. However, the estimated *lost profit* from having fewer than eight machines operating to produce gidgets is $400 per day for each machine down. (The company can sell the full output from eight operating machines, but not much more.)

The analysis of this problem will be pursued in Secs. 18.3 and 18.4.

Example 2—Which Computer?

EMERALD UNIVERSITY is making plans to lease a supercomputer to be used for scientific research by the faculty and students. Two models are being considered: one from the MBI Corporation and the other from the CRAB Company. The MBI computer costs more but is somewhat faster than the CRAB computer. In particular, if a sequence of typical jobs were run continuously for one 24-hour day, the number completed would have a Poisson distribution with a mean of 30 and 25 for the MBI and the CRAB computers, respectively. It is estimated that an average of 20 jobs will be submitted per day and that the time from one submission to the next will have an exponential distribution with a mean of 0.05 day. The leasing cost per day would be $5,000 for the MBI computer and $3,750 for the CRAB computer.

Thus, the queueing system of concern has the computer as its (single) server and the jobs to be run as its customers. Furthermore, this system fits the *M*/*M*/1 model presented at the beginning of Sec. 17.6. With 1 day as the unit of time, $\lambda = 20$ customers per day, and $\mu = 30$ and 25 customers per day with the MBI and the CRAB computers, respectively. You will see in Secs. 18.3 and 18.4 how the decision was made between the two computers.

Example 3—How Many Tool Cribs?

The MECHANICAL COMPANY is designing a new plant. This plant will need to include one or more tool cribs in the factory area to store tools required by the shop mechanics. The tools will be handed out by clerks as the mechanics arrive and request them and will be returned to the clerks when they are no longer needed. In existing plants, there have been frequent complaints from supervisors that their mechanics have had to waste too much time traveling to tool cribs and waiting to be served, so it appears that there should be *more* tool cribs and *more* clerks in the new plant. On the other hand, management is exerting pressure to reduce overhead in the new plant, and this reduction would

lead to *fewer* tool cribs and *fewer* clerks. To resolve these conflicting pressures, an OR study is to be conducted to determine just how many tool cribs and clerks the new plant should have.

Each tool crib constitutes a queueing system, with the clerks as its servers and the mechanics as its customers. Based on previous experience, it is estimated that the time required by a tool crib clerk to service a mechanic has an exponential distribution, with a mean of $\frac{1}{2}$ minute. Judging from the anticipated number of mechanics in the entire factory area, it is also predicted that they would require this service randomly but at a mean rate of 2 mechanics per minute. Therefore, it was decided to use the *M*/*M*/*s* model of Sec. 17.6 to represent each queueing system. With 1 hour as the unit of time, $\mu = 120$. If only one tool crib were to be provided, λ also would be 120. With more than one tool crib, this mean arrival rate would be divided among the different queueing systems.

The total cost to the company of each tool crib clerk is about $20 per hour. The capital recovery costs, upkeep costs, and so forth associated with each tool crib provided are estimated to be $16 per working hour. While a mechanic is busy, the value to the company of his or her output averages about $48 per hour.

Sections 18.3 and 18.4 include discussions of how this (and additional) information was used to make the required decisions.

18.2 DECISION MAKING

Queueing-type situations that require decision making arise in a wide variety of contexts. For this reason, it is not possible to present a meaningful decision-making procedure that is applicable to all these situations. Instead, this section attempts to give a broad conceptual picture of a typical approach.

Designing a queueing system typically involves making one or a combination of the following decisions:

1. Number of servers at a service facility
2. Efficiency of the servers
3. Number of service facilities.

When such problems are formulated in terms of a queueing model, the corresponding decision variables usually are s (number of servers at each facility), μ (mean service rate per busy server), and λ (mean arrival rate at each facility). The *number of service facilities* is directly related to λ because, assuming a uniform workload among the facilities, λ equals the total mean arrival rate to all facilities divided by the number of facilities.

Refer to Sec. 18.1 and note how the three examples there respectively illustrate situations involving these three decisions. In particular, the decision facing Simulation, Inc., is *how many repairers* (servers) to provide. The problem for Emerald University is *how fast a computer* (server) is needed. The problem facing Mechanical Company is *how many tool cribs* (service facilities) to install as well as *how many clerks* (servers) to provide at each facility.

The first kind of decision is particularly common in practice. However, the other two also arise frequently, particularly for the internal service systems described in Sec. 17.3. One example illustrating a decision on the efficiency of the servers is the selection of the

type of materials-handling equipment (the servers) to purchase to transport certain kinds of loads (the customers). Another such example is the determination of the size of a maintenance crew (where the entire crew is one server). Other decisions concern the number of service facilities, such as copy centers, computer facilities, tool cribs, storage areas, and so on, to distribute throughout an area.

All the specific decisions discussed here involve the general question of the *appropriate level of service* to provide in a queueing system. As mentioned at the beginning of Chap. 17, decisions regarding the amount of service capacity to provide usually are based primarily on two considerations: (1) the cost incurred by providing the service, as shown in Fig. 18.1, and (2) the amount of waiting for that service, as suggested in Fig. 18.2. Figure 18.2 can be obtained by using the appropriate waiting-time equation from queueing theory.

These two considerations create conflicting pressures on the decision maker. The objective of reducing service costs recommends a minimal level of service. On the other hand, long waiting times are undesirable, which recommends a high level of service. Therefore, it is necessary to strive for some type of compromise. To assist in finding this compromise, Figs. 18.1 and 18.2 may be combined, as shown in Fig. 18.3. The problem is thereby reduced to selecting the point on the curve of Fig. 18.3 that gives the best balance between the average delay in being serviced and the cost of providing that service. Reference to Figs. 18.1 and 18.2 indicates the corresponding level of service.

FIGURE 18.1
Service cost as a function of service level.

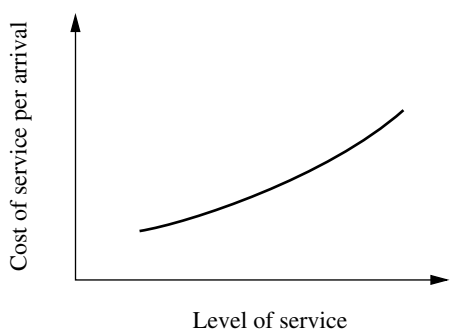
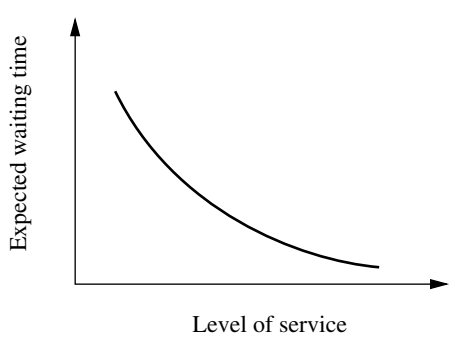


FIGURE 18.2
Expected waiting time as a function of service level.



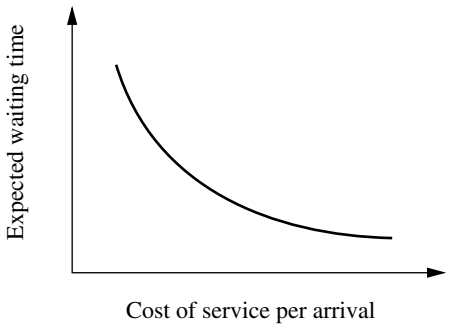


FIGURE 18.3
Relationship between average delay and service cost.

Obtaining the proper balance between delays and service costs requires answers to such questions as, How much expenditure on service is equivalent (in its detrimental impact) to a customer's being delayed 1 unit of time? Thus, to compare service costs and waiting times, it is necessary to adopt (explicitly or implicitly) a common measure of their impact. The natural choice for this common measure is cost, which then requires estimation of the cost of waiting.

Because of the diversity of waiting-line situations, no single process for estimating the cost of waiting is generally applicable. However, we shall discuss the basic considerations involved for several types of situations.

One broad category is where the customers are *external* to the organization providing the service; i.e., they are *outsiders* bringing their business to the organization. Consider first the case of *profit-making* organizations (typified by the commercial service systems described in Sec. 17.3). From the viewpoint of the decision maker, the cost of waiting probably consists primarily of the *lost profit* from *lost business.* This loss of business may occur immediately (because the customer grows impatient and leaves) or in the future (because the customer is sufficiently irritated that he or she does not come again). This kind of cost is quite difficult to estimate, and it may be necessary to revert to other criteria, such as a tolerable probability distribution of waiting times. When the customer is not a human being, but a job being performed on order, there may be more readily identifiable costs incurred, such as those caused by idle in-process inventories or increased expediting and administrative effort.

Now consider the type of situation where service is provided on a *nonprofit* basis to customers *external* to the organization (typical of social service systems and some transportation service systems described in Sec. 17.3). In this case, the cost of waiting usually is a *social cost* of some kind. Thus, it is necessary to evaluate the consequences of the waiting for the individuals involved and/or for society as a whole and to try to impute a monetary value to avoiding these consequences. Once again, this kind of cost is quite difficult to estimate, and it may be necessary to revert to other criteria.

A situation may be more amenable to estimating waiting costs if the customers are *internal* to the organization providing the service (as for the internal service systems discussed in Sec. 17.3). For example, the customers may be machines (as in Example 1) or employees (as in Example 3) of a firm. Therefore, it may be possible to identify directly some of or all the costs associated with the idleness of these customers. Typically, what

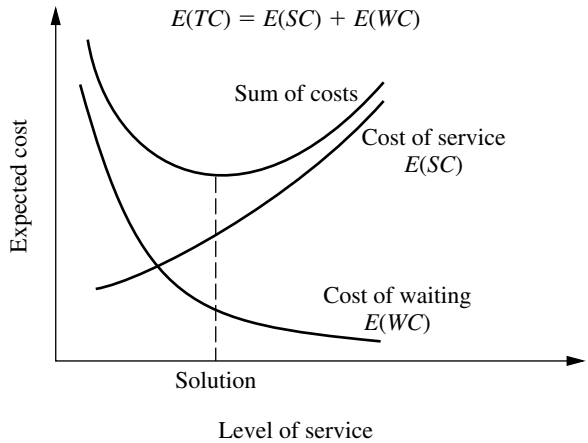


FIGURE 18.4
Conceptual solution procedure for many waiting-line problems.

is being wasted by this idleness is *productive output,* in which case the waiting cost becomes the *lost profit* from *all lost productivity.*

Given that the *cost of waiting* has been evaluated explicitly, the remainder of the analysis is conceptually straightforward. The objective is to determine the level of service that minimizes the total of the expected cost of service and the expected cost of waiting for that service. This concept is depicted in Fig. 18.4, where WC denotes *waiting cost,* SC denotes *service cost,* and TC denotes *total cost.* Thus, the mathematical statement of the objective is to

Minimize $E(\text{TC}) = E(\text{SC}) + E(\text{WC})$.

The next two sections are concerned with the application of this concept to various types of problems. Thus, Sec. 18.3 describes how E(WC) can be expressed mathematically. Section 18.4 then focuses on E(SC) to formulate the overall objective function E(TC) for several basic design problems (including some with multiple decision variables, so that the level-of-service axis in Fig. 18.4 then requires more than one dimension).

18.3 FORMULATION OF WAITING-COST FUNCTIONS

To express E(WC) mathematically, we must first formulate a *waiting-cost function* that describes how the actual waiting cost being incurred varies with the current behavior of the queueing system. The form of this function depends on the context of the individual problem. However, most situations can be represented by one of the two basic forms described next.

The *g*(*N*) Form

Consider first the situation discussed in the preceding section where the queueing system *customers* are *internal* to the organization providing the service, and so the primary cost of waiting may be the *lost profit from lost productivity.* The *rate* at which productive output is lost sometimes is essentially *proportional* to the number of customers in the queue-

ing system. However, in many cases there is not enough productive work available to keep all the members of the calling population continuously busy. Therefore, little productive output may be lost by having just a few members idle, waiting for service in the queueing system, whereas the loss may increase greatly if a few more members are made idle because they require service. Consequently, the primary property of the queueing system that determines the *current rate* at which waiting costs are being incurred is N, the number of customers in the system. Thus, the form of the waiting-cost function for this kind of situation is that illustrated in Fig. 18.5, namely, a function of N. We shall denote this form by $g(N)$.

The $g(N)$ function is constructed for a particular situation by estimating $g(n)$, the waiting-cost rate incurred when $N = n$, for $n = 1, 2, \ldots$, where $g(0) = 0$. After computing the P_n probabilities for a given design of the queueing system, we can calculate

$$E(\text{WC}) = E(g(N)).$$

Because N is a random variable, this calculation is made by using the expression for the expected value of a *function* of a *discrete* random variable

$$E(\text{WC}) = \sum_{n=0}^{\infty} g(n)P_n.$$

The Linear Case. For the special case where $g(N)$ is a *linear function* (i.e., when the waiting cost is proportional to N), then

$$g(N) = C_w N,$$

where C_w is the cost of waiting per unit time for each customer. In this case, $E(\text{WC})$ reduces to

$$E(\text{WC}) = C_w \sum_{n=0}^{\infty} nP_n = C_w L.$$

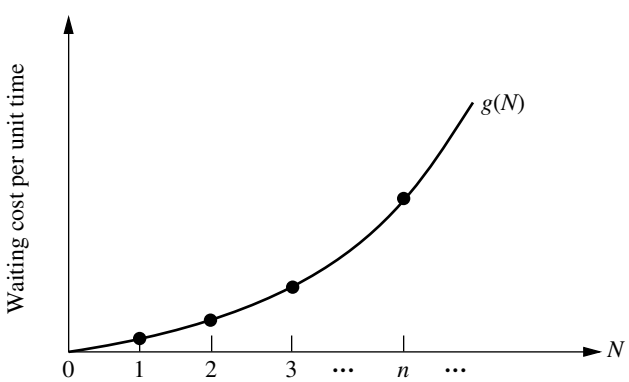


FIGURE 18.5
The waiting-cost function as a function of N.

Example 1—How Many Repairers? For Example 1 of Sec. 18.1, Simulation, Inc., has two standby widget-making machines, so there is no lost productivity as long as the number of customers (machines requiring repair) in the system does not exceed 2. However, for each *additional* customer (up to the maximum of 10 total), the estimated lost profit is \$400 per day. Therefore,

$$g(n) = \begin{cases} 0 & \text{for } n = 0, 1, 2 \\ 400(n-2) & \text{for } n = 3, 4, \ldots, 10, \end{cases}$$

as shown in Table 18.1. Consequently, after calculating the P_n probabilities as described in Sec. 18.1, $E(\text{WC})$ is calculated by summing the rightmost column of Table 18.1 for each of the two cases of interest, namely, having one repairer ($s = 1$) or two repairers ($s = 2$).

The $h(\mathcal{W})$ Form

Now consider the cases discussed in Sec. 18.2 where the queueing system *customers* are *external* to the organization providing the service. Three major types of queueing systems described in Sec. 17.3—commercial service systems, transportation service systems, and social service systems—typically fall into this category. In the case of commercial service systems, the primary cost of waiting may be the lost profit from lost future business. For transportation service systems and social systems, the primary cost of waiting may be in the form of a social cost. However, for either type of cost, its magnitude tends to be affected greatly by the size of the waiting times experienced by the customers. Thus, the primary property of the queueing system that determines the waiting cost currently being incurred is $\mathcal{W}$, the waiting time in the system for the *individual* customers. Consequently, the form of the waiting-cost function for this kind of situation is that illustrated in Fig. 18.6, namely, a function of $\mathcal{W}$. We shall denote this form by $h(\mathcal{W})$.

Note that the example of a $h(\mathcal{W})$ function shown in Fig. 18.6 is a nonlinear function where the slope keeps increasing as $\mathcal{W}$ increases. Although $h(\mathcal{W})$ sometimes is a simple linear function instead, it is fairly common to have this kind of nonlinear function. An in-

TABLE 18.1 Calculation of $E(\text{WC})$ for Example 1

		$s = 1$		$s = 2$	
$N = n$	$g(n)$	P_n	$g(n)P_n$	P_n	$g(n)P_n$
0	0	0.271	0	0.433	0
1	0	0.217	0	0.346	0
2	0	0.173	0	0.139	0
3	400	0.139	56	0.055	24
4	800	0.097	78	0.019	16
5	1,200	0.058	70	0.006	8
6	1,600	0.029	46	0.001	0
7	2,000	0.012	24	3×10^{-4}	0
8	2,400	0.003	7	4×10^{-5}	0
9	2,800	7×10^{-4}	0	4×10^{-6}	0
10	3,200	7×10^{-5}	0	2×10^{-7}	0
$E(\text{WC})$			\$281 per day		\$48 per day

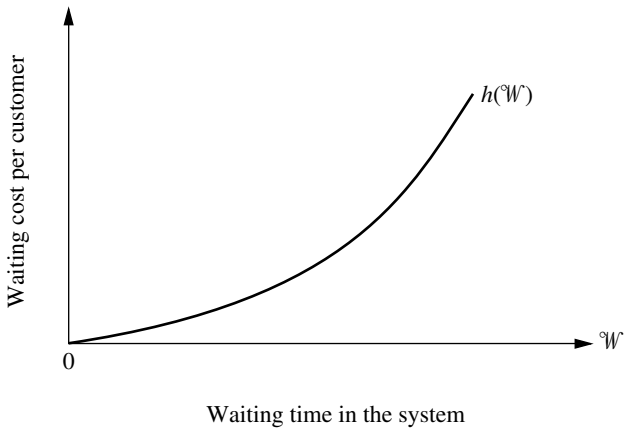


FIGURE 18.6
The waiting-cost function as a function of $\mathcal{W}$.

creasing slope reflects a situation where the *marginal cost* of extending the waiting time keeps increasing. A customer may not mind a "normal" wait of reasonable length, in which case there may be virtually no negative consequences for the organization providing the service in terms of lost profit from lost future business, a social cost, etc. However, if the wait extends even further, the customer may become increasingly exasperated, perhaps even missing deadlines. In such a situation, the negative consequences to the organization may rapidly become relatively severe.

One way of constructing the $h(\mathcal{W})$ function is to estimate $h(w)$ (the waiting cost incurred when a customer's waiting time $\mathcal{W} = w$) for several values of w and then to fit a polynomial to these points. The expectation of this function of a continuous random variable is then defined as

$$E(h(\mathcal{W})) = \int_0^\infty h(w) f_{\mathcal{W}}(w)\, dw,$$

where $f_{\mathcal{W}}(w)$ is the probability density function of $\mathcal{W}$. However, because $E(h(\mathcal{W}))$ is the expected waiting cost *per customer* and $E(\text{WC})$ is the expected waiting cost *per unit time,* these two quantities are not equal in this case. To relate them, it is necessary to multiply $E(h(\mathcal{W}))$ by the expected *number of customers per unit time* entering the queueing system. In particular, if the mean arrival rate is a constant λ, then

$$E(\text{WC}) = \lambda E(h(\mathcal{W})) = \lambda \int_0^\infty h(w) f_{\mathcal{W}}(w)\, dw.$$

Example 2—Which Computer? Because the faculty and students of Emerald University would experience different turnaround times with the two computers under consideration (see Sec. 18.1), the choice between the computers required an evaluation of the consequences of making them wait for their jobs to be run. Therefore, several leading scientists on the faculty were asked to evaluate these consequences.

The scientists agreed that one major consequence is a *delay in getting research done.* Little effective progress can be made while one is awaiting the results from a computer

run. The scientists estimated that it would be worth \$500 to reduce this delay by a day. Therefore, this component of waiting cost was estimated to be \$500 per day, that is, $500\mathcal{W}$, where $\mathcal{W}$ is expressed in days.

The scientists also pointed out that a second major consequence of waiting is a *break in the continuity of the research.* Although a short delay (a fraction of a day) causes little problem in this regard, a longer delay causes significant wasted time in having to gear up to resume the research. The scientists estimated that this wasted time would be roughly proportional to the *square* of the delay time. Dollar figures of \$100 and \$400 were then imputed to the value of being able to avoid this consequence entirely rather than having a wait of $\frac{1}{2}$ day and 1 day, respectively. Therefore, this component of the waiting cost was estimated to be $400\mathcal{W}^2$.

This analysis yields

$$h(\mathcal{W}) = 500\mathcal{W} + 400\mathcal{W}^2.$$

Because

$$f_W(w) = \mu(1 - \rho)e^{-\mu(1-\rho)w}$$

for the *M*/*M*/1 model (see Sec. 17.6) fitting this single-server queueing system,

$$E(h(\mathcal{W})) = \int_0^\infty (500w + 400w^2)\mu(1 - \rho)e^{-\mu(1-\rho)w}\, dw,$$

where $\rho = \lambda/\mu$ for a single-server system. Since $\mu(1 - \rho) = (\mu - \lambda)$, the values of μ and λ presented in Sec. 18.1 give

$$\mu(1 - \rho) = \begin{cases} 10 & \text{for MBI computer} \\ 5 & \text{for CRAB computer.} \end{cases}$$

Evaluating the integral for these two cases yields

$$E(h(\mathcal{W})) = \begin{cases} 58 & \text{for MBI computer} \\ 132 & \text{for CRAB computer.} \end{cases}$$

The result represents the expected waiting cost (in dollars) for each person arriving with a job to be run. Because $\lambda = 20$, the total expected waiting cost per day becomes

$$E(\text{WC}) = \begin{cases} \$1{,}160 \text{ per day} & \text{for MBI computer} \\ \$2{,}640 \text{ per day} & \text{for CRAB computer.} \end{cases}$$

The Linear Case. Before turning to the next example, consider now the special case where $h(\mathcal{W})$ is a linear function,

$$h(\mathcal{W}) = C_w\mathcal{W},$$

where C_w is the cost of waiting per unit time for each customer. In this case, $E(\text{WC})$ reduces to

$$E(\text{WC}) = \lambda E(C_w\mathcal{W}) = C_w(\lambda W) = C_wL.$$

Note that this result is identical to the result when $g(N)$ is a linear function. Consequently, when the total waiting cost incurred by the queueing system is simply *proportional* to the

total waiting time, it does not matter whether the $g(N)$ or the $h(W)$ form is used for the waiting-cost function.

Example 3—How Many Tool Cribs? As indicated in Sec. 18.1, the value to the Mechanical Company of a busy mechanic's output averages about \$48 per hour. Thus, $C_w = 48$. Consequently, for each tool crib the expected waiting cost per hour is

$$E(\text{WC}) = 48L,$$

where L represents the expected number of mechanics waiting (or being served) at the tool crib.

18.4 DECISION MODELS

We mentioned in Sec. 18.2 that three common decision variables in designing queueing systems are s (number of servers), μ (mean service rate for each server), and λ (mean arrival rate at each service facility). We shall now formulate models for making some of these decisions.

Model 1—Unknown *s*

Model 1 is designed for the case where both μ and λ are fixed at a particular service facility, but where a decision must be made on the number of servers to have on duty at the facility.

Formulation of Model 1.

Definition:	C_s = marginal cost of a server per unit time.	
Given:	μ, λ, C_s.	
To find:	s.	
Objective:	Minimize	$E(\text{TC}) = C_s s + E(\text{WC})$.

Because only a few alternative values of s normally need to be considered, the usual way of solving this model is to calculate $E(\text{TC})$ for these values of s and select the minimizing one. For the linear case where $E(\text{WC}) = C_w L$, an Excel template has been provided in your OR Courseware for performing these calculations when the queueing system fits the $M/M/s$ queueing model. However, as long as the queueing model is tractable, it often is not very difficult to perform these calculations yourself for other cases, as illustrated by the following example.

Example 1—How Many Repairers? For Example 1 of Sec. 18.1, each repairer (server) costs Simulation, Inc., approximately \$280 per day. Thus, with 1 day as the unit of time, $C_s = 280$. Using the values of $E(\text{WC})$ calculated in Table 18.1 then yields the results shown in Table 18.2, which indicate that the company should continue having just one repairer.

Model 2—Unknown μ and *s*

Model 2 is designed for the case where both the efficiency of service, measured by μ, and the number of servers s at a service facility need to be selected.

TABLE 18.2 Calculation of $E(TC)$ in dollars per day for Example 1

s	$C_s s$	$E(WC)$	$E(TC)$
1	\$280	\$281	\$561 per day ← minimum
2	\$560	\$ 48	\$608 per day
≥3	≥\$840	≥\$ 0	≥\$840 per day

Alternative values of μ may be available because there is a choice on the *quality* of the servers. For example, when the servers will be materials-handling units, the quality of the units to be purchased affects their service rate for moving loads.

Another possibility is that the *speed* of the servers can be adjusted mechanically. For example, the speed of machines frequently can be adjusted by changing the amount of power consumed, which also changes the cost of operation.

Still another type of example is the selection of the number of crews (the servers) and the size of each crew (which determines μ) for jointly performing a certain task. The task might be maintenance work, or loading and unloading operations, or inspection work, or setup of machines, and so forth.

In many cases, only a few alternative values of μ are available, e.g., the efficiency of the alternative types of materials-handling equipment or the efficiency of the alternative crew sizes.

Formulation of Model 2.

Definitions: $f(\mu)$ = marginal cost of server per unit time when mean service rate is μ.
A = set of feasible values of μ.
Given: $\lambda, f(\mu), A$.
To find: μ, s.
Objective: Minimize $E(\text{TC}) = f(\mu)s + E(\text{WC})$, subject to $\mu \in A$.

Example 2—Which Computer? As indicated in Sec. 18.1, $\mu = 30$ for the MBI computer and $\mu = 25$ for the CRAB computer, where 1 day is the unit of time. These computers are the only two being considered by Emerald University, so

$$A = \{25, 30\}.$$

Because the leasing cost per day is \$3,750 for the CRAB computer ($\mu = 25$) and \$5,000 for the MBI computer ($\mu = 30$),

$$f(\mu) = \begin{cases} 3{,}750 & \text{for } \mu = 25 \\ 5{,}000 & \text{for } \mu = 30. \end{cases}$$

The supercomputer chosen will be the only one available to the faculty and students, so the number of servers (supercomputers) for this queueing system is restricted to $s = 1$. Hence,

$$E(\text{TC}) = f(\mu) + E(\text{WC}),$$

where $E(\text{WC})$ is given in Sec. 18.3 for the two alternatives. Thus,

$$E(\text{TC}) = \begin{cases} 3{,}750 + 2{,}640 = \$6{,}390 \text{ per day} & \text{for CRAB computer} \\ 5{,}000 + 1{,}160 = \$6{,}160 \text{ per day} & \text{for MBI computer.} \end{cases}$$

Consequently, the decision was made to lease the MBI supercomputer.

The Application of Model 2 to Other Situations. This example illustrates a case where the number of feasible values of μ is *finite* but the value of s is fixed. If s were not fixed, a two-stage approach could be used to solve such a problem. First, for each individual value of μ, set $C_s = f(\mu)$, and solve for the value of s that minimizes $E(\text{TC})$ for model 1. Second, compare these minimum $E(\text{TC})$ for the alternative values of μ, and select the one giving the overall minimum.

When the number of feasible values of μ is *infinite* (such as when the speed of a machine or piece of equipment is set mechanically within some feasible interval), another two-stage approach sometimes can be used to solve the problem. First, for each individual value of s, *analytically* solve for the value of μ that minimizes $E(\text{TC})$. [This approach requires setting to zero the derivative of $E(\text{TC})$ with respect to μ and then solving this equation for μ, which can be done only when analytical expressions are available for both $f(\mu)$ and $E(\text{WC})$.] Second, compare these minimum $E(\text{TC})$ for the alternative values of s, and select the one giving the overall minimum.

This analytical approach frequently is relatively straightforward for the case of $s = 1$ (see Prob. 18.4-17). However, because far fewer and less convenient analytical results are available for multiple-server versions of queueing models, this approach is either difficult (requiring computer calculations with numerical methods to solve the equation for μ) or completely impossible when $s > 1$. Therefore, a more practical approach is to consider only a relatively small number of representative values of μ and to use available tabulated results for the appropriate queueing model to obtain (or approximate) $E(\text{TC})$ for these μ values.

A Special Result with Model 2. Fortunately, under certain fairly common circumstances described next, $s = 1$ (and its minimizing value of μ) *must* yield the overall minimum $E(\text{TC})$ for model 2, so $s > 1$ cases need not be considered at all.

Optimality of a Single Server. Under certain conditions, $s = 1$ necessarily is *optimal* for model 2.

The primary conditions[1] are that

1. The value of μ minimizing $E(\text{TC})$ for $s = 1$ is feasible.
2. Function $f(\mu)$ is either *linear* or *concave* (as defined in Appendix 2).

In effect, this optimality result indicates that it is better to concentrate service capacity into one fast server rather than dispersing it among several slow servers. Condition 2 says that this concentrating of a given amount of service capacity can be done without increasing the cost of service. Condition 1 says that it must be possible to make μ sufficiently large that a single server can be used to full advantage.

[1]There also are minor restrictions on the queueing model and the waiting-cost function. However, any of the constant service-rate queueing models presented in Chap. 17 for $s \geq 1$ are allowed. If the $g(N)$ form is used for the waiting-cost function, it can be any *increasing* function. If the $h(W)$ form is used, it can be any linear function or any convex function (as defined in Appendix 2), which fits most cases of interest.

TABLE 18.3 Comparison of service efficiency for Model 2 solutions

	Mean Rate of Service Completions
$N = n$	$(s, \mu) = (s^*, \mu^*)$ versus $(s, \mu) = (1, s^*\mu^*)$
$n = 0$	$0 = 0$
$n = 1, 2, \ldots, s^* - 1$	$n\mu^* < s^*\mu^*$
$n \geq s^*$	$s^*\mu^* = s^*\mu^*$

To understand why this result holds, consider any other solution to model 2, $(s, \mu) = (s^*, \mu^*)$, where $s^* > 1$. The service capacity of this system (as measured by the mean rate of service completions when all servers are working) is $s^*\mu^*$. We shall now compare this solution with the corresponding single-server solution $(s, \mu) = (1, s^*\mu^*)$ having the *same* service capacity. In particular, Table 18.3 compares the mean rate at which service completions occur for each given number of customers in the system $N = n$. This table shows that the service efficiency of the (s^*, μ^*) solution sometimes is worse but never is better than for the $(1, s^*\mu^*)$ solution because it can use the full service capacity only when there are at least s^* customers in the system, whereas the single-server solution uses the full capacity whenever there are *any* customers in the system. Because this lower service efficiency can only increase waiting in the system, $E(\text{WC})$ must be larger for (s^*, μ^*) than for $(1, s^*\mu^*)$. Furthermore, the expected service cost must be at least as large because condition 2 [and $f(0) = 0$] implies that

$$f(\mu^*)s \geq f(s^*\mu^*).$$

Therefore, $E(\text{TC})$ is larger for (s^*, μ^*) than $(1, s^*\mu^*)$. Finally, note that condition 1 implies that there is a feasible solution with $s = 1$ that is at least as good as $(1, s^*\mu^*)$. The conclusion is that *any* $s > 1$ solution *cannot* be optimal for model 2, so $s = 1$ must be optimal.[1]

This result is still of some use even when one or both conditions fail to hold. If μ cannot be made sufficiently large to permit a single server, it still suggests that a *few* fast servers should be preferred to many slow ones. If condition 2 does not hold, we still know that $E(\text{WC})$ is minimized by concentrating any given amount of service capacity into a single server, so the best $s = 1$ solution must be at least nearly optimal unless it causes a *substantial* increase in service cost.

Model 3—Unknown λ and s

Model 3 is designed especially for the case where it is necessary to select both the *number of service facilities* and the *number of servers s* at each facility. In the typical situation, a population (such as the employees in an industrial building) must be provided with a certain service, so a decision must be made as to what proportion of the population (and therefore what value of λ) should be assigned to each service facility. Examples of such facilities include employee facilities (drinking fountains, vending machines, and rest-

[1]For a rigorous proof of this result, see S. Stidham, Jr., "On the Optimality of Single-Server Queueing Systems," *Operations Research,* **18:** 708–732, 1970.

rooms), storage facilities, and reproduction equipment facilities. It may sometimes be clear that only a single server should be provided at each facility (e.g., one drinking fountain or one copy machine), but s often is also a decision variable.

To simplify our presentation, we shall require in model 3 that λ and s be the same for all service facilities. However, it should be recognized that a slight improvement in the indicated solution might be achieved by permitting minor deviations in these parameters at individual facilities. This should be investigated as part of the detailed analysis that generally follows the application of the mathematical model.

Formulation of Model 3.

Definitions:	C_s = marginal cost of server per unit time.
	C_f = fixed cost of service per service facility per unit time.
	λ_p = mean arrival rate for entire calling population.
	n = number of service facilities = λ_p/λ.
Given:	μ, C_s, C_f, λ_p.
To find:	λ, s.
Objective:	Minimize E(TC), subject to $\lambda = \lambda_p/n$, where $n = 1, 2, \ldots$.

Finding *E*(TC). It might appear at first glance that the appropriate expression for the expected total cost per unit time of all the facilities should be

$$E(\text{TC}) \stackrel{?}{=} n[(C_f + C_s s) + E(\text{WC})],$$

where E(WC) here represents the expected waiting cost per unit time for *each* facility. However, if this expression actually were valid, it would imply that $n = 1$ *necessarily* is optimal for model 3. The reasoning is completely analogous to that for the optimality of a single-server result for model 2; namely, any solution $(n, s) = (n^*, s^*)$ with $n^* > 1$ has higher service costs than the $(n, s) = (1, n^*s^*)$ solution, and it *also* has a higher expected waiting cost because it sometimes makes less effective use of the available service capacity. In particular, it sometimes has idle servers at one facility while customers are waiting at another facility, so the mean rate of service completions would be less than if the customers had access to *all* the servers at one common facility.

Because there are many situations where it obviously would *not* be optimal to have just one service facility (e.g., the number of restrooms in a 50-story building), something must be wrong with this expression. Its deficiency is that it considers only the cost of service and the cost of waiting *at the service facilities* while totally ignoring the cost of the time wasted in *traveling* to and from the facilities. Because travel time would be prohibitive with only one service facility for a large population, enough separate facilities must be distributed throughout the calling population to hold travel time down to a reasonable level.

Thus, letting the random variable T be the round-trip travel time for a customer coming to and going back from one of the service facilities, we see that the total time lost by the customer actually is $\mathcal{W} + T$. (Recall from Chap. 17 that $\mathcal{W}$ is the waiting time in the queueing system *after* the customer arrives.) Therefore, a customer's *total* cost for time lost should be based on $\mathcal{W} + T$ rather than just $\mathcal{W}$. To simplify the analysis, let us separate this total cost into the sum of the waiting-time cost based on $\mathcal{W}$ (or N) and the travel-time cost based on T. We shall also assume that the travel-time cost is proportional to T,

where C_t is the cost of each unit of travel time for each customer. For ease of presentation, suppose that the probability distribution of T is the same for each service facility, so that $C_tE(T)$ is the *expected travel cost* for each arrival at any of the service facilities. The resulting expression for $E(\text{TC})$ is

$$E(\text{TC}) = n[(C_f + C_s s) + E(\text{WC}) + \lambda C_t E(T)]$$

because λ is the expected number of arrivals *per unit time* at each facility. Consequently, by evaluating (or estimating) $E(T)$ for each case of interest, model 3 can be solved by calculating $E(\text{TC})$ for various values of s for each n and then selecting the solution giving the overall minimum.

Example 3—How Many Tool Cribs? For the new plant being designed for the Mechanical Company (see Sec. 18.1), the layout of the portion of the factory area where the mechanics will work is shown in Fig. 18.7. The three possible locations for tool cribs are identified as locations 1, 2, and 3, where access to these locations will be provided by a system of orthogonal aisles parallel to the sides of the indicated area. The coordinates are given in units of feet.

The three basic alternatives being considered are these:

Alternative 1: Have one tool crib—use location 2.
Alternative 2: Have two tool cribs—use locations 1 and 3.
Alternative 3: Have three tool cribs—use locations 1, 2, and 3.

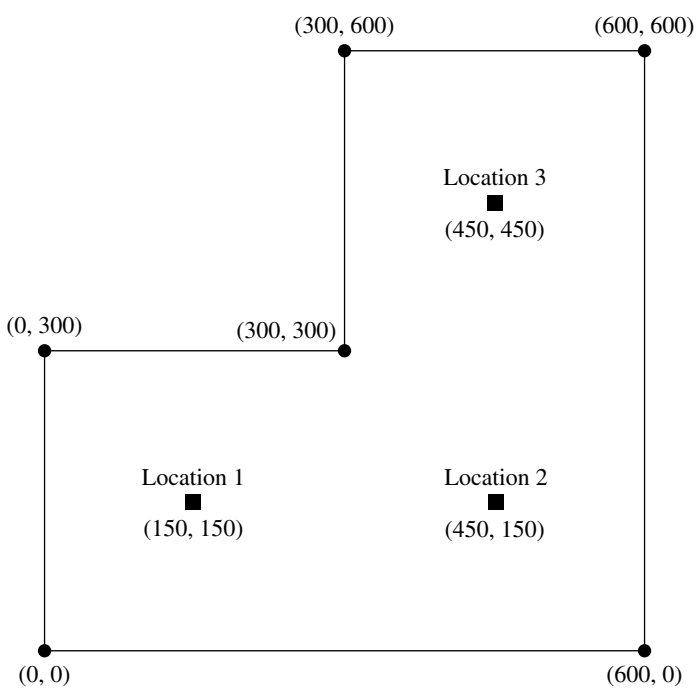


FIGURE 18.7
Layout for Example 3.

TABLE 18.4 Calculation of $E(TC)$, in dollars per hour for Example 3

n	λ	s	L	$E(T)$	$C_f + C_s s$	$E(WC)$	$\lambda C_t E(T)$	$E(TC)$
1	120	1	∞	0.04	\$36	∞	\$230.40	∞
1	120	2	1.333	0.04	\$56	\$64.00	\$230.40	\$350.40
1	120	3	1.044	0.04	\$76	\$50.11	\$230.40	\$356.51
2	60	1	1.000	0.0278	\$36	\$48.00	\$ 80.00	\$328.00
2	60	2	0.534	0.0278	\$56	\$25.63	\$ 80.00	\$323.26
3	40	1	0.500	0.02	\$36	\$24.00	\$ 38.40	\$295.20
3	40	2	0.344	0.02	\$56	\$16.51	\$ 38.40	\$332.73

The mechanics will be distributed quite uniformly throughout the area shown, and each mechanic will be assigned to the *nearest* tool crib. It is estimated that the mechanics will walk to and from a tool crib at an average speed of slightly less than 3 miles per hour. Based on this information and an estimate of the average distance traveled on each trip to and from the tool crib, $E(T)$ is estimated to be 0.04, 0.0278, and 0.02 hour for alternatives 1, 2, and 3, respectively. [A supplement to this chapter on the CD-ROM discusses the evaluation of travel time and also spells out how these particular values of $E(T)$ were obtained for this example.]

The stage now is set for using model 3 to choose from these alternatives. Most of the data required for this model are given in Sec. 18.1, namely,

$$\mu = 120 \text{ per hour}, \qquad C_f = \$16 \text{ per hour},$$
$$C_s = \$20 \text{ per hour},$$
$$\lambda_p = 120 \text{ per hour}, \qquad C_t = \$48 \text{ per hour},$$

where the $M/M/s$ model given in Sec. 17.6 is used to calculate L and so on. In addition, the end of Sec. 18.3 gives $E(WC) = 48L$ in dollars per hour. Therefore,

$$E(TC) = n\left[(16 + 20s) + 48L + \frac{120}{n}\,48E(T)\right].$$

The resulting calculation of $E(TC)$ for various s values for each n is given in Table 18.4, which indicates that the *overall minimum* $E(TC)$ of \$295.20 per hour is obtained by having three tool cribs (so $\lambda = 40$ for each), with one clerk at each tool crib.

18.5 SOME AWARD-WINNING APPLICATIONS OF QUEUEING THEORY

The prestigious *Franz Edelman Awards for Management Science Achievement* are awarded annually by the Institute of Operations Research and Management Sciences (INFORMS) for the year's best applications of OR. A rather substantial number of these awards have been given for innovative applications of queueing theory. We briefly describe some of these applications in this section.

One of the early first-prize winners (described in the November 1975 issue, Part 2, of *Interfaces*) was the *Xerox Corporation.* The company had recently introduced a major new duplicating system that was proving to be particularly valuable for its owners. Consequently, these customers were demanding that Xerox's tech reps reduce the waiting

times to repair the machines. An OR team then applied queueing theory to study how to best meet the new service requirements. This resulted in replacing the previous one-person tech rep territories by larger three-person tech rep territories. This change had the dramatic effect of both substantially reducing the average waiting times of the customers and increasing the utilization of the tech reps by over 50 percent.

In Sec. 3.5, we described an award-winning application by *United Airlines* (January 1986 issue of *Interfaces*) that resulted in annual savings of over $6 million. This application involved scheduling the work assignments of United's 4,000 reservations sales representatives and support personnel at its 11 reservations offices and the 1,000 customer service agents at its 10 largest airports. After determining how many employees are needed at each location during each half hour of the week, we discussed how linear programming was applied to design the work schedules for all the employees to meet these service requirements most efficiently. However, we never mentioned how these service requirements on the number of employees needed each half hour were determined.

We now are in a position to point out that these service requirements were determined by applying *queueing theory.* Each specific location (e.g., the check-in counters at an airport) constitutes a queueing system with the employees as the servers. After forecasting the mean arrival rate during each half hour of the week, queueing models are used to find the minimum number of servers that will provide satisfactory measures of performance for the queueing system.

L.L. Bean, Inc., the large telemarketer and mail-order catalog house, relied mainly on queueing theory for its award-winning study of how to allocate its telecommunications resources. (The article describing this study is in the January 1991 issue of *Interfaces,* and other articles giving additional information are in the November 1989 and March–April 1993 issues of this journal.) The telephone calls coming in to its call center to place orders are the customers in a large queueing system, with the telephone agents as the servers. The key questions being asked during the study were the following.

1. How many telephone trunk lines should be provided for incoming calls to the call center?
2. How many telephone agents should be scheduled at various times?
3. How many hold positions should be provided for customers waiting for a telephone agent? (Note that the limited number of hold positions causes the system to have a finite queue.)

For each interesting combination of these three quantities, queueing models provide the measures of performance of the queueing system. Given these measures, the OR team carefully assessed the cost of lost sales due to making some customers either incur a busy signal or be placed on hold too long. By adding the cost of the telemarketing resources, the team then was able to find the combination of the three quantities that minimizes the expected total cost. This resulted in cost savings of $9 to $10 million per year.

New York City has a long-standing tradition of using OR techniques in planning and operating many of its complex urban service systems. Starting in the late 1960s, award-winning studies involving queueing theory have been conducted for its Fire Department and its Police Department. (Fires and police emergencies are the customers in these respective queueing systems.) Subsequently, major OR studies (including several more involving queueing theory) have been conducted for its Department of Sanitation, Depart-

ment of Transportation, Department of Health and Hospitals, Department of Environmental Protection, Office of Management and Budget, and Department of Probation. Because of the success of these studies, many of these departments now have their own in-house OR groups.

The award-winning study in New York City that we will describe here involves its arrest-to-arraignment system. This system consists of the process from when individuals are arrested until they are arraigned (the first court appearance before an arraignment judge, who determines whether there was probable cause for the arrest). Before the study, the city's arrestees (the customers in a queueing system) were in custody waiting to be arraigned for an average of 40 hours (occasionally more than 70 hours). These waiting times were considered excessive, because the arrestees were being held in crowded, noisy conditions that were emotionally stressful, unhealthy, and often physically dangerous. Therefore, a 2-year OR study was conducted to overhaul the system. Both queueing theory and simulation (the subject of Chap. 22) were used. This led to sweeping operational and policy changes that simultaneously reduced average waiting times until arraignment to 24 hours or less and provided annual savings of $9.5 million. (See the January 1993 issue of *Interfaces* for details.)

The first prize in the 1993 competition was won by *AT&T* for a study that (like the preceding one) also combined the use of queueing theory and simulation (January–February 1994 issue of *Interfaces*). The queueing models are of both AT&T's telecommunication network and the call center environment for the typical business customers of AT&T that have such a center. The purpose of the study was to develop a user-friendly PC-based system that AT&T's business customers can use to guide them in how to design or redesign their call centers. Since call centers comprise one of the United States' fastest-growing industries, this system had been used about 2,000 times by AT&T's business customers by 1992. This resulted in more than $750 million in annual profit for these customers.

KeyCorp is one of the largest bank holding companies in the United States, with more than 1,300 branches and over 6,000 tellers. This company's award-winning OR study (January 1996 issue of *Interfaces*) focused on using queueing theory to improve the performance of each branch's queueing system where the tellers serve the customers. This resulted in developing a companywide service excellence management system (SEMS). A key part of SEMS is a performance capture system that collects data on a continuous basis for each discrete component of each teller transaction in a completely automated process. This system enables SEMS to measure branch activities and generate reports on customer waiting times, teller proficiency, and productivity levels. These reports help managers schedule tellers to better match customer arrivals. They also identify opportunities for enhancing the productivity and service provided by the tellers by redesigning the service process and providing performance standards. These efforts led to a huge 53 percent reduction in the average service times, a dramatic improvement in customer waiting times, and a major increase in the level of customer satisfaction. At the same time, SEMS is expected to reduce personnel expenses by $98 million over 5 years.

There have been many other award-winning applications of queueing theory, as well as numerous additional articles describing other successful applications. However, the several examples presented in this section hopefully have given you a feeling for the kinds of applications that are occurring and for the impact they sometimes have.

18.6 CONCLUSIONS

This chapter has discussed the application of queueing theory for *designing* queueing systems. Every individual problem has its own special characteristics, so no standard procedure can be prescribed to fit every situation. Therefore, the emphasis has been on introducing fundamental considerations and approaches that can be adapted to most cases. We have focused on three particularly common decision variables (s, μ, and λ) as a vehicle for introducing and illustrating these concepts. However, there are many other possible decision variables (e.g., the size of a waiting room for a queueing system) and many more complicated situations (e.g., designing a *priority* queueing system) that can also be analyzed in a similar way.

Another useful area for the application of queueing theory is the development of policies for *controlling* queueing systems, e.g., for *dynamically* adjusting the number of servers or the service rate to compensate for changes in the number of customers in the system. Research is being conducted in this area.

Queueing theory has proved to be a very useful tool, and we anticipate that its use will continue to grow as recognition of the many guises of queueing systems grows.

SELECTED REFERENCES

1. Allen, A. O.: *Probability, Statistics, and Queueing Theory with Computer Science Applications,* 2d ed., Academic Press, New York, 1990, chaps. 5–6.
2. Hall, R. W.: *Queueing Methods: For Services and Manufacturing,* Prentice-Hall, Englewood Cliffs, NJ, 1991.
3. Hillier, F. S., M. S. Hillier, and G. J. Lieberman: *Introduction to Management Science: A Modeling and Case Studies Approach with Spreadsheets,* Irwin/McGraw-Hill, Burr Ridge, IL, 2000, chap. 14.
4. Kleinrock, L.: *Queueing Systems, vol. II: Computer Applications,* Wiley, New York, 1976.
5. Lee, A. M.: *Applied Queueing Theory,* St. Martin's Press, New York, 1966.
6. Newell, G. F.: *Applications of Queueing Theory,* 2d ed., Chapman and Hall, London, 1982.
7. Papadopoulos, H. T., C. Heavey, and J. Browne: *Queueing Theory in Manufacturing Systems Analysis and Design,* Chapman and Hall, London, 1993.

LEARNING AIDS FOR THIS CHAPTER IN YOUR OR COURSEWARE

"Ch. 18—Application of QT" Excel File:

Same templates as listed at the end of Chap. 17, plus
Template for *M*/*M*/*s* Economic Analysis of Number of Servers

Supplement to This Chapter:

The Evaluation of Travel Time (appears on the book's website, www.mhhe.com/hillier).

See Appendix 1 for documentation of the software.

PROBLEMS

To the left of each of the following problems (or their parts), we have inserted a T whenever one of the templates for this chapter (and the preceding chapter) can be useful. An asterisk on the problem number indicates that at least a partial answer is given in the back of the book.

18.2-1. For each kind of queueing system listed in Prob. 17.3-1, briefly describe the nature of the *cost of service* and the *cost of waiting* that would need to be considered in designing the system.

18.3-1.* Suppose that a queueing system fits the *M*/*M*/1 model described in Sec. 17.6, with $\lambda = 2$ and $\mu = 4$. Evaluate the expected waiting cost per unit time $E(\text{WC})$ for this system when its waiting-cost function has the form

(a) $g(N) = 10N + 2N^2$.

(b) $h(\mathcal{W}) = 25\mathcal{W} + \mathcal{W}^3$.

18.3-2. Follow the instructions of Prob. 18.3-1 for the following waiting-cost functions.

(a) $g(N) = \begin{cases} 10N & \text{for } N = 0, 1, 2 \\ 6N^2 & \text{for } N = 3, 4, 5 \\ N^3 & \text{for } N > 5. \end{cases}$

(b) $h(\mathcal{W}) = \begin{cases} \mathcal{W} & \text{for } 0 \le \mathcal{W} \le 1 \\ \mathcal{W}^2 & \text{for } \mathcal{W} \ge 1. \end{cases}$

T **18.4-1.** Section 18.3 indicates that a linear waiting-cost function yields $E(\text{WC}) = C_w L$, where C_w is the cost of waiting per unit time for each customer. In this case, the objective for decision model 1 in Sec. 18.4 is to minimize $E(\text{TC}) = C_s s + C_w L$. The purpose of this problem is to enable you to explore the effect that the relative sizes of C_s and C_w have on the optimal number of servers.

Suppose that the queueing system under consideration fits the *M*/*M*/*s* model with $\lambda = 8$ customers per hour and $\mu = 10$ customers per hour. Use the Excel template in your OR Courseware for economic analysis with the *M*/*M*/*s* model to find the optimal number of servers for each of the following cases.

(a) $C_s = \$100$ and $C_w = \$10$.

(b) $C_s = \$100$ and $C_w = \$100$.

(c) $C_s = \$10$ and $C_w = \$100$.

T **18.4-2.*** Jim McDonald, manager of the fast-food hamburger restaurant McBurger, realizes that providing fast service is a key to the success of the restaurant. Customers who have to wait very long are likely to go to one of the other fast-food restaurants in town next time. He estimates that each minute a customer has to wait in line before completing service costs him an average of 30 cents in lost future business. Therefore, he wants to be sure that enough cash registers always are open to keep waiting to a minimum. Each cash register is operated by a part-time employee who obtains the food ordered by each customer and collects the payment. The total cost for each such employee is $9 per hour.

During lunch time, customers arrive according to a Poisson process at a mean rate of 66 per hour. The time needed to serve a customer is estimated to have an exponential distribution with a mean of 2 minutes.

Determine how many cash registers Jim should have open during lunch time to minimize his expected total cost per hour.

T **18.4-3.** The Garrett-Tompkins Company provides three copy machines in its copying room for the use of its employees. However, due to recent complaints about considerable time being wasted waiting for a copier to become free, management is considering adding one or more additional copy machines.

During the 2,000 working hours per year, employees arrive at the copying room according to a Poisson process at a mean rate of 30 per hour. The time each employee needs with a copy machine is believed to have an exponential distribution with a mean of 5 minutes. The lost productivity due to an employee spending time in the copying room is estimated to cost the company an average of $25 per hour. Each copy machine is leased for $3,000 per year.

Determine how many copy machines the company should have to minimize its expected total cost per hour.

18.4-4. A certain queueing system has a Poisson input, with a mean arrival rate of 4 customers per hour. The service-time distribution is exponential, with a mean of 0.2 hour. The marginal cost of providing each server is $20 per hour, where it is estimated that the cost that is incurred by having each customer *idle* (i.e., in the queueing system) is $120 per hour for the first customer and $180 per hour for each additional customer. Determine the number of servers that should be assigned to the system to minimize the expected total cost per hour. [*Hint:* Express $E(\text{WC})$ in terms of L, P_0, and ρ, and then use Figs. 17.6 and 17.7.]

18.4-5.* Reconsider Prob. 17.6-9. The total compensation for the new employee would be $8 per hour, which is just half that for the cashier. It is estimated that the grocery store incurs lost profit due to lost future business of $0.08 for each minute that each customer has to wait (including service time). The manager now wants to determine on an expected total cost basis whether it would be worthwhile to hire the new person.

(a) Which decision model presented in Sec. 18.4 applies to this problem? Why?

(b) Use this model to determine whether to continue the status quo or to adopt the proposal.

18.4-6. Customers arrive at a fast-food restaurant with one server according to a Poisson process at a mean rate of 30 per hour. The server has just resigned, and the two candidates for the replacement are *X* (fast but expensive) and *Y* (slow but inexpensive). Both candidates would have an exponential distribution for service times, with *X* having a mean of 1.2 minutes and *Y* having a mean of 1.5 minutes. Restaurant revenue per month is given by \$6,000/*W,* where *W* is the expected waiting time (in minutes) of a customer in the system.

Determine the upper bound on the difference in their monthly compensations that would justify hiring *X* rather than *Y.*

18.4-7. Jerry Jansen, Materials Handling Manager at the Casper-Edison Corporation's new factory, needs to make a purchasing decision. He needs to choose between two types of materials-handling equipment, a small tractor-trailer train and a heavy-duty forklift truck, for transporting heavy goods between certain producing centers in the factory. Calls for the materials-handling unit to move a load occur according to a Poisson process at a mean rate of 4 per hour. The total time required to move a load has an exponential distribution, where the expected time would be 12 minutes for the tractor-trailer train and 9 minutes for the forklift truck. The total equivalent uniform hourly cost (capital recovery cost plus operating cost) would be \$50 for the tractor-trailer train and \$150 for the forklift truck. The estimated cost of idle goods (waiting to be moved or in transit) because of increased in-process inventory is \$20 per load per hour.

Jerry also has established certain criteria that he would like the materials-handling unit to satisfy in order to keep production flowing on schedule as much as possible. He would like to average no more than half an hour for completing the move of a load after receiving the call requesting the move. He also would like the time for completing the move to be no more than 1 hour 80 percent of the time. Finally, he would like to have no more than three loads waiting to start their move at least 80 percent of the time.

T **(a)** Obtain the various measures of performance if the tractor-trailer train were to be chosen. Evaluate how well these measures meet the above criteria.

T **(b)** Repeat part (*a*) if the forklift truck were to be chosen.

(c) Compare the two alternatives in terms of their expected total cost per hour (including the cost of idle goods).

(d) Which alternative do you think Jerry should choose?

18.4-8. The Southern Railroad Company has been subcontracting for the painting of its railroad cars as needed. However, management has decided that the company can save money by doing this work itself. A decision now needs to be made to choose between two alternative ways of doing this.

Alternative 1 is to provide two paint shops, where painting is done by hand (one car at a time in each shop), for a total hourly cost of \$70. The painting time for a car would be 6 hours. Alternative 2 is to provide one spray shop involving an hourly cost of \$100. In this case, the painting time for a car (again done one at a time) would be 3 hours. For both alternatives, the cars arrive according to a Poisson process with a mean rate of 1 every 5 hours. The cost of idle time per car is \$100 per hour.

(a) Use Fig. 17.11 to estimate L, L_q, W, and W_q for Alternative 1.

(b) Find these same measures of performance for Alternative 2.

(c) Determine and compare the expected total cost per hour for these alternatives.

18.4-9. An airline maintenance base wants to make a change in its overhaul operation. The present situation is that only one airplane can be repaired at a time, and the expected repair time is 36 hours, whereas the expected time between arrivals is 45 hours. This situation has led to frequent and prolonged delays in repairing incoming planes, even though the base operates continuously. The average cost of an idle plane to the airline is \$3,000 per hour. It is estimated that each plane goes into the maintenance shop 5 times per year. It is believed that the input process for the base is essentially Poisson and that the probability distribution of repair times is Erlang, with shape parameter $k = 2$.

Alternative *A* is to provide a duplicate maintenance shop, so that two planes can be repaired simultaneously. The cost, amortized over 5 years, is \$400,000 per year for each of the airline's airplanes.

Alternative *B* is to replace the present maintenance equipment by the most efficient (and expensive) equipment available, thereby reducing the expected repair time to 18 hours. The cost, amortized over 5 years, is \$550,000 per year for each airplane.

Which alternative should the airline choose?

18.4-10.* The production of tractors at the Jim Buck Company involves producing several subassemblies and then using an assembly line to assemble the subassemblies and other parts into finished tractors. Approximately three tractors per day are produced in this way. An in-process inspection station is used to inspect the subassemblies before they enter the assembly line. At present there are two inspectors at the station, and they work together to inspect each subassembly. The inspection time has an exponential distribution, with a mean of 15 minutes. The cost of providing this inspection system is \$40 per hour.

A proposal has been made to streamline the inspection procedure so that it can be handled by only one inspector. This inspector would begin by visually inspecting the exterior of the subassembly, and she would then use new efficient equipment to complete the inspection. Although this process with just one inspector would slightly increase the mean of the distribution of inspection times from 15 minutes to 16 minutes, it also would reduce the variance of this distribution to only 40 percent of its current value.

The subassemblies arrive at the inspection station according to a Poisson process at a mean rate of 3 per hour. The cost of hav-

ing the subassemblies wait at the inspection station (thereby increasing in-process inventory and possibly disrupting subsequent production) is estimated to be $20 per hour for each subassembly.

Management now needs to make a decision about whether to continue the status quo or adopt the proposal.

T **(a)** Find the main measures of performance—L, L_q, W, W_q—for the current queueing system.

(b) Repeat part (*a*) for the proposed queueing system.

(c) What conclusions can you draw about what management should do from the results in parts (*a*) and (*b*)?

(d) Determine and compare the expected total cost per hour for the status quo and the proposal.

18.4-11. The car rental company, Try Harder, has been subcontracting for the maintenance of its cars in St. Louis. However, due to long delays in getting its cars back, the company has decided to open its own maintenance shop to do this work more quickly. This shop will operate 42 hours per week.

Alternative 1 is to hire two mechanics (at a cost of $1,500 per week each), so that two cars can be worked on at a time. The time required by a mechanic to service a car has an Erlang distribution, with a mean of 5 hours and a shape parameter of $k = 8$.

Alternative 2 is to hire just one mechanic (for $1,500 per week) but to provide some additional special equipment (at a capitalized cost of $1,250 per week) to speed up the work. In this case, the maintenance work on each car is done in two stages, where the time required for each stage has an Erlang distribution with the shape parameter $k = 4$, where the mean is 2 hours for the first stage and 1 hour for the second stage.

For both alternatives, the cars arrive according to a Poisson process at a mean rate of 0.3 car per hour (during work hours). The company estimates that its net lost revenue due to having its cars unavailable for rental is $150 per week per car.

(a) Use Fig. 17.13 to estimate L, L_q, W, and W_q for alternative 1.

(b) Find these same measures of performance for alternative 2.

(c) Determine and compare the expected total cost per week for these alternatives.

18.4-12. A certain small car-wash business is currently being analyzed to see if costs can be reduced. Customers arrive according to a Poisson process at a mean rate of 15 per hour, and only one car can be washed at a time. At present the time required to wash a car has an exponential distribution, with a mean of 4 minutes. It also has been noticed that if there are already 4 cars waiting (including the one being washed), then any additional arriving customers leave and take their business elsewhere. The lost incremental profit from each such lost customer is $6.

Two proposals have been made. Proposal 1 is to add certain equipment, at a capitalized cost of $6 per hour, which would reduce the expected washing time to 3 minutes. In addition, each arriving customer would be given a guarantee that if she had to wait longer than $\frac{1}{2}$ hour (according to a time slip she receives upon arrival) before her car is ready, then she receives a free car wash (at a marginal cost of $4 for the company). This guarantee would be well posted and advertised, so it is believed that no arriving customers would be lost.

Proposal 2 is to obtain the most advanced equipment available, at an increased cost of $20 per hour, and each car would be sent through two cycles of the process in succession. The time required for a cycle has an exponential distribution, with a mean of 1 minute, so total expected washing time would be 2 minutes. Because of the increased speed and effectiveness, it is believed that essentially no arriving customers would be lost.

The owner also feels that because of the loss of customer goodwill (and consequent lost future business) when customers have to wait, a cost of $0.20 for each minute that a customer has to wait before her car wash begins should be included in the analysis of all alternatives.

Evaluate the expected total cost per hour E(TC) of the status quo, proposal 1, and proposal 2 to determine which one should be chosen.

18.4-13.* The Seabuck and Roper Company has a large warehouse in southern California to store its inventory of goods until they are needed by the company's many furniture stores in that area. A single crew with four members is used to unload and/or load each truck that arrives at the loading dock of the warehouse. Management currently is downsizing to cut costs, so a decision needs to be made about the future size of this crew.

Trucks arrive at the loading dock according to a Poisson process at a mean rate of 1 per hour. The time required by a crew to unload and/or load a truck has an exponential distribution (regardless of crew size). The mean of this distribution with the four-member crew is 15 minutes. If the size of the crew were to be changed, it is estimated that the mean service rate of the crew (now $\mu = 4$ customers per hour) would be proportional to its size.

The cost of providing each member of the crew is $20 per hour. The cost that is attributable to having a truck not in use (i.e., a truck standing at the loading dock) is estimated to be $30 per hour.

(a) Identify the customers and servers for this queueing system. How many servers does it currently have?

T **(b)** Use the appropriate Excel template to find the various measures of performance for this queueing system with four members on the crew. (Set $t = 1$ hour in the Excel template for the waiting-time probabilities.)

T **(c)** Repeat (*b*) with three members.

T **(d)** Repeat part (*b*) with two members.

(e) Should a one-member crew also be considered? Explain.

(f) Given the previous results, which crew size do you think management should choose?

(g) Use the cost figures to determine which crew size would minimize the expected total cost per hour.

(h) Assume now that the mean service rate of the crew is proportional to the square root of its size. What should the size be to minimize expected total cost per hour?

18.4-14. Trucks arrive at a warehouse according to a Poisson process with a mean rate of 4 per hour. Only one truck can be loaded at a time. The time required to load a truck has an exponential distribution with a mean of $10/n$ minutes, where n is the number of loaders ($n = 1, 2, 3, \ldots$). The costs are (i) \$18 per hour for each loader and (ii) \$20 per hour for each truck being loaded or waiting in line to be loaded. Determine the number of loaders that minimizes the expected hourly cost.

18.4-15. A company's machines break down according to a Poisson process at a mean rate of 3 per hour. Nonproductive time on any machine costs the company \$60 per hour. The company employs a maintenance person who repairs machines at a mean rate of μ machines per hour (when continuously busy) if the company pays that person a wage of $\$5\mu$ per hour. The repair time has an exponential distribution.

Determine the hourly wage that minimizes the company's total expected cost.

18.4-16. Jake's Machine Shop contains a grinder for sharpening the machine cutting tools. A decision must now be made on the speed at which to set the grinder.

The grinding time required by a machine operator to sharpen the cutting tool has an exponential distribution, where the mean $1/\mu$ can be set at 0.5 minute, 1 minute, or 1.5 minutes, depending upon the speed of the grinder. The running and maintenance costs go up rapidly with the speed of the grinder, so the estimated cost per minute is \$1.60 for providing a mean of 0.5 minute, \$0.40 for a mean of 1.0 minute, and \$0.20 for a mean of 1.5 minutes.

The machine operators arrive randomly to sharpen their tools at a mean rate of 1 every 2 minutes. The estimated cost of an operator being away from his or her machine to the grinder is \$0.80 per minute.

T **(a)** Obtain the various measures of performance for this queueing system for each of the three alternative speeds for the grinder. (Set $t = 5$ minutes in the Excel template for the waiting time probabilities.)

(b) Use the cost figures to determine which grinder speed minimizes the expected total cost per minute.

18.4-17. Consider the special case of model 2 where (1) any $\mu > \lambda/s$ is feasible and (2) both $f(\mu)$ and the waiting-cost function are linear functions, so that

$$E(\text{TC}) = C_r s\mu + C_w L,$$

where C_r is the marginal cost per unit time for each unit of a server's mean service rate and C_w is the cost of waiting per unit time for each customer. The optimal solution is $s = 1$ (by the optimality of a single-server result), and

$$\mu = \lambda + \sqrt{\frac{\lambda C_w}{C_r}}$$

for any queueing system fitting the $M/M/1$ model presented in Sec. 17.6.

Show that this μ is indeed optimal for the $M/M/1$ model.

18.4-18. Greg is making plans to open a new fast-food restaurant soon. He is estimating that customers will arrive randomly (a Poisson process) at a mean rate of 150 per hour during the busiest times of the day. He is planning to have three employees directly serving the customers. He now needs to make a decision about how to organize these employees.

Option 1 is to have three cash registers with one employee at each to take the orders and get the food and drinks. In this case, it is estimated that the average time to serve each customer would be 1 minute, and the distribution of service times is assumed to be exponential.

Option 2 is to have one cash register with the three employees working together to serve each customer. One would take the order, a second would get the food, and the third would get the drinks. Greg estimates that this would reduce the average time to serve each customer down to 20 seconds, with the same assumption of exponential service times.

Greg wants to choose the option that would provide the best service to his customers. However, since Option 1 has three cash registers, both options would serve the customers at a mean rate of 3 per minute when everybody is busy serving customers, so it is not clear which option is better.

T **(a)** Use the main measures of performance—L, L_q, W, W_q—to compare the two options.

(b) Explain why these comparisons make sense intuitively.

(c) Which measure do you think would be most important to Greg's customers? Why? Which option is better with respect to this measure?

18.4-19. Consider a harbor with a single dock for unloading ships. The ships arrive according to a Poisson process at a mean rate of λ ships per week, and the service-time distribution is exponential with a mean rate of μ unloadings per week. Assume that harbor facilities are owned by the shipping company, so that the objective is to balance the cost associated with idle ships with the cost of running the dock. The shipping company has no control over the arrival rate λ (that is, λ is fixed); however, by changing the size of the unloading crew, and so on, the shipping company can adjust the value of μ as desired.

Suppose that the expected cost per unit time of running the unloading dock is $D\mu$. The waiting cost for each idle ship is some constant (C) times the *square* of the total waiting time (including loading time). The shipping company wishes to adjust μ so that the expected total cost (including the waiting cost for idle ships) per unit time is minimized. Derive this optimal value of μ in terms of D and C.

18.4-20. Consider a queueing system with two types of customers. Type 1 customers arrive according to a Poisson process with a mean rate of 5 per hour. Type 2 customers also arrive according to a Poisson process with a mean rate of 5 per hour. The system has two servers, and both serve both types of customers. For types 1 and 2, service times have an exponential distribution with a mean of 10 minutes. Service is provided on a first-come-first-served basis.

Management now wants you to compare this system's design of having both servers serve both types of customers with the alternative design of having one server serve just type 1 customers and the other server serve just type 2 customers. Assume that this alternative design would not change the probability distribution of service times.

(a) Without doing any calculations, indicate which design would give a smaller expected total number of customers in the system. What result are you using to draw this conclusion?

T **(b)** Verify your conclusion in part (*a*) by finding the expected total number of customers in the system under the original design and then under the alternative design.

18.4-21. Reconsider Prob. 17.6-33.

(a) Formulate part (*a*) to fit as closely as possible a special case of one of the decision models presented in Sec. 18.4. (Do not solve.)

(b) Describe Alternatives 2 and 3 in queueing theory terms, including their relationship (if any) to the decision models presented in Sec. 18.4. Briefly indicate why, in comparison with Alternative 1, each of these other alternatives might decrease the total number of operators (thereby increasing their utilization) needed to achieve the required production rate. Also point out any dangers that might prevent this decrease.

18.4-22. George is planning to open a drive-through photo-developing booth with a single service window that will be open approximately 200 hours per month in a busy commercial area. Space for a drive-through lane is available for a rental of $200 per month per car length. George needs to decide how many car lengths of space to provide for his customers.

Excluding this rental cost for the drive-through lane, George believes that he will average a profit of $4 per customer served (nothing for a drop-off of film and $8 when the photographs are picked up). He also estimates that customers will arrive randomly (a Poisson process) at a mean rate of 20 per hour, although those who find the drive-through lane full will be forced to leave. Half of the customers who find the drive-through lane full wanted to drop off film and the other half wanted to pick up their photographs. The half who wanted to drop off film will take their business elsewhere instead. The other half of the customers who find the drive-through lane full will not be lost because they will keep trying later until they get in and pick up their photographs. George assumes that the time required to serve a customer will have an exponential distribution with a mean of 2 minutes.

T **(a)** Find L and the mean rate at which customers are lost when the number of car lengths of space provided is 2, 3, 4, and 5.

(b) Calculate W from L for the cases considered in part (*a*).

(c) Use the results from part (*a*) to calculate the decrease in the mean rate at which customers are lost when the number of car lengths of space provided is increased from 2 to 3, from 3 to 4, and from 4 to 5. Then calculate the increase in expected profit per hour (excluding space rental costs) for each of these three cases.

(d) Compare the increases in expected profit found in part (*c*) with the cost per hour of renting each car length of space. What conclusion do you draw about the number of car lengths of space that George should provide?

18.4-23. Consider a factory whose floor area is a square with 600 feet on each side. Suppose that one service facility of a certain kind is provided in the center of the factory. The employees are distributed uniformly throughout the factory, and they walk to and from the facility at an average speed of 3 miles per hour along a system of orthogonal aisles.

Find the expected travel time $E(T)$ per arrival.

18.4-24. A certain large shop doing light fabrication work uses a single central storage facility (dispatch station) for material in in-process storage. The typical procedure is that each employee personally delivers his finished work (by hand, tote box, or hand cart) and receives new work and materials at the facility. Although this procedure worked well in earlier years when the shop was smaller, it appears that it may now be advisable to divide the shop into two semi-independent parts, with a separate storage facility for each one. You have been assigned the job of comparing the use of two facilities and of one facility from a cost standpoint.

The factory has the shape of a rectangle 150 by 100 yards. Thus, by letting 1 yard be the unit of distance, the (x, y) coordinates of the corners are (0, 0), (150, 0), (150, 100), and (0, 100). With this coordinate system, the existing facility is located at (50, 50), and the location available for the second facility is (100, 50).

Each facility would be operated by a single clerk. The time required by a clerk to service a caller has an exponential distribution, with a mean of 2 minutes. Employees arrive at the present facility according to a Poisson input process at a mean rate of 24 per hour. The employees are rather uniformly distributed throughout

the shop, and if the second facility were installed, each employee would normally use the nearer of the two facilities. Employees walk at an average speed of about 5,000 yards per hour. All aisles are parallel to the outer walls of the shop. The net cost of providing each facility is estimated to be about $20 per hour, plus $15 per hour for the clerk. The estimated total cost of an employee being idled by traveling or waiting at the facility is $25 per hour.

Given the preceding cost factors, which alternative minimizes the expected total cost?

18.4-25.* Consider the formulation of the County Hospital emergency room problem as a preemptive priority queueing system, as presented in Sec. 17.8. Suppose that the following inputted costs are assigned to making patients wait (*excluding* treatment time): $10 per hour for stable cases, $1,000 per hour for serious cases, and $100,000 per hour for critical cases. The cost associated with having an additional doctor on duty would be $40 per hour. Referring to Table 17.4, determine on an expected-total-cost basis whether there should be one or two doctors on duty.

T **18.4-26.** The Becker Company factory has been experiencing long delays in jobs going through the turret lathe department because of inadequate capacity. The head of this department contends that five machines are required, as opposed to the three machines now in place. However, because of pressure from management to hold down capital expenditures, only one additional machine will be authorized unless there is solid evidence that a second one is necessary.

This shop does three kinds of jobs, namely, government jobs, commercial jobs, and standard products. Whenever a turret lathe operator finishes a job, he starts a government job if one is waiting; if not, he starts a commercial job if any are waiting; if not, he starts on a standard product if any are waiting. Jobs of the same type are taken on a first-come-first-served basis.

Although much overtime work is required currently, management wants the turret lathe department to operate on an 8-hour, 5-day-per-week basis. The probability distribution of the time required by a turret lathe operator for a job appears to be approximately exponential, with a mean of 10 hours. Jobs come into the shop randomly (a Poisson process) at a mean rate of 6 per week for governments jobs, 4 per week for commercial jobs, and 2 per week for standard products. (These figures are expected to remain the same for the indefinite future.)

Management feels that the average waiting time before work begins in the turret lathe department should not exceed 0.25 (working) day for government jobs, 0.5 day for commercial jobs, and 2 days for standard products.

(a) Determine how many additional turret lathes need to be obtained to satisfy these management guidelines.

(b) It is worth about $750, $450, and $150 to avoid a delay of 1 additional (working) day in a government, commercial, and standard job, respectively. The incremental capitalized cost of providing each turret lathe (including the operator and so on) is estimated to be $250 per working day. Determine the number of additional turret lathes that should be obtained to minimize the expected total cost.

CASE 18.1 QUEUEING QUANDARY[1]

Never dull. That is how you would describe your job at the centralized records and benefits administration center for Cutting Edge, a large company manufacturing computers and computer peripherals. Since opening the facility six months ago, you and Mark Lawrence, the Director of Human Resources, have endured one long roller coaster ride. Receiving the go-ahead from corporate headquarters to establish the centralized records and benefits administration center was definitely an up. Getting caught in the crossfire of angry customers (all employees of Cutting Edge) because of demand overload for the records and benefits call center was definitely a down. Accurately forecasting the demand for the call center provided another up.

And today you are faced with another down. Mark approaches your desk with a not altogether attractive frown on his face.

[1]The scenario in this case is a sequel, a few months later, to the scenario introduced in Case 20.1. However, this case can be considered completely independently of Case 20.1.

He begins complaining immediately, "I just don't understand. The forecasting job you did for us two months ago really allowed us to understand the weekly demand for the center, but we still have not been able to get a grasp on the staffing problem. We used both historical data and your forecasts to calculate the average weekly demand for the call center. We transformed this average weekly demand into average hourly demand by dividing the weekly demand by the number of hours in the workweek. We then staffed the center to meet this average hourly demand by taking into account the average number of calls a representative is able to handle per hour.

But something is horribly wrong. Operational data records show that over thirty percent of the customers wait over four minutes for a representative to answer the call! Customers are still sending me numerous complaints, and executives from corporate headquarters are still breathing down my neck! I need help!"

You calm Mark down and explain to him that you think you know the problem: the number of calls received in a certain hour can be much greater (or much less) than the average because of the stochastic nature of the demand. In addition, the number of calls a representative is able to handle per hour can be much less (or much greater) than the average depending upon the types of calls received.

You then tell him to have no fear, you have the problem under control. You have been reading about the successful application of queueing theory to the operation of call centers, and you decide that the queueing models you learned in school will help you determine the appropriate staffing level.

(a) You ask Mark to describe the demand and service rate. He tells you that calls are randomly received by the call center and that the center receives an average of 70 calls per hour. The computer system installed to answer and hold the calls is so advanced that its capacity far exceeds the demand. Because the nature of a call is random, the time required to process a call is random, where the time frequently is small but occasionally can be much longer. On average, however, representatives can handle 6 calls per hour. Which queueing model seems appropriate for this situation? Given that slightly more than 35 percent of customers wait over 4 minutes before a representative answers the call, use this model to estimate how many representatives Mark currently employs.

(b) Mark tells you that he will not be satisfied unless 95 percent of the customers wait only 1 minute or less for a representative to answer the call. Given this customer service level and the average arrival rates and service rates from part (*a*), how many representatives should Mark employ?

(c) Each representative receives an annual salary of $30,000, and Mark tells you that he simply does not have the resources available to hire the number of representatives required to achieve the customer service level desired in part (*b*). He asks you to perform sensitivity analysis. How many representatives would he need to employ to ensure that 80 percent of customers wait 1 minute or less? How many would he need to employ to ensure that 95 percent of customers wait 90 seconds or less? How would you recommend Mark choose a customer service level? Would the decision criteria be different if Mark's call center were to serve external customers (not connected to the company) instead of internal customers (employees)?

(d) Mark tells you that he is not happy with the number of representatives required to achieve a high customer service level. He therefore wants to explore alternatives to simply hiring additional representatives. The alternative he considers is instituting a training program that

will teach representatives to more efficiently use computer tools to answer calls. He believes that this alternative will increase the average number of calls a representative is able to handle per hour from 6 calls to 8 calls. The training program will cost $2,500 per employee per year since employees' knowledge will have to be updated yearly. How many representatives will Mark have to employ and train to achieve the customer service level desired in part (*b*)? Do you prefer this alternative to simply hiring additional representatives? Why or why not?

(e) Mark realizes that queueing theory helps him only so much in determining the number of representatives needed. He realizes that the queueing models will not provide accurate answers if the inputs used in the models are inaccurate. What inputs do you think need reevaluation? How would you go about estimating these inputs?