



CHAPTER FOURTEEN

QUEUING MODELS AND CAPACITY PLANNING

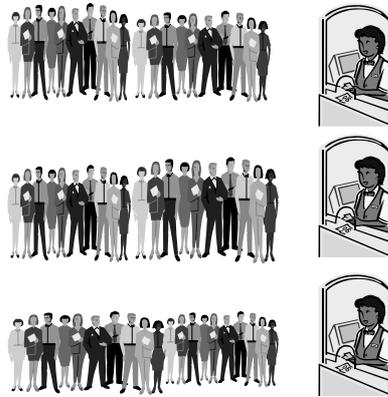
Queuing theory is a mathematical approach to the analysis of waiting lines. Waiting lines in health care organizations can be found wherever either patients or customers arrive randomly for services, such as walk-in patients and emergency room arrivals, or phone calls from physician offices to health maintenance organizations (HMO) for approvals. Patients arriving for health care services with appointments are not considered as waiting lines, even if they wait to see their health care provider. Most sorts of health care service systems have the capacity to serve more patients than they are called to over the long term. Therefore, customer waiting lines are a *short-term* phenomenon, and the employees who serve customers, or caregivers who serve patients, are frequently inactive while they wait for customers to arrive.

If service capacity is increased, waiting lines should become smaller, but then employees (called **servers**) will be idle more often as they wait for customers—or, in health care, patients (see Figure 14.1). A health care manager can examine the trade-off between capacity and service delays using **queuing analysis**. Specifically, when considering improvements in services, the health care manager weighs the cost of providing a given level of service against the potential costs from having patients wait.

Why must we wait in lines? The following example illustrates another waiting phenomenon. A hospital ER may have the capacity to handle an average of fifty patients an hour, and yet may have waiting lines even though the average number

FIGURE 14.1. QUEUE PHENOMENON.

Location: Hospital Outpatient Pharmacy
 Day & Time: Monday, 11:00 AM
 Average Number in Line: 20
 Idle Servers: 0
 Average Wait Time: 15 Minutes



Location: Hospital Outpatient Pharmacy
 Day & Time: Monday, 3:30 PM
 Average Number in Line: 0
 Idle Servers: 3
 Average Wait Time: 0 Minutes
 Hourly Wage Per Idle Pharmacist: \$40



of patients is only thirty-five an hour. The key word is *average*. In reality, patients arrive at random intervals rather than at evenly spaced intervals, and some patients require more intensive treatment (longer service time) than others. In other words, both arrivals and length of service times exhibit great variability. As a result, the ER becomes *temporarily* overloaded at times, and patients have to wait. At other times, the ER is idle because there are no patients. Although a system may be underloaded from a macro viewpoint (long-term), variability in patient arrivals and medical service times sometimes causes the system to be overloaded from a micro standpoint (short-term). In systems where variability can be minimized—because of scheduled arrivals or constant service times—waiting lines should not ordinarily form. With the diversity of services and the arrival patterns in the health care sector, however, that condition is unattainable in many areas of delivery.

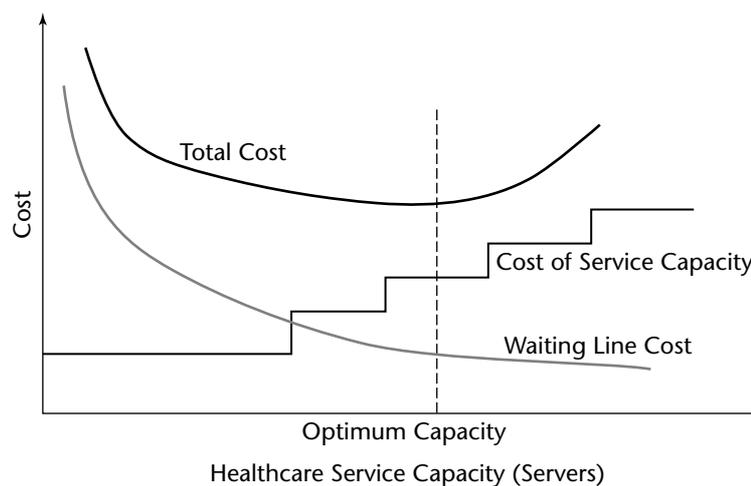
The goal of queuing is to minimize total costs. The two basic costs mentioned above are those associated with patients or customers having to wait for service and those associated with capacity. **Capacity costs** are the costs of maintaining the ability to provide service. For example, physicians' and nurses' salaries, as well

as other fixed costs, must be paid whether the ER is idle or not. **Waiting costs** include the salaries paid to employees while they wait for service from other employees (for example, a physician in a group practice, waiting for an exam room to be cleaned and readied for the next patient, or waiting for an x-ray or test result); the cost of waiting space (such as the size of a doctor's waiting room); and also the loss of business a health care organization can suffer when patients refuse to wait and go elsewhere in the future. Of course, society, too, incurs costs for more critical care when a patient has not been received soon enough because of congested waiting times or limited capacity.

It is difficult to accurately pin down the cost to the health care organization of patients' waiting time, so health care managers often treat waiting times or line lengths as a policy variable. An acceptable extent of waiting is specified, and the health care manager directs that capacity be established to meet that level. The goal of queuing analysis is to balance the cost of providing a level of health care service capacity with the cost to the health care organization of keeping patients waiting. The concept is illustrated in Figure 14.2.

Note that as service capacity increases, so does its cost; service capacity costs are shown as incremental (rising in steps for given service levels). As capacity increases, however, the number of patients waiting and the time they wait tend to decrease, so waiting costs decrease. A total cost curve is then added to the graph to reflect the trade-off between those two costs. The goal of the analysis is to identify the level of service capacity that will minimize total cost.

FIGURE 14.2. HEALTH CARE SERVICE CAPACITY AND COSTS.



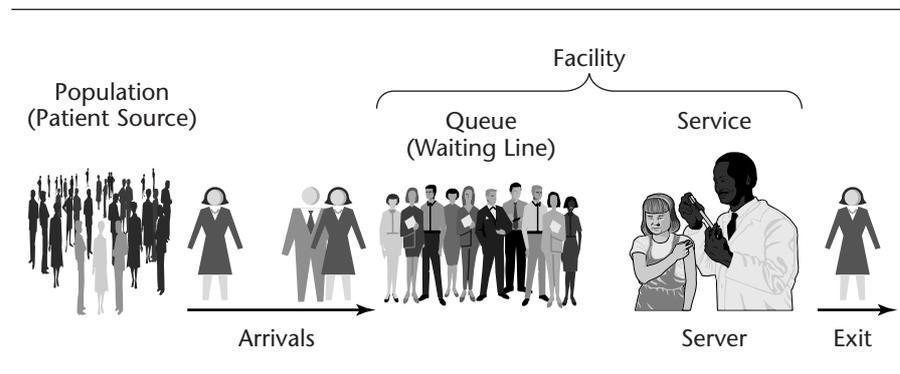
Queuing System Characteristics

A health care manager can choose among many queuing models. Obviously, choosing the appropriate one is the key to solving the problem successfully. Model choice depends on the characteristics of the system under investigation. The main queuing model characteristics are: 1) the population source; 2) number of servers; 3) arrival patterns and service patterns; and 4) queue discipline.

Figure 14.3 illustrates a flu inoculation process as a simple queuing model: patients come from a population, enter on a queue (waiting line) for service, receive flu injections from a health care provider (server), and leave the system.

Population Source. The first characteristic to look at when analyzing a queuing problem is whether or not the potential number of patients is limited, that is, whether the population source is infinite or finite. In an **infinite source** situation, patient arrivals are unrestricted, and can greatly exceed system capacity at any time. An infinite source exists when service (access) is unrestricted, such as at a public hospital ER. When potential patients are limited to small numbers, a **finite source** situation exists, for example when a mental health caseworker is assigned forty clients. When one or many clients leave or are added to the caseworker's assignment load, the probability changes of help being needed—a client needing therapeutic service. As seen in that example, then, finite source models require a formulation different than that of infinite source models. Other types of finite population situations are a health care facility (such as a PPO) contracted to serve the members of a given health insurance plan, or a physician practice with two thousand patients. For most of these queuing situations, however, infinite

FIGURE 14.3. QUEUING CONCEPTUALIZATION OF FLU INOCULATIONS.

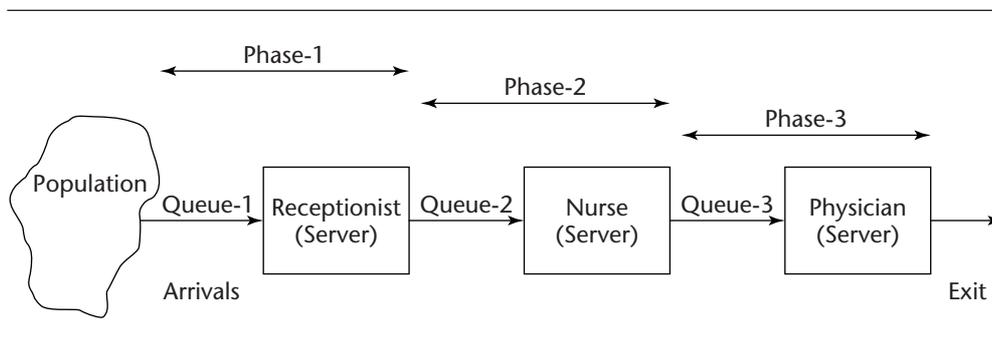


source models could be used, since the patient base is large enough not to cause any major shift in probabilities, and also would cause no significant errors. Hence, we will consider only infinite source models, as being usually more applicable to queuing and capacity problems in health care.

Number of Servers. The capacity of queuing systems is determined by the capacity of each server (also known as a line or **channel**) and the number of servers being used. It is generally assumed that each channel can handle one customer at a time. Health care systems can be conceptualized as **single-line** or **multiple-line**, and may consist of **phases** (steps in a queuing system). Examples of single-line systems in health care facilities are rare. The flu inoculation example best illustrates one, in which a single health care provider carries out both administration (paperwork for consent, fee collection) and clinical care (inoculation) as a single server. In contrast, many solo health care providers (physicians, dentists, therapists) have offices with receptionists and nurses or other assistants; those are examples of single-line, multiphase systems. Figure 14.4 shows the conceptualization of a single-line, multiphase system. Patients arrive to see a receptionist, and if others are before them they wait until a receptionist is available [first queue]; eventually they reach a receptionist, process initial paperwork, and wait; to see a nurse or physician assistant for initial examination of vital signs (blood pressure, temperature, complaint, and history taking) [second queue]; and they wait again until a physician is available [third queue].

Multiple-line systems are found in many health care facilities: hospitals, outpatient clinics, emergency services, and so on. Multiple-line queue systems can be either single-phase or multiphase. A single-phase, multiple-line system would be illustrated by an extension of flu inoculation to more than one server (three

FIGURE 14.4. CONCEPTUALIZATION OF A SINGLE-LINE, MULTIPHASE SYSTEM.



nurses giving inoculations and patients forming a single queue to wait (see Figure 14.5). In actuality, most health care services are multiple-line, multiphase systems. For example, a non-urgent arrival to an emergency room can be conceptualized in several phases: 1) initial evaluation; 2) diagnostic tests; and 3) clinical interventions. Although the phases will vary from patient to patient, because each one receives care from several staff members in succession, the configuration in this case is a multiple-line queuing system. The lower half of Figure 14.5 illustrates the multiple-line, multiphase queuing example.

Arrival Patterns. Waiting lines occur because random, highly variable arrival and service patterns cause systems to be temporarily overloaded. Hospital emergency rooms are very typical examples of erratic arrival patterns causing such variability. The arrival patterns might be different on mornings and afternoons, and even more so after physician offices close, in the evenings. In general, queues are more prevalent in evening hours and on weekends. Figure 14.6 illustrates the random

FIGURE 14.5. MULTIPLE-LINE QUEUING SYSTEM.

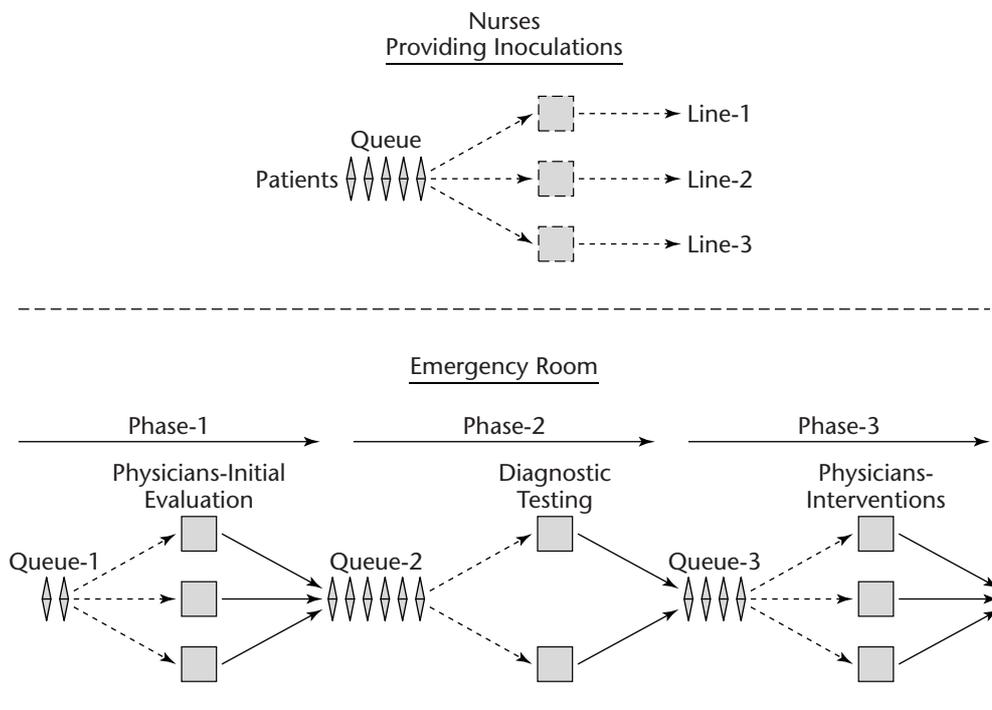
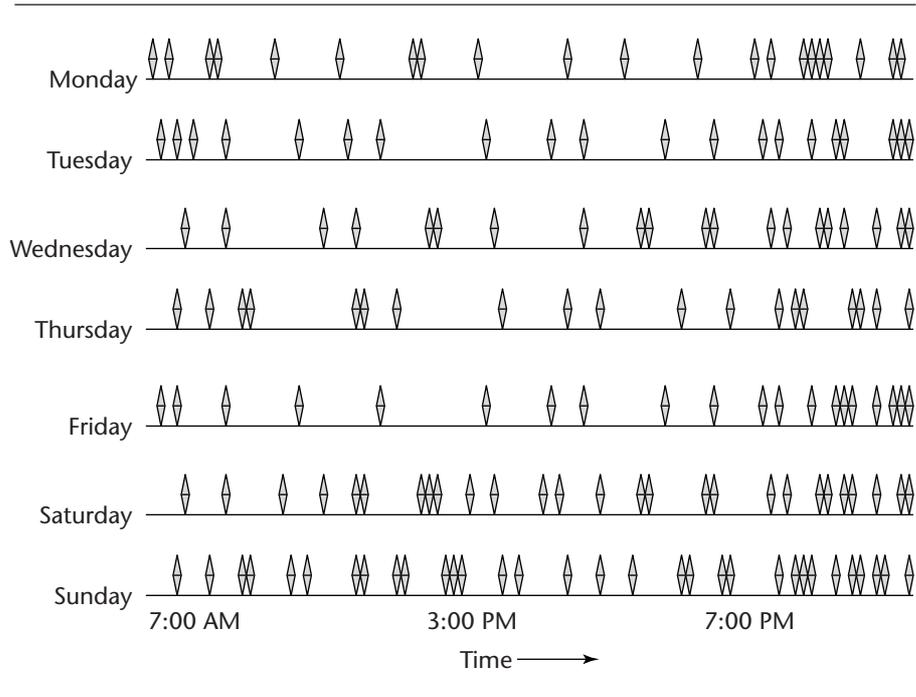


FIGURE 14.6. EMERGENCY ROOM ARRIVAL PATTERNS.



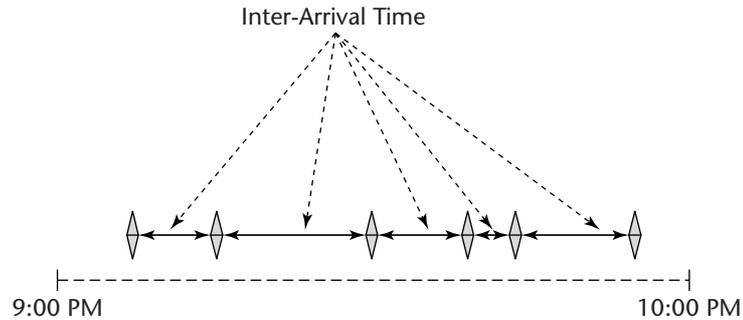
behavior of arrivals at various times of the day and days of the week. The patterns on weekends are distributed more densely than those on weekdays, as are those during evening hours.

Besides that, in any block of time there are no discernible patterns, so the random nature of the arrivals—their numbers and the times between the arrivals—has to be measured. The variability can often be described by theoretical distributions.

The most commonly used models assume that the patient arrival rate can be described by a Poisson distribution, and that the time between arrivals, inter-arrival time, can be described by a negative exponential distribution. Figure 14.7 conceptualizes the arrival rate and inter-arrival times.

Arrival rate is determined as the average arrivals for a given time period, as illustrated in Figure 14.7. During the hour of 9:00–10:00 P.M., there are six arrivals, and if that held true over the same time period for a number of days, then we could say that the average arrival rate is six patients per hour. The spacing between the arrivals, inter-arrival time, does not occur uniformly. The first patient

FIGURE 14.7. MEASURES OF ARRIVAL PATTERNS.



arrives at ten minutes after the hour, the second fifteen minutes later, and so on. Such patterns are often characterized by negative exponential distribution. The mean of the negative exponential distribution, average inter-arrival time, can be obtained by dividing the average number of patients arrived into the time period—in this case sixty minutes. Hence the average inter-arrival time for this example is $60 \div 6 = 10$ minutes, which is interpreted as: patients are arriving on average ten minutes apart. One can convert arrival rate to inter-arrival time or vice versa, since arrival rates can be described with a Poisson distribution, which is more practical to use than negative exponential distribution. Poisson distribution, Figure 14.8, is a discrete distribution that shows the probability of arrivals in a given time period, where the mean and variance of the Poisson distribution are the same.

Service Patterns. Service to the arriving patients is another element that exhibits variability. Because of the varying nature of illnesses and patient conditions, the time required for clinical attention (service times) varies from patient to patient. Figure 14.9 illustrates a service pattern for ER patients when patient A requires over 100 minutes of direct clinical attention, but patient C requires about 25 minutes. As in the inter-arrival time, service time also can be described by negative exponential distribution. However, service rate and service times are also interchangeably used, so that the Poisson distribution can characterize the service rate.

In summary, the Poisson and the negative exponential distribution are alternate ways of presenting the same information. If service time is exponential, then the service rate is Poisson. Further, if the customer arrival rate is Poisson, then the inter-arrival rate (the time between arrivals) is exponential. In another example, if a lab processes ten customers per hour (rate), the average service time is six

FIGURE 14.8. POISSON DISTRIBUTION.

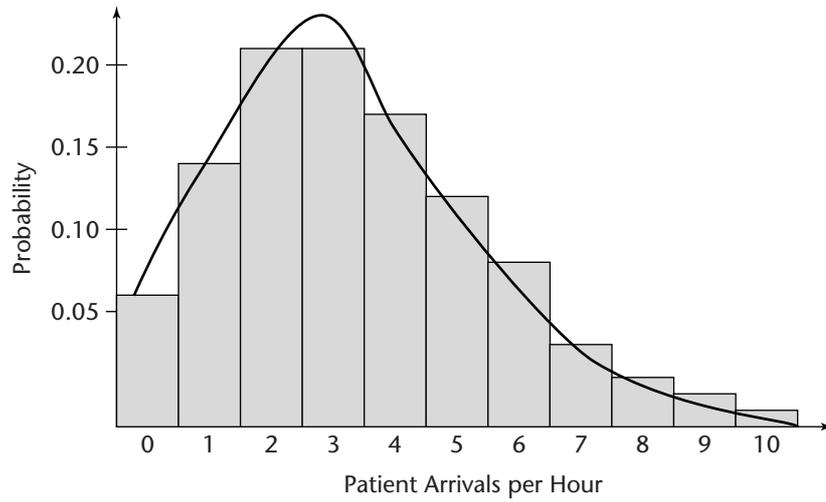
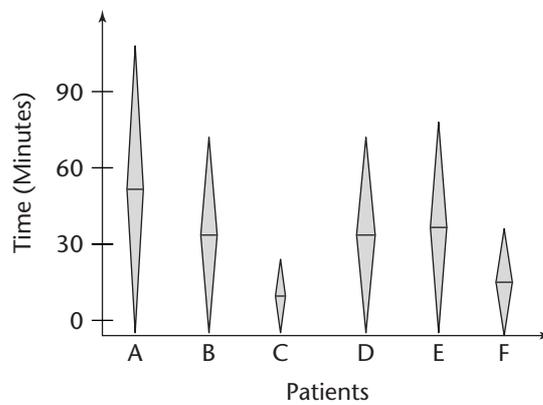


FIGURE 14.9. SERVICE TIME FOR ER PATIENTS.



minutes. If the arrival rate is twelve per hour, then the average time between arrivals is five minutes. Thus, service and arrival rates are described by the Poisson distribution, and inter-arrival times and service times are described by a negative exponential distribution.

Queue Characteristics. Queues can be infinitely long or with limited capacity. A flu shot clinic with patients forming a queue around the block can be described as an infinite queue, whereas a physician office with fifteen chairs in a waiting room is an example of a limited-capacity queue.

A queue can be formed as a single line to one or more server(s), or it can be formed as separate lines for each server. In the second type, patients may jump from queue to queue to gain advantage in reaching a service point, but often lose more time because of service variability. Patients who arrive and see big lines (the flu shot example) may change their minds and not join the queue, but go elsewhere to obtain service; this is called **balking**. If they do join the queue and are dissatisfied with the waiting time, they may leave the queue; this is called **reneging**.

Queue discipline refers to the order in which customers are processed. The assumption that service is provided on a **first-come, first-served** basis is probably the most commonly encountered rule. First-come first-served, which is seen in many businesses, has special adaptations in health care queue discipline: shortest processing time first (for example, in the operating room simple or small surgeries may be scheduled first); reservation first (in the physician office); critical first (in the emergency room). Let us examine the example of the emergency room, which does not serve on a first-come basis. Patients do not all represent the same risk (or waiting costs); those with the highest risk (the most seriously ill) are processed first under a triage system, even though other patients may have arrived earlier.

Queuing models are identified by their characteristics. From a methods perspective, a nomenclature of A/B/C/D/E is used to describe them. Exhibit 14.1

EXHIBIT 14.1. QUEUING MODEL CLASSIFICATION.

- A: Specification of arrival process, measured by inter-arrival time or arrival rate.
 - M: Negative exponential or Poisson distribution.
 - D: Constant value.
 - K: Erlang distribution.
 - G: A general distribution with known mean and variance.
 - B: Specification of service process, measured by inter-service time or service rate.
 - M: Negative exponential or Poisson distribution.
 - D: Constant value.
 - K: Erlang distribution.
 - G: A general distribution with known mean and variance.
 - C: Specification of number of servers—"s".
 - D: Specification of queue or the maximum numbers allowed in a queuing system.
 - E: Specification of customer population.
-

provides details for each component of the nomenclature. The last two components, D and E, of the nomenclature are not used unless there is a specific waiting room capacity or a limited population of patients. Two examples of nomenclature in use are: 1) a queuing model with Poisson arrival and service rates with three servers is described by M/M/3. A physician office with waiting room capacity of fifteen, five physicians, and Poisson arrival and service rates is described by M/M/5/15.

Since infinite-patient-source models are our main focus, the last section of the nomenclature, “E,” will be omitted in the ensuing discussions.

Measures of Queuing System Performance

The health care manager must consider five typical measures when evaluating existing or proposed service systems. Those measures are:

1. Average number of patients waiting (in queue or in the system).
2. Average time the patients wait (in queue or in the system).
3. Capacity utilization.
4. Costs of a given level of capacity.
5. Probability that an arriving patient will have to wait for service.

The system utilization measure reflects the extent to which the servers are busy rather than idle. On the surface, it might seem that health care managers would seek 100 percent system utilization. However, increases in system utilization are achieved only at the expense of increases in both the length of the waiting line and the average waiting time, with values becoming exceedingly large as utilization approaches 100 percent. Under normal circumstances, 100 percent utilization may not be realistic; a health care manager should try to achieve a system that minimizes the sum of waiting costs and capacity costs. In queue modeling, the health care manager also must ensure that average arrival and service rates are stable, indicating that the system is in a **steady state**, a fundamental assumption.

Typical Infinite-Source Models

This section provides examples of the two commonly used models:

1. Single channel, M/M/s.
2. Multiple channel, M/M/s > 1.

where “s” designates the number of channels (servers).

EXHIBIT 14.2. QUEUING MODEL NOTATION.

λ	arrival rate
μ	service rate
L_q	average number of customers waiting for service
L	average number of customers in the system (waiting or being served)
W_q	average time customers wait in line
W	average time customers spend in the system
ρ	system utilization
$1/\mu$	service time
P_0	probability of zero units in system
P_n	probability of n units in system

These models assume steady state conditions and a Poisson arrival rate. The most commonly used symbols in queuing models are shown in Exhibit 14.2.

Model Formulations

Five key relationships provide the basis for queuing formulations and are common for all infinite-source models:

1. The average number of patients being served is the ratio of arrival to service rate.

$$r = \frac{\lambda}{\mu} \quad [14.1]$$

2. The average number of patients in the system is the average number in line plus the average number being served.

$$L = L_q + r \quad [14.2]$$

3. The average time in line is the average number in line divided by the arrival rate.

$$W_q = \frac{L_q}{\lambda} \quad [14.3]$$

4. The average time in the system is the sum of the time in line plus the service time.

$$W = W_q + \frac{1}{\mu} \quad [14.4]$$

5. System utilization is the ratio of arrival rate to service capacity.

$$\rho = \frac{\lambda}{s\mu} \quad [14.5]$$

Single Channel, Poisson Arrival, and Exponential Service Time (M/M/1). The simplest model represents a system that has one server (or possibly a single surgical team). The queue discipline is first-come, first-served, and it is assumed that the customer arrival rate can be approximated by a Poisson distribution, and service time by a negative exponential distribution, or Poisson service rate. The length of queue can be endless, just as the demand for medical services is. The formulas (performance measures) for the single-channel model are as follows:

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} \quad [14.6]$$

$$P_0 = 1 - \frac{\lambda}{\mu} \quad [14.7]$$

$$P_n = P_0 \left(\frac{\lambda}{\mu} \right)^n \quad [14.8]$$

or

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

Once arrival (λ) and service (μ) rates are determined, length of the queue (L_q), probability of no arrival (P_0), and n arrivals (P_n) can be determined easily from the formulas.

EXAMPLE 14.1

A hospital is exploring the level of staffing needed for a booth in the local mall, where they would test and provide information on diabetes. Previous experience has shown that, on average, every fifteen minutes a new person approaches the booth. A nurse can complete testing and answering questions, on average, in twelve minutes. If there is a single nurse at the booth, calculate system performance measures including the probability of idle time and of one or two persons waiting in the queue. What happens to the utilization rate if another workstation and nurse are added to the unit?

Solution: Arrival rate: $\lambda = 1(\text{hour}) \div 15 = 60(\text{minutes}) \div 15 = 4$ persons per hour. Service rate: $\mu = 1(\text{hour}) \div 12 = 60(\text{minutes}) \div 12 = 5$ persons per hour. Using formula [14.1], we get:

$$r = \frac{\lambda}{\mu} = \frac{4}{5} = .8 \text{ average persons served at any given time.}$$

Then using formula [14.6], we obtain

$$L_q = \frac{4^2}{5(5 - 4)} = 3.2 \text{ persons waiting in the queue.}$$

Formula [14.2] helps us to calculate number of persons in the system as:

$$L = L_q + \frac{\lambda}{\mu} = 3.2 + .8 = 4 \text{ persons.}$$

Using formulas [14.3] and [14.4] we obtain wait times:

$$W_q = \frac{L_q}{\lambda} = \frac{3.2}{4} = 0.1067 = 6.4 \text{ minutes of waiting time in the queue}$$

$$W = W_q + \frac{1}{\mu} = 6.4 + \frac{60}{5} = 6.4 + 12 = 18.4 \text{ minutes in the system (waiting and service).}$$

Using formulas [14.7] and [14.8], we calculate queue lengths of zero, one, and two persons:

$$P_0 = 1 - \frac{\lambda}{\mu} = 1 - \frac{4}{5} = 1 - .8 = .2, \text{ 20 percent probability of idle time}$$

$$P_1 = P_0 \left(\frac{\lambda}{\mu} \right)^1 = (.2) \left(\frac{4}{5} \right)^1 = (.2)(.8)^1 = (.2)(.8) = .16 \text{ or 16\%}$$

$$P_2 = P_0 \left(\frac{\lambda}{\mu} \right)^2 = (.2) \left(\frac{4}{5} \right)^2 = (.2)(.8)^2 = (.2)(.64) = .128 \text{ or 12.8\%}.$$

Finally, using formula [14.5] for utilization of servers:

Current system utilization ($s = 1$);

$$\rho = \frac{\lambda}{s\mu} = \frac{4}{1 * 5} = 80\%.$$

System utilization with an additional nurse ($s = 2$);

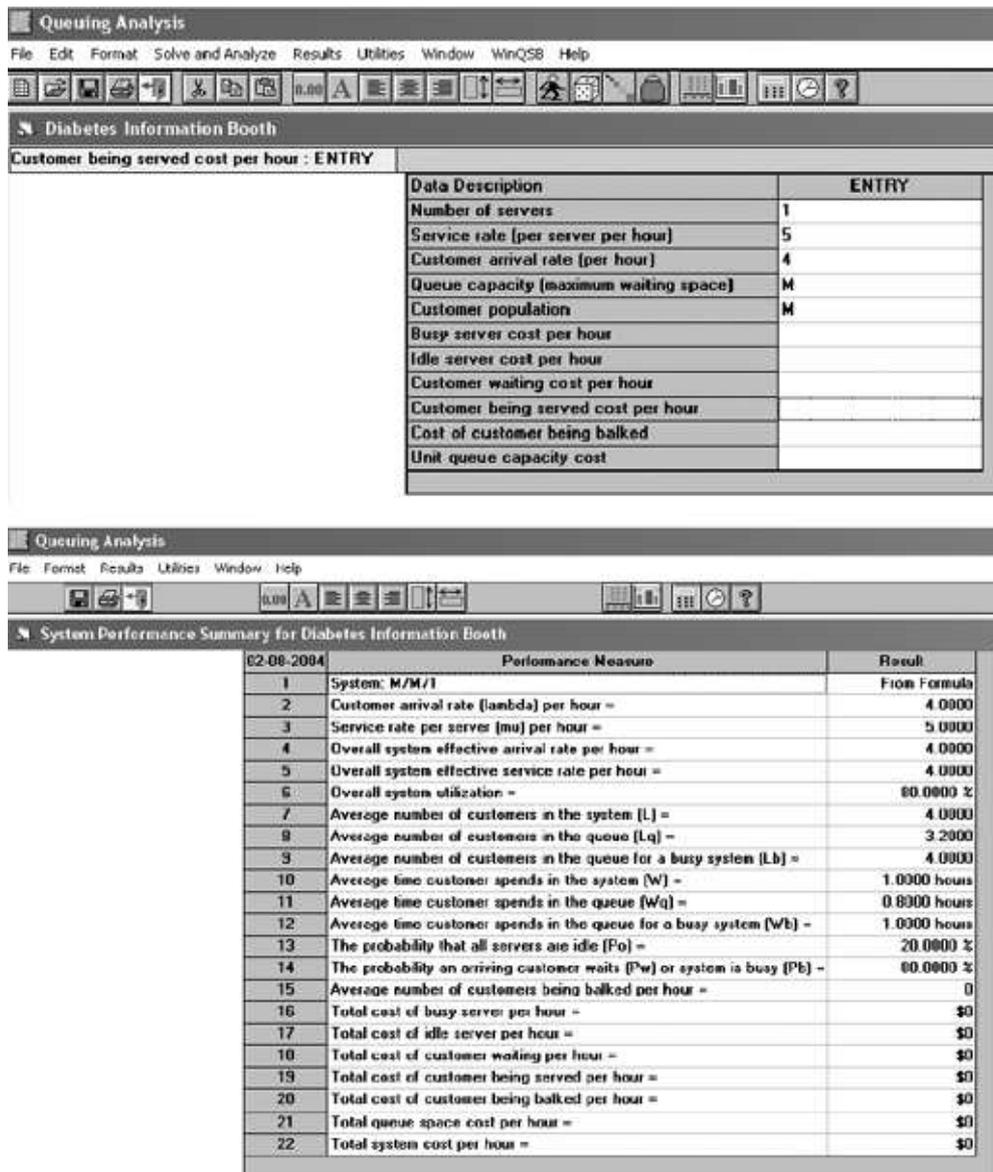
$$\rho = \frac{\lambda}{s\mu} = \frac{4}{2 * 5} = 40\%.$$

System utilization decreases as we add more resources to it. ■

In M/M/1 queue models, arrival time cannot be greater than service time. Since there is only one server, the system can tolerate up to 100% utilization. If arrival rates are more than service rates, then a multi-channel queue system is appropriate.

Software Solution. Using WinQSB software, a simple M/M/1 queue problem set-up and solution are shown in Figure 14.10. Readers can observe the same system performance measures as obtained via the formulas above, from the WinQSB

FIGURE 14.10. WINQSB SETUP AND SOLUTION FOR DIABETES INFORMATION BOOTH PROBLEM.



Source: Screen shots reprinted by permission from Microsoft Corporation and Yih-Long Chang (author of WinQSB).

FIGURE 14.11. WINQSB SYSTEM PROBABILITY SUMMARY FOR DIABETES INFORMATION BOOTH.

n	Estimated Probability of n Customers in the System	Cumulative Probability
0	0.2000	0.2000
1	0.1600	0.3600
2	0.1280	0.4880
3	0.1024	0.5904
4	0.0819	0.6723
5	0.0655	0.7379
6	0.0524	0.7903
7	0.0419	0.8322
8	0.0336	0.8658
9	0.0268	0.8926
10	0.0215	0.9141

Source: Screen shots reprinted by permission from Microsoft Corporation and Yih-Long Chang (author of WinQSB).

output. The probabilities for numbers of persons in the system at any given time are displayed in Figure 14.11.

Queuing analysis formulations for more than one server and other extensions require intensive formulations. Hand-solving such problems is beyond both the intent of this text and the time available to health care managers. However, using WinQSB, one can employ such higher-order models for their capacity formulations and for measuring existing and redesigned systems' performance.

Multi-Channel, Poisson Arrival, and Exponential Service Time (M/M/s > 1).

Expanding on Example 14.1: The hospital found that among the elderly, this free service had gained popularity, and now, during weekday afternoons, arrivals occur on average every 6 minutes 40 seconds (or 6.67 minutes), making the effective arrival rate 9 per hour. To accommodate the demand, the booth is staffed with two nurses working during weekday afternoons at the same average service rate. What are the system performance measures for this situation?

Solution: This is an M/M/2 queuing problem. The WinQSB solution provided in Figure 14.12 shows the 90% utilization. It is noteworthy that now each person has to wait on average one hour before they can be served by any of the nurses. On the basis of these results, the health care manager may consider expanding the booth further during those hours.

FIGURE 14.12. WINQSB SYSTEM PERFORMANCE FOR EXPANDED DIABETES INFORMATION BOOTH.

02-08-2004	Performance Measure	Result
1	System: M/M/2	From Formula
2	Customer arrival rate (λ) per hour =	9.0000
3	Service rate per server (μ) per hour =	5.0000
4	Overall system effective arrival rate per hour =	9.0000
5	Overall system effective service rate per hour =	9.0000
6	Overall system utilization =	90.0000 %
7	Average number of customers in the system (L) =	9.4737
8	Average number of customers in the queue (Lq) =	7.6737
9	Average number of customers in the queue for a busy system (Lb) =	9.0000
10	Average time customer spends in the system (W) =	1.0526 hours
11	Average time customer spends in the queue (Wq) =	0.8526 hours
12	Average time customer spends in the queue for a busy system (Wb) =	1.0000 hours
13	The probability that all servers are idle (Po) =	5.2632 %
14	The probability an arriving customer waits (Pw) or system is busy (Pb) =	85.2632 %
15	Average number of customers being balked per hour =	0
16	Total cost of busy server per hour =	\$0
17	Total cost of idle server per hour =	\$0
18	Total cost of customer waiting per hour =	\$0
19	Total cost of customer being served per hour =	\$0
20	Total cost of customer being balked per hour =	\$0
21	Total queue space cost per hour =	\$0
22	Total system cost per hour =	\$0

Source: Screen shots reprinted by permission from Microsoft Corporation and Yih-Long Chang (author of WinQSB).

The M/M/3 solution of adding another workstation staffed with a nurse is shown in Figure 14.13. Increasing the capacity of the system from two to three servers improves the system performance measures significantly. Now, with three nurses, the average wait is reduced from .8526 hour (51 minutes) to .0591 hour (3.5 minutes), and the total time spent in the system is now 15.5 minutes, compared to 63 minutes (1.0526 hours) with two nurses. Of course, the expansion also reduces congestion in system utilization, which used to be 90 percent, and is now at 60 percent. While these are improvements in the system, the probability of idle time for the nurses increases from 5.2 percent to 14.5 percent.

Up to this point we have explored system performance measures, but not considered costs. Health care information booths are marketing tools for health care organizations and as such should be assessed for cost-effectiveness. The health care manager should assess the impact of not serving potential patients appropriately (long waiting times in booths creating dissatisfied customers) against

FIGURE 14.13. WINQSB SYSTEM PERFORMANCE SUMMARY FOR EXPANDED DIABETES INFORMATION BOOTH WITH M/M/3.

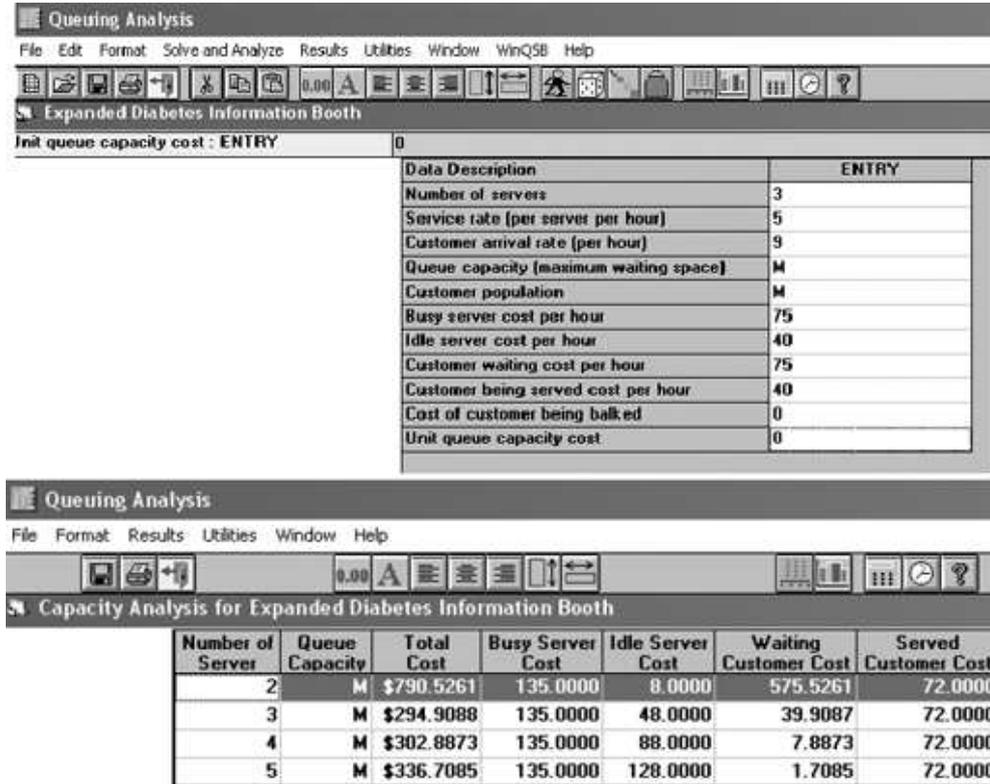
02-08-2004	Performance Measure	Result
1	System: M/M/3	From Formula
2	Customer arrival rate (λ) per hour =	9.0000
3	Service rate per server (μ) per hour =	5.0000
4	Overall system effective arrival rate per hour =	9.0000
5	Overall system effective service rate per hour =	9.0000
6	Overall system utilization =	60.0000 %
7	Average number of customers in the system (L) =	2.3321
8	Average number of customers in the queue (Lq) =	0.5321
9	Average number of customers in the queue for a busy system (Lb) =	1.5000
10	Average time customer spends in the system (W) =	0.2591 hours
11	Average time customer spends in the queue (Wq) =	0.0591 hours
12	Average time customer spends in the queue for a busy system (Wb) =	0.1667 hours
13	The probability that all servers are idle (Po) =	14.5985 %
14	The probability an arriving customer waits (Pw) or system is busy (Pb) =	35.4744 %
15	Average number of customers being balked per hour =	0
16	Total cost of busy server per hour =	\$0
17	Total cost of idle server per hour =	\$0
18	Total cost of customer waiting per hour =	\$0
19	Total cost of customer being served per hour =	\$0
20	Total cost of customer being balked per hour =	\$0
21	Total queue space cost per hour =	\$0
22	Total system cost per hour =	\$0

Source: Screen shots reprinted by permission from Microsoft Corporation and Yih-Long Chang (author of WinQSB).

the capacity costs of the information booth (setting up terminals and staffing the booth with nurses).

Assuming an operational cost per hour of \$40 (for the idle or busy server), and customer waiting costs (or busy-server cost) of \$75/hour, one can evaluate the best capacity alternative for this problem. Figure 14.14 shows the data entry and solution to the optimum capacity calculation. The total costs column for each server number is the sum of costs associated with servers and customers. Here, the total cost of two servers ($s = 2$) is \$790.53; with three servers the total cost goes down to \$298.88 per hour; and when capacity increased to four servers, the total cost per hour increases to \$302.89. Table 14.1 provides a summary of the M/M/s queue performance results under three capacities ($s = 2, 3, 4$) so the health care manager can evaluate them and conclude on a capacity. Three servers not only provide the minimum total cost per hour for this system, but also suggest a reasonable waiting time, queue length, and utilization rate. Thus, the optimal solution for the “diabetes information booth” capacity decision is three servers.

FIGURE 14.14. WINQSB CAPACITY ANALYSIS.



Source: Screen shots reprinted by permission from Microsoft Corporation and Yih-Long Chang (author of WinQSB).

TABLE 14.1. SUMMARY ANALYSIS FOR M/M/S QUEUE FOR DIABETES INFORMATION BOOTH.

Performance Measure	2 Stations	3 Stations	4 Stations
Patient arrival rate	9	9	9
Service rate	5	5	5
Overall system utilization	90%	60%	45%
L (system)	9.5	2.3	1.9
L_q	7.7	0.5	0.1
W (system) – in hours	1.05	.26	0.21
W_q – in hours	0.85	.06	0.01
P_o (idle)	5.3%	14.6%	16.2%
P_w (busy)	85.3%	35.5%	12.8%
Average number of patients balked	0	0	0
Total system cost in \$ per hour	790.53	294.91	302.89

Summary

The realities of health care organizations can be abstracted and analyzed using various queuing models, of which M/M/s is the most common. The key to this abstraction is to identify the bottleneck in operations and evaluate that portion of the operation. For example, an emergency room may be responding to the needs of patients adequately during weekdays, but difficulties may be arising over the weekends and in certain hours of the evening. Then two separate models can be identified to solve the capacity requirements for those particular time periods by measuring the arrival rates at those times and the other particulars (costs).

Queue discipline is another factor especially important in health care. Health care managers must look at multiple priorities and process patient service according to the most urgent and least urgent. This problem, too, can be evaluated as separate problems in queuing situations—with varying arrival and service rates. That is, even in the same system, queue problems can be identified for different categories of patients. The case study provided at the end of the chapter is an example for this situation.

Exercises

Exercise 14.1

People call a suburban hospital's health hot line at the rate of eighteen per hour on Monday mornings; this can be described by Poisson distribution. Providing general information or channeling to other resources takes an average of three minutes per call and varies exponentially. There is one nurse agent on duty on Mondays. Determine each of the following:

- a. System utilization.
- b. Average number in line.
- c. Average time in line.
- d. Average time in the system.

Exercise 14.2

One physician on duty full time works in a hospital emergency room. Previous experience has shown that emergency patients arrive according to a Poisson distribution with an average rate of four per hour. The physician can provide emergency treatment for approximately six patients per hour. The distribution of the physician's service time is approximately a negative exponential. Assume that the queue length can be infinite with FCFS discipline. Answer the following questions.

- a. Determine the arrival and service rates.
- b. Calculate the average probability of the system utilization and idle time.
- c. Calculate the probability of no patients in the system, and the probability of three patients in the system.
- d. What are the average numbers of patients in the waiting line (L_q) and in the system (L)?
- e. What are the average times that patients will spend in the waiting line (W_q) and in the system (W)?

Exercise 14.3

On average, six nurses work per shift at a community hospital emergency service. Patients arrive at the emergency service according to Poisson distribution with a mean of six per hour. Service time is exponential, with a mean of thirty minutes per patient. Assume that there is one patient per nurse. Find each of the performance measures listed below, using WinQSB.

- a. Compute the average number of patients in a queue.
- b. Compute the probability of zero units in the system.
- c. Compute the average waiting time for patients in a queue and in the system.
- d. Compute the system utilization rate.
- e. On the weekend, emergency service averages four patients per hour, and the service rate is expected to be forty minutes. How many nurses will be needed to achieve an average time in line of thirty minutes or less?

Exercise 14.4

Ocean View General Hospital operates five cardiac catheterization labs. The hours of operation are ideally 7:00 A.M. to 4:30 P.M., but because of the nature of the work, the day doesn't end until all scheduled cases are completed. Patients are scheduled in the labs in ninety-minute time slots. Although each cardiologist performs at his or her own rate, the average time requirement for a diagnostic study is sixty minutes, and an interventional case including a stent requires about ninety minutes.

Patients are rarely scheduled more than three to four days in advance, and most are scheduled about forty-eight hours before the procedure. The patient mix is 60 percent outpatient and 40 percent inpatient. Outpatients are asked to arrive two hours in advance of their scheduled time, in order to prepare them for the catheterization lab, but also to provide flexibility in the schedule if a physician finishes early and cases can be moved up. The hardest part of managing this area is the unpredictable nature of the schedule. Emergency patients with an acute myocardial infarction have priority and are immediately taken into the lab, bumping scheduled patients.

Each lab is staffed with a team of three or four members, who are responsible for the care of the patient during the procedure and also for room turnover. They are not responsible for recovery of the patient or for the line-extraction process. This system enables them to turn the room over for the next patient in fifteen to twenty minutes postprocedure, increasing lab throughput.

The recovery room has fourteen staff assigned to cardiac catheterization patients. The lab performs 8,052 procedures per year and is open for 234 days a year. Catheterization labs are open for six hours of operations. Patients stop arriving two hours before the last scheduled case. The average procedure time is 87 minutes.

The following costs are associated with catheterization lab:

1. Cost of the catheterization team idle: the four-member team of Registered Cardiovascular Invasive Specialists with an hourly rate of \$22.00/hour. Total: \$88.00/hour.
2. Cost of waiting: the cost of care provided in the preprocedure area. A two-person team of an RN and an EMT are able to care for six preprocedure patients waiting to go to the catheterization lab. Hourly salaries: RN: \$28.00 and EMT: \$12.00. Total: \$6.66/hour.
3. Cost of customers being served: the average cost of performing a cardiac catheterization procedure is \$800.00.
4. Cost of customers being balked: the average reimbursement for a cardiac catheterization procedure is \$1,500.

Using WinQSB, determine the optimum capacity for the Ocean View General Hospital's catheterization labs.

Exercise 14.5

A major operation in an outpatient medical office is answering the telephones. This is especially true in primary care, such as pediatrics. Patients mostly use the telephone to communicate with the physician's office. In pediatrics, such interactions include calling for appointments, refills, medical advice, referrals, and forms (for example: school forms, camp forms.) Because of the frequent use of the telephone in outpatient pediatrics, it is an important focus for assessing productivity and efficiency.

A pediatric practice consists of nine physicians and two nurse practitioners. The practice has two offices. The patient population is approximately ten thousand children, with nearly fifty thousand visits per year. The phone system consists of sixteen telephone lines, most of them at the main office.

As the practice has grown, there have been increasing complaints from patients about wait time on the phone lines. All incoming calls are routed to the main office. When a patient dials the practice's office telephone number, a voicemail system directs the caller to press a number according to the purpose of the call (for example, "Press 'one' for appointments.") The system also distributes the phone calls according to whether the person calling is a patient, physician, laboratory, or hospital.

During the winter months, when the volume of sick patients is highest, a patient's wait can sometimes be as long as ten to fifteen minutes on the appointment line before speaking to a person. Since most customer service guidelines recommend telephone hold times no longer than one minute, this is an area that greatly needs improvement.

Telephone calls form a single waiting line and are served on a first-come, first-served basis. Arrival rates can be described by Poisson distribution, and service times can be described by negative exponential distribution. With these characteristics, a multiple-channel model for queuing analysis is most appropriate.

The queuing analysis of the practice's phone system can be divided into three parts of the workday, which lasts from 8:00 A.M. to 5:00 P.M. For the first hour of the day (8:00 A.M. to 9:00 A.M.) there are usually three receptionists working to answer telephone calls only. For the last hour of the day (4:00 P.M. to 5:00 P.M.), there are usually five receptionists answering phones as well as checking patients in and out. For the bulk of the day, there are usually six receptionists working. The use of fewer servers during the first and last hours is primarily because fewer patients are being seen during those hours, so fewer servers are needed for checking patients in and out.

To determine the customer arrival rate (or phone calls/hour), incoming monthly phone call data for the previous year were obtained from the telephone company (Table EX 14.5.1.)

TABLE EX 14.5.1

Month	Phone Calls
January	6,640
February	6,756
March	6,860
April	6,226
May	6,671
June	7,168
July	6,802
August	6,971
September	7,205
October	6,944
November	6,623
December	6,875
Total	81,741

From examining previous studies of the office's phone call volume distribution, it is estimated that 30% of the phone calls occur between 8 A.M. and 9 A.M.; 40% between 9 A.M. and 4 P.M., and the remaining 30% arriving from 4 P.M. to 5 P.M. (Table 14.5.2).

TABLE EX 14.5.2

Customer Arrival Rates (λ)	
8:00 A.M. to 9:00 A.M.	31 phone calls/hr
9:00 A.M. to 4:00 P.M.	42 phone calls/hr
4:00 P.M. to 5:00 P.M.	31 phone calls/hr

To estimate the service rate (or phone calls/hour/receptionist), several sample studies were performed by an office administrator. It is important to note that the receptionists perform functions other than answering phones, such as checking patients in and out. Therefore, the number of phone calls that a server can answer per hour depends on the other responsibilities that the person has that day. In order to arrive at a service rate, the assumption was made that the average maximum of phone calls per hour for the sample days would represent the servers operating at the maximum phone-call-answering capacity when having other responsibilities. While this assumption may underestimate actual server rate, for purposes of this study, the conservative estimate is acceptable in the absence of further data.

There is one exception to this assumption. During the first hour of the day, from 8:00 A.M. to 9:00 A.M., patients are not yet being seen in the office. Therefore, during that hour the servers have a faster telephone service rate, since they have no other primary duties (Table EX 14.5.3). From samples studied, we have determined that the maximum service capability when only answering phones is approximately four minutes per phone call, or fifteen calls per hour per server. This number was used for the service rate for the first hour (8:00 A.M. to 9:00 P.M.).

TABLE EX 14.5.3

Service Rate (μ)	
8:00 A.M. to 9:00 A.M.	15 phone calls/hr*
9:00 A.M. to 4:00 P.M.	8 phone calls/hr
4:00 P.M. to 5:00 P.M.	8 phone calls/hr

Cost studies were then performed based on the financial data from the previous year. Capacity costs were calculated based on salary and benefits per server and a percentage of the equipment maintenance, phone line costs, rent, and other capital expenditures (Table EX 14.5.4). With a total of fifty employees and a total of thirty full-time equivalents (FTEs), the portion of capital expenditures was determined as 1/30 of costs. Phone line charges were determined by a per-line charge, since one server would utilize one line each day.

TABLE EX 14.5.4

Total Hourly Cost for Busy Server Summary	
Salary	\$13.00
Benefit	\$3.75
Telephone Charges	\$4.73
Capital Expenses	\$4.83
Total Hourly Cost for Busy Server	\$26.31

Capacity cost or busy server cost would be equivalent to idle server cost. Regardless of whether or not the receptionist is answering the phone, she is paid the same salary and benefits and is using the same space and utilities. In addition, the practice must pay the phone line and equipment maintenance charges, regardless of usage.

For calculation purposes, a value was assigned to the cost to the customer of waiting. A value of \$50/hour was assigned to customer waiting costs. In reality, though, customer waiting costs are likely to vary with the length of time waited, with a steep exponential increase in cost to the patient for longer times waited.

The cost of being balked would represent a lost patient if a patient's call was not answered. In pediatrics, the patients generally prefer continuity of care throughout their child's life. Therefore, a truly balked customer might represent a child's lifetime worth of visits. However, one might also define a balked customer as one who will not come for a visit that day because the phone call was not answered promptly. This person would be likely to return to the practice on another day if he or she established a doctor-patient relationship with the practice. Therefore, for the purposes of this model, it is assumed that the cost of being balked is the lost revenue from an office visit, which is approximately \$80. (See Table EX 14.5.5)

TABLE EX 14.5.5

Cost Summary	
Busy server cost/hr	\$26.31
Idle server cost/hr	\$26.31
Customer waiting cost/hr	\$50
Cost of customer being balked	\$80

Using WinQSB, perform a queuing analysis for the pediatric practice's telephone system to determine the optimal server capacity for the volume of phone calls that they receive. Are there enough servers/receptionists and enough phone lines?

Exercise 14.6

An outpatient clinic that is open two hundred days/year receives twelve thousand visits per year, or approximately sixty patients per day. These visits are divided over two wings, for thirty patients per wing. Appointments are made for two three-hour sessions per day. Thus, the patient arrival rate averages five patients per hour, $(\lambda) = 5$ patients/hour.

By observing the check-in and checkout processes, the service rate can be determined. Each check-in requires that the patient be retrieved from the waiting area, contact and insurance information be reviewed, and possibly a copay collected. This process takes approximately ten minutes, so the service rate for check-ins is six patients per hour. Thus, the service rate for check-ins is $(\mu) = 6$ patients/hour.

Checkout takes approximately twenty minutes per patient and involves scheduling follow-up appointments, ordering tests, and answering questions. Hence the service rate for check-outs is $(\mu) = 3$ patients/hour.

Currently, three employees perform the administrative duties. One is assigned check-in duties, and the other two mostly check out patients. All are cross-trained on both roles, and in reality the staff varies from day to day in who performs which role.

To address the long wait times, the clinic administrator wants to evaluate hiring additional staff members. Assuming that both service rates approximate Poisson distribution, and using WinQSB, calculate the optimal staffing pattern for the clinic and the system performance measures.

Exercise 14.7

Emergency room use at "SAVE-ME!!" Hospital peaks on Saturday nights during the period from 7:00 P.M. to 2:00 A.M. Historically, the hospital has provided space for three stations (examining rooms) for non-emergency cases and two stations for emergency cases during that period. Non-emergency patients are examined on a first-come, first-served basis, and emergency cases are treated on a most-serious, first-served basis, after a triage nurse has screened all cases. An area competitor hospital recently announced discontinuation of emergency services within six months. "SAVE-ME!!" estimates that current arrival patterns during the 7:00 P.M.–2:00 A.M. period would increase by one-third for non-emergency cases and would double for emergency cases. The hospital wants to know how additional resources in the ER might reduce congestion and waiting time, as well as the overall cost of operations, for non-emergency and for emergency patients.

The past year's operating data were gathered from the information systems; they included records of arrival and service times. Preliminary examination of the data revealed little seasonal variation in ER use for that year, and ER personnel stated that their protocols and procedures had remained relatively constant since the reorganization of the ER two years ago.

The arrival pattern of patients, tabulated for twenty Saturday nights (total of one hundred hours), showed that 900 non-emergency and 150 emergency patients came to the ER during that time. The arrival pattern approximates a Poisson distribution. "SAVE-ME!!" has a queue holding capacity of two for emergency cases, and their non-emergency room has twelve chairs in the waiting area. After a lengthy time-motion study, the average service time was found to be thirty minutes per non-emergency and seventy-five minutes per emergency patient. A separate study conducted by the finance/accounting department provided estimates for relevant costs as shown in Table EX 14.7.

- a. Using WinQSB, analyze both the emergency and the non-emergency capacity requirements for current conditions and for six months later, and fill in the Performance Evaluation Table on the following page.

TABLE EX 14.7

Cost Type\Patient Type	Non-Emergency	Emergency
Busy server cost/hr	100	200
Idle server cost/hr	450	800
Customer waiting cost/hr	200	400
Customer being served cost/hr	100	300
Cost of customer being balked	600	1,200
Unit queue capacity cost (holding or seating)	25	50

PERFORMANCE EVALUATION TABLE

Performance Measure	Non-Emergency			Emergency		
	Current Capacity 3-Stations	Optimal Capacity ?-Stations	6-Months Optimal ? Stations	Current Capacity 2-Stations	Optimal Capacity ? Stations	6-Months Optimal ? Stations
Patient arrival rate						
Service rate						
Overall system utilization						
L (system)						
L_q						
W (system)						
W_q						
P_o (idle)						
P_w						
Avg. # patients balked						
Total system cost						

Note: Replace “?” marks on the table with optimal capacity.

- b. Recommend the number of examining rooms for current and the future conditions on the basis of the above performance-evaluation statistics.

Hint:

1. Evaluate non-emergency, after data entry:
 - a) Use Solve and Analyze, and Solve the Performance, and print the results (from those, fill the first column in the Performance Evaluation Table above).
 - b) Use Solve and Analyze, and Perform Capacity Analysis, setting Number of Servers for “Start from” to 1, “End at” to 10, “Step” to 1; set Queue Capacity for “Start from” to 12, “End at” to 12, “Step” to 1.

Observe the Total Cost column on the results table, and determine the optimal server capacity based on the lowest total cost. You can set the cursor to the “total cost” column and click on “graph” to observe the lowest cost server capacity; this will be the Optimal Capacity.

2. Plug in the optimal capacity (number of servers) into the original data entry, and repeat Step 1 a) above. (Then fill in the second column of the Performance Evaluation Table.)
3. Increase the arrival rate for non-emergency cases by one-third and repeat Steps 1-b and 2. Report only the optimal capacity performance statistics on column 3 of the Performance Evaluation Table.
4. Repeat Steps 1 through 3 for emergency cases, and fill the remaining three columns for Emergency, on the Performance Evaluation Table. (Note: Use Solve and Analyze and Perform Capacity Analysis. Set Number of Servers "Start from" to 1, "End at" to 6, "Step" to 1. Set Queue Capacity "Start from" to 2, "End at" to 2, "Step" to 1.) Remember to increase the arrival rate twice for the six-month situation.