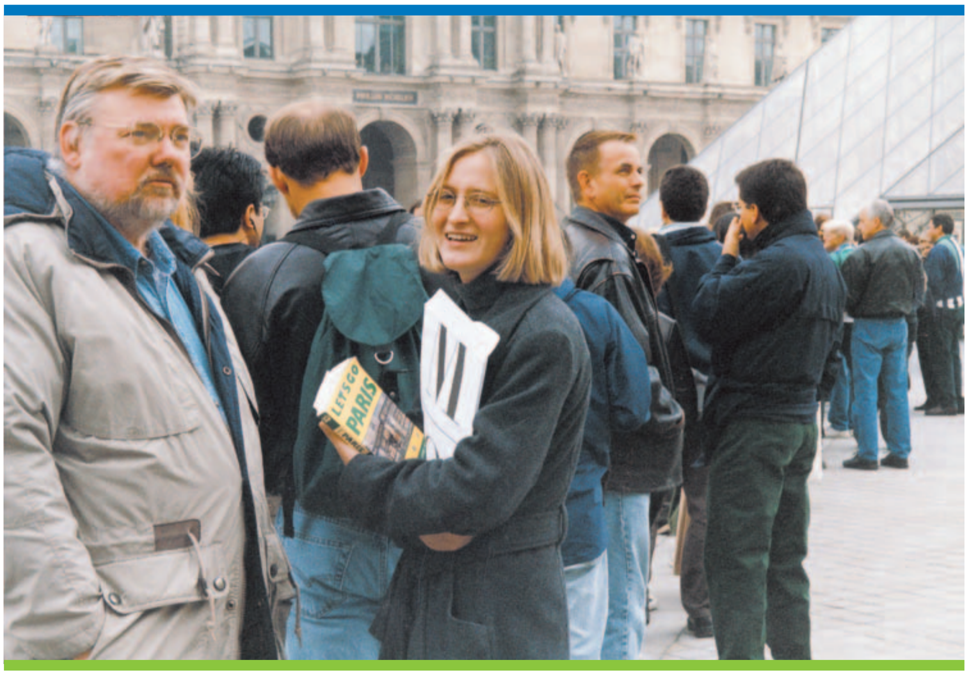
technical note six
WAITING LINE MANAGEMENT

technical note



We've all had to wait in lines and know that whatever line we choose, the other ones will go faster. Here people line up for the entrance to the Louvre in Paris, France.

Understanding waiting lines or **queues** and learning how to manage them is one of the most important areas in operations management. It is basic to creating schedules, job design, inventory levels, and so on. In our service economy we wait in line every day, from driving to work to checking out at the supermarket. We also encounter waiting lines at factories—jobs wait in lines to be worked on at different machines, and machines themselves wait their turn to be overhauled. In short, waiting lines are pervasive.

Queues

In this technical note we discuss the basic elements of waiting line problems and provide standard steady-state formulas for solving them. These formulas, arrived at through queuing theory, enable planners to analyze service requirements and establish service facilities appropriate to stated conditions. Queuing theory is broad enough to cover such dissimilar delays as those encountered by customers in a shopping mall or aircraft in a holding pattern awaiting landing slots. Recently, Internet access providers have had problems providing enough modem telephone lines for subscribers dialing into the Internet. This problem can also be analyzed by queuing models.

ECONOMICS OF THE WAITING LINE PROBLEM

The central problem in virtually every waiting line situation is a trade-off decision. The manager must weigh the added cost of providing more rapid service (more traffic lanes, additional landing strips, more checkout stands) against the inherent cost of waiting.

Frequently, the cost trade-off decision is straightforward. For example, if we find that the total time our employees spend in the line waiting to use a copying machine would otherwise be spent in productive activities, we could compare the cost of installing one additional machine to the value of employee time saved. The decision could then be reduced to dollar terms and the choice easily made.

Vol. IX
“Queuing—Apropos
Technology NAR”

On the other hand, suppose that our waiting line problem centers on demand for beds in a hospital. We can compute the cost of additional beds by summing the costs for building construction, additional equipment required, and increased maintenance. But what is on the other side of the scale? Here we are confronted with the problem of trying to place a dollar figure on a patient's need for a hospital bed that is unavailable. While we can estimate lost hospital income, what about the human cost arising from this lack of adequate hospital care?

COST-EFFECTIVENESS BALANCE

Exhibit TN6.1 shows the essential trade-off relationship under typical (steady-state) customer traffic conditions. Initially, with minimal service capacity, the waiting line cost is at a maximum. As service capacity is increased, there is a reduction in the number of customers in the line and in their waiting times, which decreases waiting line cost. The variation in this function is often represented by the negative exponential curve. The cost of installing service capacity is shown simplistically as a linear rather than step function. The aggregate or total cost is shown as a U-shaped curve, a common approximation in such equilibrium problems. The idealized optimal cost is found at the crossover point between the service capacity and waiting line curves.

THE PRACTICAL VIEW OF WAITING LINES

Before we proceed with a technical presentation of waiting line theory, it is useful to look at the intuitive side of the issue to see what it means. Exhibit TN6.2 shows arrivals at a service facility (such as a bank) and service requirements at that facility (such as tellers and loan officers). One important variable is the number of arrivals over the hours that the service system is open. From the service delivery viewpoint, customers demand varying amounts of service, often exceeding normal capacity. We can control arrivals in a variety of ways. For example, we can have a short line (such as a drive-in at a fast-food restaurant with only several spaces), we can establish specific hours for specific customers, or we can run specials. For the server, we can affect service time by using faster or slower servers, faster or slower machines, different tooling, different material, different layout, faster setup time, and so on.

EXHIBIT TN6.1

Service Capacity versus Waiting Line Trade-Off

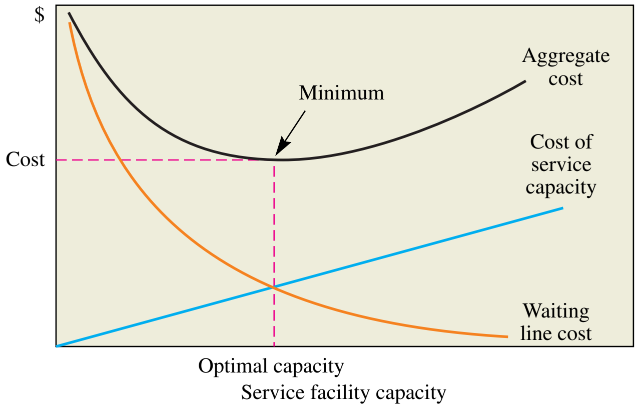
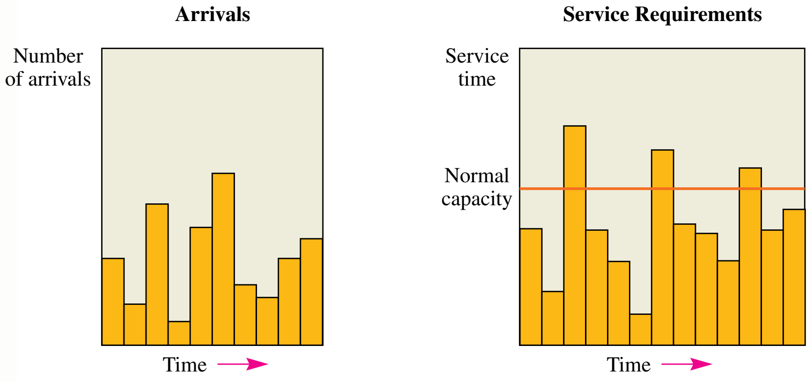


EXHIBIT TN6.2

Arrival and Service Profiles



Suggestions for Managing Queues

The following are some useful suggestions for managing queues that go beyond the quantitative waiting line models.

1 **Determine an acceptable waiting time for your customers.** How long do your customers expect to wait? Set operational objectives based on what is acceptable.

2 **Try to divert your customer's attention when waiting.** Providing music, a video, or some other form of entertainment may help to distract the customers from the fact they are waiting.

3 **Inform your customers of what to expect.** This is especially important when the waiting time will be longer than normal. Tell them why the waiting time is longer than normal and what you are doing to alleviate the queue.

4 **Keep employees not serving the customers out of sight.** Nothing is more frustrating to someone waiting in line than to see employees, who potentially could be serving those in line, working on other activities.

5 **Segment customers.** If a group of customers needs something that can be done very quickly, give them a special line so they do not have to wait for the slower customers.

6 **Train your servers to be friendly.** Greeting the customer by name, or providing some other special attention, can go a long way toward overcoming the negative feeling of a long wait. [Hint: Rather than servers being told to just "be friendly," psychologists suggest they be told when to invoke specific friendly actions such as smiling—when greeting customers, when taking orders, and when giving change (in a convenience store). Tests using such specific behavioral actions have shown significant increases in perceived friendliness of the servers in the eyes of the customer.]

7 **Encourage customers to come during the slack periods.** Inform customers of times when they usually would not have to wait; also tell them when the peak periods are—this may help to smooth the load.

8 **Take a long-term perspective toward getting rid of the queues.** Develop plans for alternative ways to serve your customers. Where appropriate, develop plans for automating or speeding up the process in some manner. This is not to say you want to eliminate personal attention; some customers expect this.

Source: Based on K. Katz, B. M. Larson, and R. C. Larson, "Prescription for the Waiting-in-Line Blues," *Sloan Management Review,* Winter 1991, pp. 51–52.

The essential point is waiting lines are *not* a fixed condition of a productive system but are to a very large extent within the control of the system management and design. Professor Richard Larson (the famous "wait-watcher") and his colleagues offer useful suggestions for managing queues based on their research in the banking industry. (See the box titled "Suggestions for Managing Queues.")

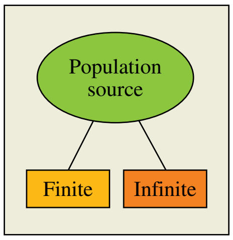
The Queuing System

Queuing system

The **queuing system** consists essentially of three major components: (1) the source population and the way customers arrive at the system, (2) the servicing system, and (3) the condition of the customers exiting the system (back to source population or not?), as seen in Exhibit TN6.3. The following sections discuss each of these areas.

Customer Arrivals

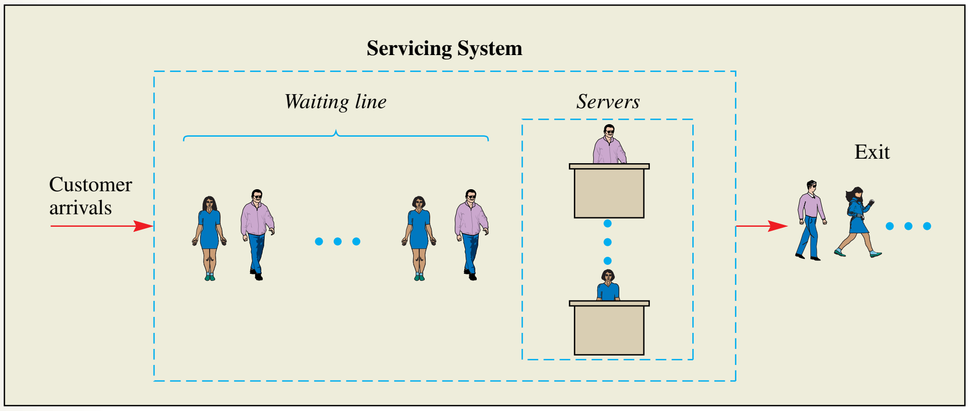
Arrivals at a service system may be drawn from a *finite* or an *infinite* population. The distinction is important because the analyses are based on different premises and require different equations for their solution.



Finite Population A *finite population* refers to the limited-size customer pool that will use the service and, at times, form a line. The reason this finite classification is important is that when a customer leaves its position as a member for the population (a machine breaking down and requiring service, for example), the size of the user group is reduced by one, which reduces the probability of the next occurrence. Conversely, when a customer is serviced and returns to the user group, the population increases and the probability of a user requiring service also increases. This finite class of problems requires a separate set of formulas from that of the infinite population case.

EXHIBIT TN6.3

Components of a Queuing System



As an example, consider a group of six machines maintained by one repairperson. When one machine breaks down, the source population is reduced to five, and the chance of one of the remaining five breaking down and needing repair is certainly less than when six machines were operating. If two machines are down with only four operating, the probability of another breakdown is again changed. Conversely, when a machine is repaired and returned to service, the machine population increases, thus raising the probability of the next breakdown. A finite population model with one server that can be used in such cases is presented in Exhibits TN6.8 and TN6.10.

Infinite Population An *infinite population* is large enough in relation to the service system so that the population size caused by subtractions or additions to the population (a customer needing service or a serviced customer returning to the population) does not significantly affect the system probabilities. If, in the preceding finite explanation, there were 100 machines instead of six, then if one or two machines broke down, the probabilities for the next breakdowns would not be very different and the assumption could be made without a great deal of error that the population (for all practical purposes) was infinite. Nor would the formulas for "infinite" queuing problems cause much error if applied to a physician with 1,000 patients or a department store with 10,000 customers.

DISTRIBUTION OF ARRIVALS

When describing a waiting system, we need to define the manner in which customers or the waiting units are arranged for service.

Arrival rate

Waiting line formulas generally require an **arrival rate,** or the number of units per period (such as an average of one every six minutes). A *constant* arrival distribution is periodic, with exactly the same time between successive arrivals. In productive systems, the only arrivals that truly approach a constant interval period are those subject to machine control. Much more common are *variable* (random) arrival distributions.

In observing arrivals at a service facility, we can look at them from two viewpoints: First, we can analyze the time between successive arrivals to see if the times follow some statistical distribution. Usually we assume that the time between arrivals is exponentially distributed. Second, we can set some time length (T) and try to determine how many arrivals might enter the system within T. We typically assume that the number of arrivals per time unit is Poisson distributed.

Exponential distribution

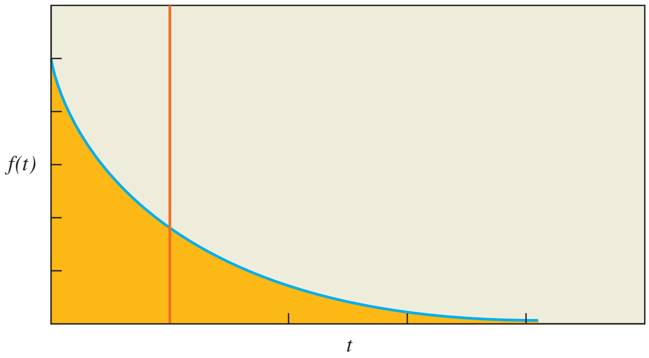
Exponential Distribution In the first case, when arrivals at a service facility occur in a purely random fashion, a plot of the interarrival times yields an **exponential distribution** such as that shown in Exhibit TN6.4. The probability function is

[TN6.1] $$f(t) = \lambda e^{-\lambda t}$$

where λ is the mean number of arrivals per time period.

EXHIBIT TN6.4

Exponential Distribution



The cumulative area beneath the curve in Exhibit TN6.4 is the summation of equation (TN6.1) over its positive range, which is $e^{-\lambda t}$. This integral allows us to compute the probabilities of arrivals within a specified time. For example, for the case of single arrivals to a waiting line ($\lambda = 1$), the following table can be derived either by solving $e^{-\lambda t}$ or by using Appendix F. Column 2 shows the probability that it will be more than t minutes until the next arrival. Column 3 shows the probability of the next arrival within t minutes (computed as 1 minus column 2)

(1) t (Minutes)	(2) Probability That the Next Arrival Will Occur in t Minutes or More (from Appendix F or Solving e^{-t})	(3) Probability That the Next Arrival Will Occur in t Minutes or Less [1 − Column (2)]
0	1.00	0
0.5	0.61	0.39
1.0	0.37	0.63
1.5	0.22	0.78
2.0	0.14	0.86

Poisson Distribution In the second case, where one is interested in the number of arrivals during some time period T, the distribution appears as in Exhibit TN6.5 and is obtained by finding the probability of exactly n arrivals during T. If the arrival process is random, the distribution is the **Poisson,** and the formula is

Poisson distribution

[TN6.2]
$$P_T(n) = \frac{(\lambda T)^n e^{-\lambda T}}{n!}$$

Equation (TN6.2) shows the probability of exactly n arrivals in time T.[1] For example, if the mean arrival rate of units into a system is three per minute ($\lambda = 3$) and we want to find the probability that exactly five units will arrive within a one-minute period ($n = 5$, $T = 1$), we have

$$P_1(5) = \frac{(3 \times 1)^5 e^{-3\times 1}}{5!} = \frac{3^5 e^{-3}}{120} = 2.025e^{-3} = 0.101$$

That is, there is a 10.1 percent chance that there will be five arrivals in any one-minute interval.

Although often shown as a smoothed curve, as in Exhibit TN6.5, the Poisson is a discrete distribution. (The curve becomes smoother as n becomes large.) The distribution is discrete because n refers, in our example, to the number of arrivals in a system, and this must be an integer. (For example, there cannot be 1.5 arrivals.)

Also note that the exponential and Poisson distributions can be derived from one another. The mean and variance of the Poisson are equal and denoted by λ. The mean of the

EXHIBIT TN6.5

Poisson Distribution for $\lambda T = 3$

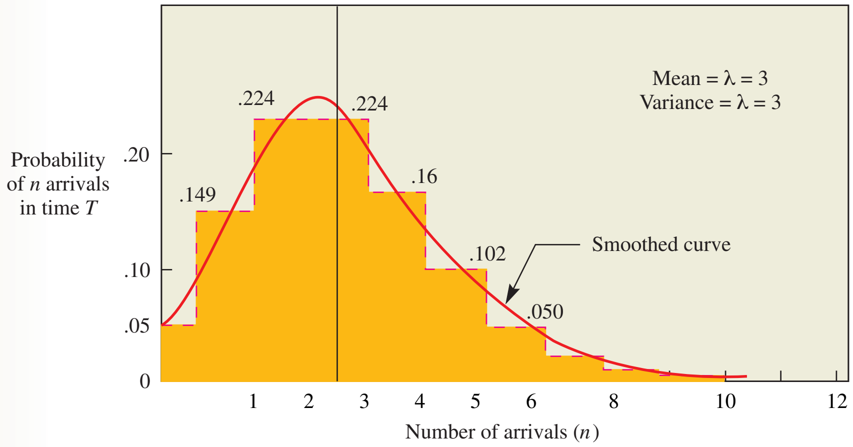
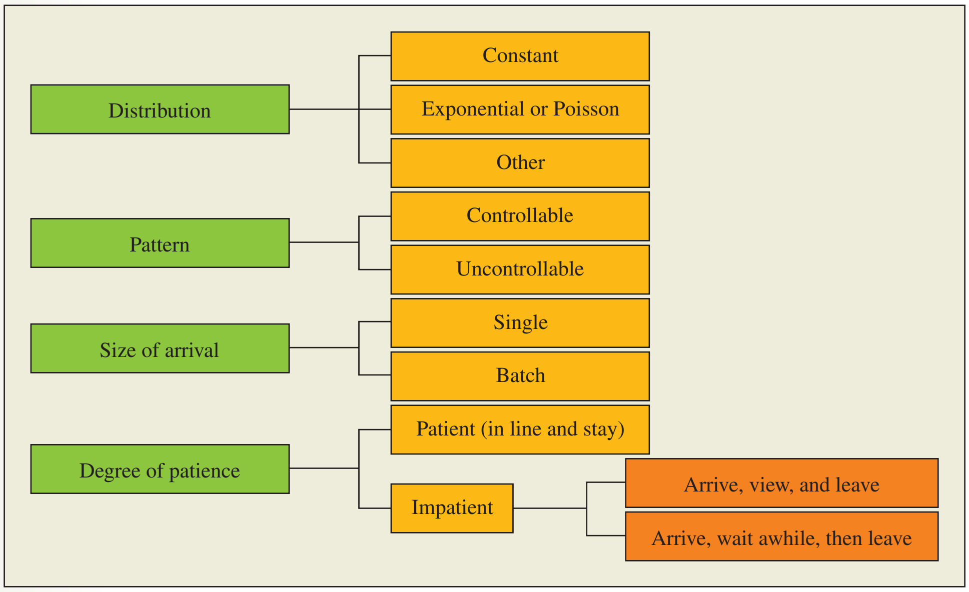


EXHIBIT TN6.6

Customer Arrivals in Queues



exponential is $1/\lambda$ and its variance is $1/\lambda^2$. (Remember that the time between arrivals is exponentially distributed and the number of arrivals per unit of time is Poisson distributed.)

Other arrival characteristics include arrival patterns, size of arrival units, and degree of patience. (See Exhibit TN6.6.)

Arrival patterns The arrivals at a system are far more controllable than is generally recognized. Barbers may decrease their Saturday arrival rate (and supposedly shift it to other days of the week) by charging an extra $1 for adult haircuts or charging adult prices for children's haircuts. Department stores run sales during the off-season or hold one-day-only sales in part for purposes of control. Airlines offer excursion and off-season rates for similar reasons. The simplest of all arrival-control devices is the posting of business hours.

Some service demands are clearly uncontrollable, such as emergency medical demands on a city's hospital facilities. But even in these situations, arrivals at emergency rooms in specific hospitals are controllable to some extent by, say, keeping ambulance drivers in the service region informed of the status of their respective host hospitals.

Size of arrival units A *single arrival* may be thought of as one unit. (A unit is the smallest number handled.) A single arrival on the floor of the New York Stock Exchange

FACTORS SUCH AS LINE LENGTH, CAPACITY, AND NUMBER OF LINES MUST BE CONSIDERED WHEN DESIGNING A QUEUING SYSTEM. MULTIPLE LINES WITH PEOPLE COUNTERS TO KEEP TRACK OF THE PARK'S CAPACITY ARE USED AT THE ENTRANCE TO SIX FLAGS MAGIC MOUNTAIN, VALENCIA, CA.

(NYSE) is 100 shares of stock; a single arrival at an egg-processing plant might be a dozen eggs or a flat of 2½ dozen; a single arrival at a restaurant is a single person.

A *batch arrival* is some multiple of the unit, such as a block of 1,000 shares on the NYSE, a case of eggs at the processing plant, or a party of five at a restaurant.

Degree of patience A *patient* arrival is one who waits as long as necessary until the service facility is ready to serve him or her. (Even if arrivals grumble and behave impatiently, the fact that they wait is sufficient to label them as patient arrivals for purposes of waiting line theory.)

There are two classes of *impatient* arrivals. Members of the first class arrive, survey both the service facility and the length of the line, and then decide to leave. Those in the second class arrive, view the situation, join the waiting line, and then, after some period of time, depart. The behavior of the first type is termed *balking,* while the second is termed *reneging*.

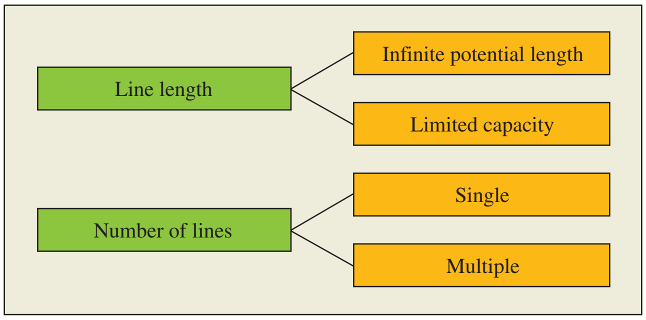
THE QUEUING SYSTEM: FACTORS

The queuing system consists primarily of the waiting line(s) and the available number of servers. Here we discuss issues pertaining to waiting line characteristics and management, line structure, and service rate. Factors to consider with waiting lines include the line length, number of lines, and queue discipline.

Length In a practical sense, an infinite line is simply one that is very long in terms of the capacity of the service system. Examples of *infinite potential length* are a line of vehicles backed up for miles at a bridge crossing and customers who must form a line around the block as they wait to purchase tickets at a theater.

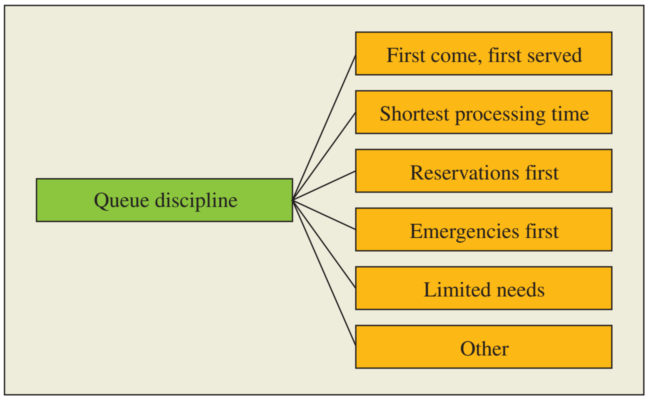
Gas stations, loading docks, and parking lots have *limited line capacity* caused by legal restrictions or physical space characteristics. This complicates the waiting line problem not only in service system utilization and waiting line computations but also in the shape of the actual arrival distribution. The arrival denied entry into the line because of lack of space may rejoin the population for a later try or may seek service elsewhere. Either action makes an obvious difference in the finite population case.

Number of lines A single line or single file is, of course, one line only. The term *multiple lines* refers to the single lines that form in front of two or more servers or to single lines that converge at some central redistribution point. The disadvantage of multiple lines in a busy facility is that arrivals often shift lines if several previous services have been of short duration or if those customers currently in other lines appear to require a short service time.



Queue discipline A queue discipline is a priority rule or set of rules for determining the order of service to customers in a waiting line. The rules selected can have a dramatic effect on the system's overall performance. The number of customers in line, the average waiting time, the range of variability in waiting time, and the efficiency of the service facility are just a few of the factors affected by the choice of priority rules.

Probably the most common priority rule is first come, first served (FCFS). This rule states that customers in line are served on the basis of their chronological arrival; no other characteristics have any bearing on the selection process. This is popularly accepted as the fairest rule, although in practice it discriminates against the arrival requiring a short service time.



Reservations first, emergencies first, highest-profit customer first, largest orders first, best customers first, longest waiting time in line, and soonest promised date are other examples of priority rules. There are two major practical problems in using any rule: One is ensuring that customers know and follow the rule. The other is ensuring that a system exists to enable employees to manage the line (such as take-a-number systems).

Service rate

Service Time Distribution Another important feature of the waiting structure is the time the customer or unit spends with the server once the service has started. Waiting line formulas generally specify **service rate** as the capacity of the server in number of units per time period (such as 12 completions per hour) and *not* as service time, which might average five minutes each. A constant service time rule states that each service takes exactly the same time. As in constant arrivals, this characteristic is generally limited to machine-controlled operations.

When service times are random, they can be approximated by the exponential distribution. When using the exponential distribution as an approximation of the service times, we will refer to μ as the average number of units or customers that can be served per time period.

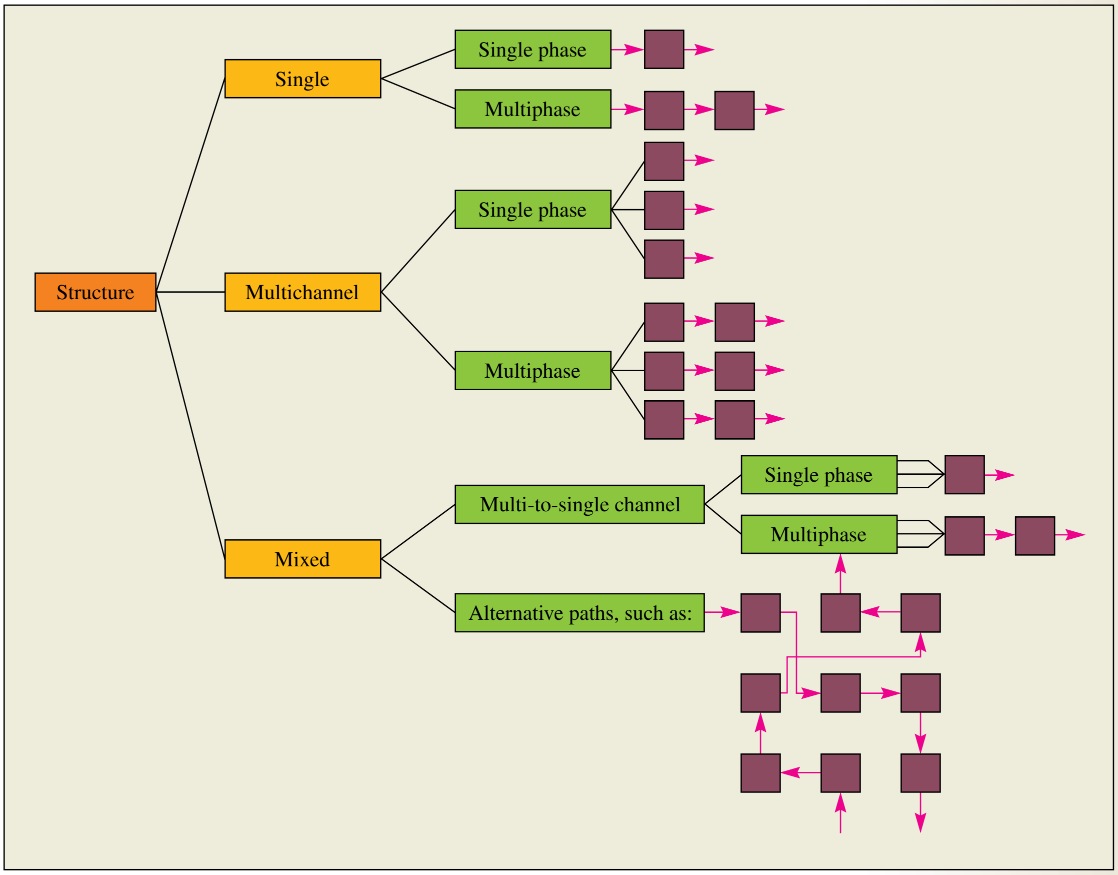
Line Structures As Exhibit TN6.7 shows, the flow of items to be serviced may go through a single line, multiple lines, or some mixtures of the two. The choice of format depends partly on the volume of customers served and partly on the restrictions imposed by sequential requirements governing the order in which service must be performed.

1 **Single channel, single phase** This is the simplest type of waiting line structure, and straightforward formulas are available to solve the problem for standard distribution patterns of arrival and service. When the distributions are nonstandard, the problem is easily solved by computer simulation. A typical example of a single-channel, single-phase situation is the one-person barbershop.

2 **Single channel, multiphase** A car wash is an illustration because a series of services (vacuuming, wetting, washing, rinsing, drying, window cleaning, and parking) is performed in a fairly uniform sequence. A critical factor in the single-channel case with service

EXHIBIT TN6.7

Line Structures



in series is the amount of buildup of items allowed in front of each service, which in turn constitutes separate waiting lines.

3 **Multichannel, single phase** Tellers' windows in a bank and checkout counters in high-volume department stores exemplify this type of structure. The difficulty with this format is that the uneven service time given each customer results in unequal speed or flow among the lines. This results in some customers being served before others who arrived earlier, as well as in some degree of line shifting. Varying this structure to ensure the servicing of arrivals in chronological order would require forming a single line, from which, as a server becomes available, the next customer in the queue is assigned.

The major problem of this structure is that it requires rigid control of the line to maintain order and to direct customers to available servers. In some instances, assigning numbers to customers in order of their arrival helps alleviate this problem.

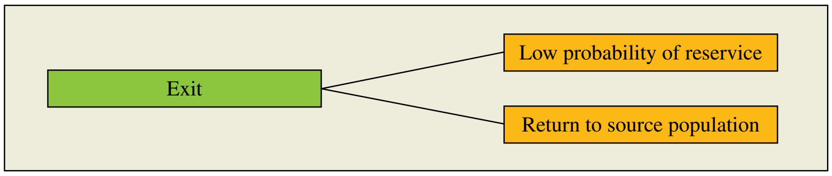
4 **Multichannel, multiphase** This case is similar to the preceding one except that two or more services are performed in sequence. The admission of patients in a hospital follows this pattern because a specific sequence of steps is usually followed: initial contact at the admissions desk, filling out forms, making identification tags, obtaining a room assignment, escorting the patient to the room, and so forth. Because several servers are usually available for this procedure, more than one patient at a time may be processed.

5 **Mixed** Under this general heading we consider two subcategories: (1) multiple-to-single channel structures and (2) alternative path structures. Under (1), we find either lines that merge into one for single-phase service, as at a bridge crossing where two lanes merge into one, or lines that merge into one for multiphase service, such as subassembly lines feeding into a main line. Under (2), we encounter two structures that differ in directional flow requirements. The first is similar to the multichannel–multiphase case, except that

(a) there may be switching from one channel to the next after the first service has been rendered and (b) the number of channels and phases may vary—again—after performance of the first service.

EXIT

Once a customer is served, two exit fates are possible: (1) The customer may return to the source population and immediately become a competing candidate for service again or (2) there may be a low probability of reservice. The first case can be illustrated by a machine that has been routinely repaired and returned to duty but may break down again; the second can be illustrated by a machine that has been overhauled or modified and has a low probability of reservice over the near future. In a lighter vein, we might refer to the first as the "recurring-common-cold case" and to the second as the "appendectomy-only-once case."



It should be apparent that when the population source is finite, any change in the service performed on customers who return to the population modifies the arrival rate at the service facility. This, of course, alters the characteristics of the waiting line under study and necessitates reanalysis of the problem.

WAITING LINE MODELS

Excel: Queue

● ● ● In this section we present four sample waiting line problems followed by their solutions. Each has a slightly different structure (see Exhibit TN6.8) and solution equation (see Exhibit TN6.10). There are more types of models than these four, but the formulas and solutions become quite complicated, and those problems are generally solved using computer simulation (see Technical Note 15). Also, in using these formulas, keep in mind that they are steady-state formulas derived on the assumption that the process under study is ongoing. Thus, they may provide inaccurate results when applied to processes where the arrival rates and/or service rates change over time. The Excel Spreadsheet Queue.xls, developed by John McClain of Cornell University and included on the CD-ROM, can be used to solve these problems.

EXHIBIT TN6.8

Properties of Some Specific Waiting Line Models

Model	Layout	Service Phase	Source Population	Arrival Pattern	Queue Discipline	Service Pattern	Permissible Queue Length	Typical Example
1	Single channel	Single	Infinite	Poisson	FCFS	Exponential	Unlimited	Drive-in teller at bank; one-lane toll bridge
2	Single channel	Single	Infinite	Poisson	FCFS	Constant	Unlimited	Roller coaster rides in amusement park
3	Multichannel	Single	Infinite	Poisson	FCFS	Exponential	Unlimited	Parts counter in auto agency
4	Single channel	Single	Finite	Poisson	FCFS	Exponential	Unlimited	Machine breakdown and repair in a factory

Here is a quick preview of our four problems to illustrate each of the four waiting line models in Exhibits TN6.8 and TN6.10. Exhibit TN6.9 defines the notations used in Exhibit TN6.10.

EXHIBIT TN6.9

Notations for Equations (Exhibit TN6.10)

INFINITE QUEUING NOTATION: MODELS 1–3

λ = Arrival rate

μ = Service rate

$\frac{1}{\mu}$ = Average service time

$\frac{1}{\lambda}$ = Average time between arrivals

ρ = Ratio of total arrival rate to service rate for a single server $\left(\frac{\lambda}{\mu}\right)^*$

L_q = Average number waiting in line

L_s = Average number in system (including any being served)

W_q = Average time waiting in line

W_s = Average total time in system (including time to be served)

n = Number of units in the system

S = Number of identical service channels

P_n = Probability of exactly n units in system

P_w = Probability of waiting in line

FINITE QUEUING NOTATION: MODEL 4

D = Probability that an arrival must wait in line

F = Efficiency factor, a measure of the effect of having to wait in line

H = Average number of units being serviced

J = Population source less those in queuing system ($N - n$)

L = Average number of units in line

S = Number of service channels

n = Average number of units in queuing system (including the one being served)

N = Number of units in population source

P_n = Probability of exactly n units in queuing system

T = Average time to perform the service

U = Average time between customer service requirements

W = Average waiting time in line

X = Service factor, or proportion of service time required

* For single-server queues this is equivalent to utilization.

EXHIBIT TN 6.10

Equations for Solving Four Model Problems

Model 1

$$\begin{cases} L_q = \dfrac{\lambda^2}{\lambda(\mu - \lambda)} & W_q = \dfrac{L_q}{\lambda} \qquad P_n = \left(1 - \dfrac{\lambda}{\mu}\right)\left(\dfrac{\lambda}{\mu}\right)^n \qquad P_o = \left(1 - \dfrac{\lambda}{\mu}\right) \\ L_s = \dfrac{\lambda}{\mu - \lambda} & W_s = \dfrac{L_s}{\lambda} \qquad \rho = \dfrac{\lambda}{\mu} \end{cases} \tag{TN6.3}$$

Model 2

$$\begin{cases} L_q = \dfrac{\lambda^2}{2\mu(\mu - \lambda)} & W_q = \dfrac{\lambda}{2\mu(\mu - \lambda)} = \dfrac{L_q}{\lambda} \\ L_s = L_q + \dfrac{\lambda}{\mu} & W_s = W_q + \dfrac{1}{\mu} = \dfrac{L_s}{\lambda} \end{cases} \tag{TN6.4}$$

(Exhibit TN6.11 provides the value of L_q given $\rho = \lambda/\mu$ and the number of servers S.)

Model 3

$$\begin{cases} L_s = L_q + \lambda/\mu & W_s = L_s/\lambda \\ W_q = L_q/\lambda & P_w = L_q\left(\dfrac{S}{\rho} - 1\right) \end{cases} \tag{TN6.5}$$

Model 4 is a finite queuing situation that is most easily solved by using finite tables. These tables, in turn, require the manipulation of specific terms.

Model 4

$$\begin{cases} X = \dfrac{T}{T + U} \qquad H = FNX \qquad L = N(1 - F) \qquad n = L + H \\ P_n = \dfrac{N!}{(N - n)!} X^n P_0 \qquad J = NF(1 - X) \\ W = \dfrac{L(T + U)}{N - L} = \dfrac{LT}{H} \qquad F = \dfrac{T + U}{T + U + W} \end{cases} \tag{TN6.6}$$

Problem 1: Customers in line. A bank wants to know how many customers are waiting for a drive-in teller, how long they have to wait, the utilization of the teller, and what the service rate would have to be so that 95 percent of the time there will not be more than three cars in the system at any time.

Problem 2: Equipment selection. A franchise for Robot Car Wash must decide which equipment to purchase out of a choice of three. Larger units cost more but wash cars faster. To make the decision, costs are related to revenue.

Problem 3: Determining the number of servers. An auto agency parts department must decide how many clerks to employ at the counter. More clerks cost more money, but there is a savings because mechanics wait less time.

Problem 4: Finite population source. Whereas the previous models assume a large population, finite queuing employs a separate set of equations for those cases where the calling customer population is small. In this last problem, mechanics must service four weaving machines to keep them operating. Based on the costs associated with machines being idle and the costs of mechanics to service them, the problem is to decide how many mechanics to use.

EXAMPLE TN6.1: Customers in Line

Western National Bank is considering opening a drive-through window for customer service. Management estimates that customers will arrive at the rate of 15 per hour. The teller who will staff the window can service customers at the rate of one every three minutes.

Part 1 Assuming Poisson arrivals and exponential service, find

1 Utilization of the teller.
2 Average number in the waiting line.
3 Average number in the system.
4 Average waiting time in line.
5 Average waiting time in the system, including service.

SOLUTION—Part 1

1 The average utilization of the teller is (using Model 1)

$$\rho = \frac{\lambda}{\mu} = \frac{15}{20} = 75 \text{ percent}$$

2 The average number in the waiting line is

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{(15)^2}{20(20 - 15)} = 2.25 \text{ customers}$$

3 The average number in the system is

$$L_s = \frac{\lambda}{\mu - \lambda} = \frac{15}{20 - 15} = 3 \text{ customers}$$

4 Average waiting time in line is

$$W_q = \frac{L_q}{\lambda} = \frac{2.25}{15} = 0.15 \text{ hour, or 9 minutes}$$

5 Average waiting time in the system is

$$W_s = \frac{L_s}{\lambda} = \frac{3}{15} = 0.2 \text{ hour, or 12 minutes}$$

Part 2 Because of limited space availability and a desire to provide an acceptable level of service, the bank manager would like to ensure, with 95 percent confidence, that no more than three cars will be in the system at any time. What is the present level of service for the three-car limit? What level of teller use must be attained and what must be the service rate of the teller to ensure the 95 percent level of service?

SOLUTION—Part 2

The present level of service for three or fewer cars is the probability that there are 0, 1, 2, or 3 cars in the system. From Model 1, Exhibit TN6.10.

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

$$\begin{array}{ll}
\text{at } n = 0,\ P_0 = (1 - 15/20) & (15/20)^0 = 0.250 \\
\text{at } n = 1,\ P_1 = (1/4) & (15/20)^1 = 0.188 \\
\text{at } n = 2,\ P_2 = (1/4) & (15/20)^2 = 0.141 \\
\text{at } n = 3,\ P_3 = (1/4) & (15/20)^3 = \underline{0.106} \\
 & \qquad\quad\ 0.685 \quad \text{or} \quad 68.5 \text{ percent}
\end{array}$$

The probability of having more than three cars in the system is 1.0 minus the probability of three or fewer cars ($1.0 - 0.685 = 31.5$ percent).

For a 95 percent service level of three or fewer cars, this states that $P_0 + P_1 + P_2 + P_3 =$ 95 percent.

$$0.95 = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^0 + \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^1 + \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^2 + \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^3$$

$$0.95 = \left(1 - \frac{\lambda}{\mu}\right)\left[1 + \frac{\lambda}{\mu} + \left(\frac{\lambda}{\mu}\right)^2 + \left(\frac{\lambda}{\mu}\right)^3\right]$$

We can solve this by trial and error for values of λ/μ. If $\lambda/\mu = 0.50$,

$$0.95 \overset{?}{=} 0.5(1 + 0.5 + 0.25 + 0.125)$$

$$0.95 \neq 0.9375$$

With $\lambda/\mu = 0.45$,

$$0.95 \overset{?}{=} (1 - 0.45)(1 + 0.45 + 0.203 + 0.091)$$

$$0.95 \neq 0.96$$

With $\lambda/\mu = 0.47$,

$$0.95 \overset{?}{=} (1 - 0.47)(1 + 0.47 + 0.221 + 0.104) = 0.9512$$

$$0.95 \approx 0.95135$$

Therefore, with the utilization $\rho = \lambda/\mu$ of 47 percent, the probability of three or fewer cars in the system is 95 percent.

To find the rate of service required to attain this 95 percent service level, we simply solve the equation $\lambda/\mu = 0.47$, where $\lambda =$ number of arrivals per hour. This gives $\mu = 32$ per hour. That is, the teller must serve approximately 32 people per hour (a 60 percent increase over the original 20-per-hour capability) for 95 percent confidence that not more than three cars will be in the system. Perhaps service may be speeded up by modifying the method of service, adding another teller, or limiting the types of transactions available at the drive-through window. Note that with the condition of 95 percent confidence that three or fewer cars will be in the system, the teller will be idle 53 percent of the time. ●

EXAMPLE TN6.2: Equipment Selection

The Robot Company franchises combination gas and car wash stations throughout the United States. Robot gives a free car wash for a gasoline fill-up or, for a wash alone, charges $0.50. Past experience shows that the number of customers that have car washes following fill-ups is about the same as for a wash alone. The average profit on a gasoline fill-up is about $0.70, and the cost of the car wash to Robot is $0.10. Robot stays open 14 hours per day.

Robot has three power units and drive assemblies, and a franchisee must select the unit preferred. Unit I can wash cars at the rate of one every five minutes and is leased for $12 per day. Unit II, a larger unit, can wash cars at the rate of one every four minutes but costs $16 per day. Unit III, the largest, costs $22 per day and can wash a car in three minutes.

The franchisee estimates that customers will not wait in line more than five minutes for a car wash. A longer time will cause Robot to lose the gasoline sales as well as the car wash sale.

If the estimate of customer arrivals resulting in washes is 10 per hour, which wash unit should be selected?

SOLUTION

Using unit I, calculate the average waiting time of customers in the wash line (μ for unit I = 12 per hour). From the Model 2 equations (Exhibit TN6.10),

$$L_q = \frac{\lambda^2}{2\mu(\mu - \lambda)} = \frac{10^2}{2(12)(12-10)} = 2.08333$$

$$W_q = \frac{L_q}{\lambda} = \frac{2.08333}{10} = 0.208 \text{ hour, or } 12\tfrac{1}{12} \text{ minutes}$$

For unit II at 15 per hour,

$$L_q = \frac{10^2}{2(15)(15-10)} = 0.667$$

$$W_q = \frac{0.667}{10} = 0.0667 \text{ hour, or 4 minutes}$$

If waiting time is the only criterion, unit II should be purchased. But before we make the final decision, we must look at the profit differential between both units.

With unit I, some customers would balk and renege because of the $12\frac{1}{12}$-minute wait. And, although this greatly complicates the mathematical analysis, we can gain some estimate of lost sales with unit I by increasing $W_q = 5$ minutes or $\frac{1}{12}$ hour (the average length of time customers will wait) and solving for λ. This would be the effective arrival rate of customers:

$$W_q = \frac{L_q}{\lambda} = \left(\frac{\lambda^2/2\mu(\mu-\lambda)}{\lambda}\right)$$

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

$$\lambda = \frac{2W_q\mu^2}{1+2W_q\mu} = \frac{2\left(\frac{1}{12}\right)(12)^2}{1+2\left(\frac{1}{12}\right)(12)} = 8 \text{ per hour}$$

Therefore, because the original estimate of λ was 10 per hour, an estimated 2 customers per hour will be lost. Lost profit of 2 customers per hour × 14 hours × $\frac{1}{2}$($0.70 fill-up profit + $0.40 wash profit) = $15.40 per day.

Because the additional cost of unit II over unit I is only $4 per day, the loss of $15.40 profit obviously warrants installing unit II.

The original five-minute maximum wait constraint is satisfied by unit II. Therefore unit III is not considered unless the arrival rate is expected to increase. ●

EXAMPLE TN6.3: Determining the Number of Servers

In the service department of the Glenn-Mark Auto Agency, mechanics requiring parts for auto repair or service present their request forms at the parts department counter. The parts clerk fills a request while the mechanic waits. Mechanics arrive in a random (Poisson) fashion at the rate of 40 per hour, and a clerk can fill requests at the rate of 20 per hour (exponential). If the cost for a parts clerk is \$6 per hour and the cost for a mechanic is \$12 per hour, determine the optimum number of clerks to staff the counter. (Because of the high arrival rate, an infinite source may be assumed.)

SOLUTION

First, assume that three clerks will be used because having only one or two clerks would create infinitely long lines (since $\lambda = 40$ and $\mu = 20$). The equations for Model 3 from Exhibit TN6.10 will be used here. But first we need to obtain the average number in line using the table of Exhibit TN6.11. Using the table and values $\lambda/\mu = 2$ and $S = 3$, we obtain $L_q = 0.8888$ mechanic.

At this point, we see that we have an average of 0.8888 mechanic waiting all day. For an eight-hour day at \$12 per hour, there is a loss of mechanic's time worth 0.8888 mechanic × \$12 per hour × 8 hours = \$85.32.

Our next step is to reobtain the waiting time if we add another parts clerk. We then compare the added cost of the additional employee with the time saved by the mechanics. Again, using the table of Exhibit TN6.11 but with $S = 4$, we obtain

$L_q = 0.1730$ mechanic in line

0.1730 × \$12 × 8 hours = \$16.61 cost of a mechanic waiting in line

Value of mechanics' time saved is \$85.32 − \$16.61	= \$68.71
Cost of an additional parts clerk is 8 hours × \$6/hour	= 48.00
Cost of reduction by adding fourth clerk	= \$20.71

This problem could be expanded to consider the addition of runners to deliver parts to mechanics; the problem then would be to determine the optimal number of runners. This, however, would have to include the added cost of lost time caused by errors in parts receipts. For example, a mechanic would recognize a wrong part at the counter and obtain immediate correction, whereas the parts runner might not. ●

EXAMPLE TN6.4: Finite Population Source

Studies of a bank of four weaving machines at the Loose Knit textile mill have shown that, on average, each machine needs adjusting every hour and that the current servicer averages $7\frac{1}{2}$ minutes per adjustment. Assuming Poisson arrivals, exponential service, and a machine idle time cost of \$40 per hour, determine if a second servicer (who also averages $7\frac{1}{2}$ minutes per adjustment) should be hired at a rate of \$7 per hour.

SOLUTION

This is a finite queuing problem that can be solved by using finite queuing tables. (See Exhibit TN6.12.) The approach in this problem is to compare the cost of machine downtime (either waiting in line or being serviced) and of one repairer, to the cost of machine downtime and two repairers. We do this by finding the average number of machines that are in the service system and multiply this number by the downtime cost per hour. To this we add the repairers' cost.

Before we proceed, we first define some terms:

N = Number of machines in the population

S = Number of repairers

T = Time required to service a machine

U = Average time a machine runs before requiring service

X = Service factor, or proportion of service time required for each machine ($X = T/(T + U)$)

L = Average number of machines waiting in line to be serviced

H = Average number of machines being serviced

EXHIBIT TN6.11

Expected Number of People Waiting in Line (L_q) for Various Values of S and λ/μ

	NUMBER OF SERVICE CHANNELS, S														
λ/μ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0.10	0.0111														
0.15	0.0264	0.0006													
0.20	0.0500	0.0020													
0.25	0.0833	0.0039													
0.30	0.1285	0.0069													
0.35	0.1884	0.0110													
0.40	0.2666	0.0166													
0.45	0.3681	0.0239	0.0019												
0.50	0.5000	0.0333	0.0030												
0.55	0.6722	0.045	0.0043												
0.60	0.9090	0.0593	0.0061												
0.65	1.2071	0.0767	0.0084												
0.70	1.6333	0.0976	0.0112												
0.75	2.2500	0.1227	0.0147												
0.80	3.2000	0.1523	0.0189												
0.85	4.8165	0.1873	0.0239	0.0031											
0.90	8.1000	0.2285	0.0300	0.0041											
0.95	18.0500	0.2767	0.0371	0.0053											
1.0		0.3333	0.0454	0.0067											
1.2		0.6748	0.0940	0.0158											
1.4		1.3449	0.1778	0.0324	0.0059										
1.6		2.8441	0.3128	0.0604	0.0121										
1.8		7.6731	0.5320	0.1051	0.0227	0.0047									
2.0			0.8888	0.1730	0.0390	0.0090									
2.2			1.4907	0.2770	0.066	0.0158									
2.4			2.1261	0.4205	0.1047	0.0266	0.0065								
2.6			4.9322	0.6581	0.1609	0.0425	0.0110								
2.8			12.2724	1.0000	0.2411	0.0659	0.0180								
3.0				1.5282	0.3541	0.0991	0.0282	0.0077							
3.2				2.3855	0.5128	0.1452	0.0427	0.0122							
3.4				3.9060	0.7365	0.2085	0.0631	0.0189							
3.6				7.0893	1.0550	0.2947	0.0912	0.0283	0.0084						
3.8				16.9366	1.5181	0.4114	0.1292	0.0412	0.0127						
4.0					2.2164	0.5694	0.1801	0.0590	0.0189						
4.2					3.3269	0.7837	0.2475	0.0827	0.0273	0.0087					
4.4					5.2675	1.0777	0.3364	0.1142	0.0389	0.0128					
4.6					9.2885	1.4857	0.4532	0.1555	0.0541	0.0184					
4.8					21.6384	2.0708	0.6071	0.2092	0.0742	0.0260					
5.0						2.9375	0.8102	0.2785	0.1006	0.0361	0.0125				
5.2						4.3004	1.0804	0.3680	0.1345	0.0492	0.0175				
5.4						6.6609	1.4441	0.5871	0.1779	0.0663	0.0243	0.0085			
5.6						11.5178	1.9436	0.6313	0.2330	0.0683	0.0330	0.0119			
5.8						26.3726	2.6481	0.8225	0.3032	0.1164	0.0443	0.0164			
6.0							3.6878	1.0707	0.3918	0.1518	0.0590	0.0224			
6.2							5.2979	1.3967	0.5037	0.1964	0.0775	0.0300	0.0113		
6.4							8.0768	1.8040	0.6454	0.2524	0.1008	0.0398	0.0153		
6.6							13.7992	2.4198	0.8247	0.3222	0.1302	0.0523	0.0205		
6.8							31.1270	3.2441	1.0533	0.4090	0.1666	0.0679	0.0271	0.0105	
7.0								4.4471	1.3471	0.5172	0.2119	0.0876	0.0357	0.0141	
7.2								6.3133	1.7288	0.6521	0.2677	0.1119	0.0463	0.0187	
7.4								9.5102	2.2324	0.8202	0.3364	0.1420	0.0595	0.0245	0.0097
7.6								16.0379	2.9113	1.0310	0.4211	0.1789	0.0761	0.0318	0.0129
7.8								35.8956	3.8558	1.2972	0.5250	0.2243	0.0966	0.0410	0.0168
8.0									5.2264	1.6364	0.6530	0.2796	0.1214	0.0522	0.0220
8.2									7.3441	2.0736	0.8109	0.3469	0.1520	0.0663	0.0283
8.4									10.9592	2.6470	1.0060	0.4288	0.1891	0.0834	0.0361
8.6									18.3223	3.4160	1.2484	0.5236	0.2341	0.1043	0.0459
8.8									40.6824	4.4805	1.5524	0.6501	0.2885	0.1208	0.0577
9.0										6.0183	1.9366	0.7980	0.3543	0.1603	0.0723
9.2										8.3869	2.4293	0.9788	0.4333	0.1974	0.0899
9.4										12.4183	3.0732	1.2010	0.5267	0.2419	0.1111
9.6										20.6160	3.9318	1.4752	0.5437	0.2952	0.1367
9.8										45.4769	5.1156	1.8165	0.7827	0.3699	0.16731
10											6.8210	2.2465	0.9506	0.4352	0.2040

EXHIBIT TN6.12

Finite Queuing Tables

POPULATION 4

X	*S*	*D*	*F*	*X*	*S*	*D*	*F*	*X*	*S*	*D*	*F*
.015	1	.045	.999		1	.479	.899	.400	3	.064	.992
.022	1	.066	.998	.180	2	.088	.991		2	.372	.915
.030	1	.090	.997		1	.503	.887		1	.866	.595
.034	1	.102	.996	.190	2	.098	.990	.420	3	.074	.990
.038	1	.114	.995		1	.526	.874		2	.403	.903
.042	1	.126	.994	.200	3	.008	.999		1	.884	.572
.046	1	.137	.993		2	.108	.988	.440	3	.085	.986
.048	1	.143	.992	.200	1	.549	.862		2	.435	.891
.052	1	.155	.991	.210	3	.009	.999		1	.900	.551
.054	1	.161	.990		2	.118	.986	.460	3	.097	.985
.058	1	.173	.989		1	.572	.849		2	.466	.878
.060	1	.179	.988	.220	3	.011	.999		1	.914	.530
.062	1	.184	.987		2	.129	.984	.480	3	.111	.983
.064	1	.190	.986		1	.593	.835		2	.498	.864
.066	1	.196	.985	.230	3	.012	.999	.480	1	.926	.511
.070	2	.014	.999		2	.140	.982	.500	3	.125	.980
	1	.208	.984		1	.614	.822		2	.529	.850
.075	2	.016	.999	.240	3	.014	.999		1	.937	.492
	1	.222	.981		2	.151	.980	.520	3	.141	.976
.080	2	.018	.999		1	.634	.808		2	.561	.835
	1	.237	.978	.250	3	.016	.999		1	.947	.475
.085	2	.021	.999		2	.163	.977	.540	3	.157	.972
	1	.251	.975		1	.654	.794		2	.592	.820
.090	2	.023	.999	.260	3	.018	.998		1	.956	.459
	1	.265	.972		2	.175	.975	.560	3	.176	.968
.095	2	.026	.999		1	.673	.780		2	.623	.805
	1	.280	.969	.270	3	.020	.998		1	.963	.443
.100	2	.028	.999		2	.187	.972	.580	3	.195	.964
	1	.294	.965		1	.691	.766		2	.653	.789
.105	2	.031	.998	.280	3	.022	.99s8		1	.969	.429
	1	.308	.962		2	.200	.968	.600	3	.216	.959
.110	2	.034	.998		1	.708	.752		2	.682	.774
	1	.321	.958	.290	3	.024	.998		1	.975	.415
.115	2	.037	.998		2	.213	.965	.650	3	.275	.944
	1	.335	.954		1	.725	.738		2	.752	.734
.120	2	.041	.997	.300	3	.027	.997		1	.985	.384
	1	.349	.950		2	.226	.962	.700	3	.343	.926
.125	2	.044	.997		1	.741	.724		2	.816	.695
	1	.362	.945	.310	3	.030	.997		1	.991	.357
.130	2	.047	.997		2	.240	.958	.750	3	.422	.905
	1	.376	.941		1	.756	.710		2	.871	.657
.135	2	.051	.996	.320	3	.033	.997		1	.996	.333
	1	.389	.936		2	.254	.954	.800	3	.512	.880
.140	2	.055	.996		1	.771	.696		2	.917	.621
	1	.402	.931	.330	3	.036	.996		1	.998	.312
.145	2	.058	.995		2	.268	.950	.850	3	.614	.852
	1	.415	.926		1	.785	.683		2	.954	.587
.150	2	.062	.995	.340	3	.039	.996		1	.999	.294
	1	.428	.921		2	.282	.945	.900	3	.729	.821
.155	2	.066	.994		1	.798	.670		2	.979	.555
	1	.441	.916	.360	3	.047	.994	.950	3	.857	.786
.160	2	.071	.994		2	.312	.936		2	.995	.526
	1	.454	.910		1	.823	.644				
.165	2	.075	.993	.380	3	.055	.993				
	1	.466	.904		2	.342	.926				
.170	2	.079	.993		1	.846	.619				

The values to be determined from the finite tables are

D = Probability that a machine needing service will have to wait

F = Efficiency factor, which measures the effect of having to wait in line to be serviced

The tables are arranged according to three variables: N, population size; X, service factor; and S, the number of service channels (repairers in this problem). To look up a value, first find the table for the correct N size, then search the first column for the appropriate X, and finally find the line for S. Then read off D and F. (In addition to these values, other characteristics about a finite queuing system can be found by using the finite formulas.)

To solve the problem, consider Case I with one repairer and Case II with two repairers.

Case I: One repairer. From problem statement,

$N = 4$

$S = 1$

$T = 7\frac{1}{2}$ minutes

$U = 60$ minutes

$$X = \frac{T}{T+U} = \frac{7.5}{7.5+60} = 0.111$$

From Exhibit TN6.12, which displays the table for $N = 4$, F is interpolated as being approximately 0.957 at $X = 0.111$ and $S = 1$.

The number of machines waiting in line to be serviced is L, where

$$L = N(1-F) = 4(1-0.957) = 0.172 \text{ machine}$$

The number of machines being serviced is H, where

$$H = FNX = 0.957(4)(0.111) = 0.425 \text{ machine}$$

Exhibit TN6.13 shows the cost resulting from unproductive machine time and the cost of the repairer.

Case II: Two repairers. From Exhibit TN6.12, at $X = 0.111$ and $S = 2$, $F = 0.998$.

The number of machines waiting in line, L, is

$$L = N(1-F) = 4(1-0.998) = 0.008 \text{ machine}$$

The number of machines being serviced, H, is

$$H = FNX = 0.998(4)(0.111) = 0.443 \text{ machine}$$

The costs for the machines being idle and for the two repairers are shown in Exhibit TN6.13. The final column of that exhibit shows that retaining just one repairer is the best choice. ●

EXHIBIT TN6.13

A Comparison of Downtime Costs for Service and Repair of Four Machines

NUMBER OF REPAIRERS	NUMBER OF MACHINES DOWN $(H + L)$	COST PER HOUR FOR MACHINES DOWN $[(H + L) \times \$40/\text{HOUR}]$	COST OF REPAIRERS ($7/HOUR EACH)	TOTAL COST PER HOUR
1	0.597	$23.88	$ 7.00	$30.88
2	0.451	18.04	14.00	32.04

APPROXIMATING CUSTOMER WAITING TIME[2]

● ● ● Good news for managers. All you need is the mean and standard deviation to compute average waiting time! Some good research has led to a "quick and dirty" mathematical approximation to the queuing models illustrated earlier in the technical note. What's nice about the approximation is that it does not assume a particular arrival rate or service distribution. All that is needed is the mean and standard deviation of the interarrival time and the service time. We will not burden you will all the details of how the approximations were derived, just how to use the formulas.

First, you will need to collect some data on your service time. The service time is the amount of time that it takes to serve each customer. Keep in mind that you want to collect your data during a period of time that fairly represents what you expect to happen during the period that you are concerned about. For example, if you want to know how many bank tellers you should have to service customers on Friday around the lunch period, collect your data during that period. This will ensure that the transactions being performed are similar to those that you expect in the future. You can use a stop watch to time how long it takes to serve each customer. Using these data, calculate the mean and standard deviation of the service time.

Recall from your statistics that the mean is

[TN6.7]
$$\overline{X} = \sum_{i=1}^{N} x_i \Big/ N$$

where x_i = observed value and N = total number of observed values.

The standard deviation is

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{X})^2}{N-1}}$$

Next, capture data on the amount of time between the arrivals of each new customer during the period of time you are studying. This is called the interarrival time. From the data, calculate the mean and standard deviation of the interarrival time. From these calculations, we have

$\overline{X}_s$ = Mean service time
$\overline{X}_a$ = Mean interarrival time
S_s = Standard deviation of the service time sample
S_a = Standard deviation of the interarrival time sample

Next, define the following:

$$C_s = \text{Coefficient of variation of service time} = \frac{S_s}{\overline{X}_s}$$

$$C_a = \text{Coefficient of variation of interarrival time} = \frac{S_a}{\overline{X}_a}$$

$$\lambda = \text{Customer arrival rate} = \frac{1}{\overline{X}_a}$$

$$\mu = \text{Customer service rate} = \frac{1}{\overline{X}_s}$$

Now, we can calculate some statistics about our system. First, define S as the number of servers that we intend to use. Then,

$$\rho = \text{Utilization of the servers} = \frac{\lambda}{S\mu}$$

$$L_q = \text{Expected length of the waiting line} = \frac{\rho^{\sqrt{2(S+1)}}}{1-\rho} \times \frac{C_a^2 + C_s^2}{2}$$

$$L_s = \text{Expected number of people in the system} = L_q + S\rho$$

$$W_q = \text{Expected time waiting in line} = \frac{L_q}{\lambda}$$

$$W_s = \text{Expected time in the system} = \frac{L_s}{\lambda}$$

The utilization (ρ) is the percentage of time that the servers are expected to be busy. Often companies that provide high service target this number at between 70 and 80 percent depending on the amount of variance there is in the customer arrival and service rates. L_q is how long the queue is expected to be, and W_q is how long a customer is expected to have to wait in the queue. L_s and W_s are the expected number of customers in the system and the expected time that a customer is in the system. These statistics consider that the total number of customers and the total waiting time must include those that are actually being served.

EXAMPLE TN6.5: Waiting Line Approximation

Let's consider an example of a call center that takes orders for a mail order business. During the peak period, the average time between call arrivals ($\overline{X}_a$) is 0.5 minute with a standard deviation (σ_a) of 0.203 minute. The average time to service a call ($\overline{X}_s$) is 4 minutes and the standard deviation of the service time (σ_s) is 2.5 minutes. If the call center is using 9 operators to service calls, how long would you expect customers to wait before being serviced? What would be the impact of adding an additional operator?

SOLUTION

Excel: Waiting Line Approximation

W_q is the time that we expect a customer to wait before being served. The best way to do these calculations is with a spreadsheet. The spreadsheet "Waiting Line Approximation.xls" on the CD-ROM can be easily used. The following steps are needed for the calculation of the customer wait time.

Step 1. Calculate expected customer arrival rate (λ), service rate per server (μ), coefficient of variation for the interarrival time (C_a) and service time (C_s).

$$\lambda = \frac{1}{\overline{X}_a} = \frac{1}{.5} = 2 \text{ customers per minute}$$

$$\mu = \frac{1}{\overline{X}_s} = \frac{1}{4} = .25 \text{ customer per minute}$$

$$C_a = \frac{S_a}{\overline{X}_a} = \frac{.203}{.5} = .406$$

$$C_s = \frac{S_s}{\overline{X}_s} = \frac{2.5}{4} = .625$$

Step 2. Calculate the expected server utilization (σ).

$$\rho = \frac{\lambda}{S\mu} = \frac{2}{9 \times .25} = .888889 \qquad \text{(Operators are expected to be busy 89 percent of the time.)}$$

Step 3. Calculate the expected number of people waiting (L_q) and the length of the wait (W_q).

$$L_q = \frac{\rho^{\sqrt{2(S+1)}}}{1-\rho} \times \frac{C_a^2 + C_s^2}{2} = \frac{.888889^{\sqrt{2(9+1)}}}{1-.888889} \times \frac{.406^2 + .625^2}{2} = 1.476064 \text{ customers}$$

(This is the number of customers that we expect to be waiting on hold.)

$$W_q = \frac{L_q}{\lambda} = \frac{1.476064}{2} = .738032 \text{ minute}$$

On average, we expect customers to wait 44 seconds (.738032 × 60) before talking to an operator.

For 10 operators, the calculations are as follows:

$$\rho = \frac{\lambda}{S\mu} = \frac{2}{10 \times .25} = .8 \qquad \text{(Operators are expected to be busy 80 percent of the time.)}$$

$$L_q = \frac{\rho^{\sqrt{2(S+1)}}}{1-\rho} \times \frac{C_a^2 + C_s^2}{2} = \frac{.8^{\sqrt{2(10+1)}}}{1-.8} \times \frac{.406^2 + .625^2}{2} = .487579 \text{ customer}$$

$$W_q = \frac{L_q}{\lambda} = \frac{0.487579}{2} = 0.24379 \text{ minute}$$

With 10 operators, the waiting time is cut one-third to 14.6 seconds. If you add two operators (bringing the total to 11), the waiting time in queue is 6.4 seconds. Adding the first additional operator has a significant impact on customer wait time.

This approximation is useful for many typical queuing situations. It is easy to implement using a spreadsheet such as the "Waiting Line Approximation.xls" spreadsheet on the CD-ROM included with the book. Keep in mind that the approximation assumes that the population to be served is large and customers arrive one at a time. The approximation can be useful for a quick analysis of a queuing situation.

COMPUTER SIMULATION OF WAITING LINES

Some waiting line problems that seem simple on first impression turn out to be extremely difficult or impossible to solve. Throughout this supplement we have been treating waiting line situations that are independent; that is, either the entire system consists of a single phase, or else each service that is performed in a series is independent. (This could happen if the output of one service location is allowed to build up in front of the next one so that this, in essence, becomes a calling population for the next service.) When a series of services is performed in sequence where the output rate of one becomes the input rate of the next, we can no longer use the simple formulas. This is also true for any problem where conditions do not meet the requirements of the equations, as specified in Exhibit TN6.9. The technique best suited to solving this type of problem is computer simulation. We treat the topic of modeling and simulation in Technical Note 15.

CONCLUSION

Waiting line problems both challenge and frustrate those who try to solve them. The basic objective is to balance the cost of waiting with the cost of adding more resources. For a service system this means that the utilization of a server may be quite low to provide a short waiting time to the customer. One main concern in dealing with waiting line problems is which procedure or priority rule to use in selecting the next product or customer to be served.

Many queuing problems appear simple until an attempt is made to solve them. This technical note has dealt with the simpler problems. When the situation becomes more complex, when there are multiple phases, or where services are performed only in a particular sequence, computer simulation is necessary to obtain the optimal solution.

Key Terms

Queue A line of waiting persons, jobs, things, or the like.

Queuing system Consists of three major components: (1) the source population and the way customers arrive at the system, (2) the serving systems, and (3) how customers exit the system.

Arrival rate The expected number of customers that arrive each period.

Exponential distribution A probability distribution often associated with interarrival times.

Poisson distribution Probability distribution often used to describe the number of arrivals during a given time period.

Service rate The capacity of a server measured in number of units that can be processed over a given time period.

Formula Review

Exponential distribution

[TN6.1]

$$f(t) = \lambda e^{-\lambda t}$$

Poisson distribution

[TN6.2]

$$P_T(n) = \frac{(\lambda T)^n e^{-\lambda T}}{n!}$$

Model 1 (See Exhibit TN6.7.)

[TN6.3]

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} \qquad W_q = \frac{L_q}{\lambda} \qquad P_n = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n \qquad P_o = \left(1 - \frac{\lambda}{\mu}\right)$$

$$L_s = \frac{\lambda}{\mu - \lambda} \qquad W_s = \frac{L_s}{\lambda} \qquad \rho = \frac{\lambda}{\mu}$$

Model 2

[TN6.4]

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

$$L_s = L_q + \frac{\lambda}{\mu} \qquad W_s = \frac{L_s}{\lambda}$$

Model 3

[TN6.5]

$$L_s = L_q + \lambda/\mu \qquad W_s = L_s/\lambda$$

$$W_q = L_q/\lambda \qquad P_w = L_q\left(\frac{S}{\rho} - 1\right)$$

Model 4

[TN6.6]

$$X = \frac{T}{T + U} \qquad H = FNX \qquad L = N(1 - F) \qquad n = L + H$$

$$P_n = \frac{N!}{(N - n)!} X^n P_0 \qquad J = NF(1 - X)$$

$$W = \frac{L(T + U)}{N - L} = \frac{LT}{H} \qquad F = \frac{T + U}{T + U + W}$$

Waiting time approximation

$$\overline{X} = \sum_{i=1}^{N} x_i \Big/ N \qquad s = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \overline{X})^2}{N-1}}$$

$$C_s = \frac{S_s}{\overline{X}_s} \qquad C_a = \frac{S_a}{\overline{X}_a} \qquad \lambda = \frac{1}{\overline{X}_a} \qquad \mu = \frac{1}{\overline{X}_s}$$

[TN6.7]

$$\rho = \frac{\lambda}{S\mu}$$

$$L_q = \frac{\rho^{\sqrt{2(S+1)}}}{1-\rho} \times \frac{C_a^2 + C_s^2}{2}$$

$$L_s = L_q + S\rho$$

$$W_q = \frac{L_q}{\lambda}$$

$$W_s = \frac{L_s}{\lambda}$$

Solved Problems

SOLVED PROBLEM 1

Quick Lube Inc. operates a fast lube and oil change garage. On a typical day, customers arrive at the rate of three per hour, and lube jobs are performed at an average rate of one every 15 minutes. The mechanics operate as a team on one car at a time.

Assuming Poisson arrivals and exponential service, find

a. Utilization of the lube team.
b. The average number of cars in line.
c. The average time a car waits before it is lubed.
d. The total time it takes to go through the system (that is, waiting in line plus lube time).

Solution

$\lambda = 3, \mu = 4$

a. Utilization $\rho = \frac{\lambda}{\mu} = \frac{3}{4} = 75\%$.

b. $L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{3^2}{4(4-3)} = \frac{9}{4} = 2.25$ cars in line.

c. $W_q = \frac{L_q}{\lambda} = \frac{2.25}{3} = .75$ hour, or 45 minutes.

d. $W_s = \frac{L_s}{\lambda} = \frac{\lambda}{\mu - \lambda} \Big/ \lambda = \frac{3}{4-3} \Big/ 3 = 1$ hour (waiting + lube).

SOLVED PROBLEM 2

American Vending Inc. (AVI) supplies vended food to a large university. Because students often kick the machines out of anger and frustration, management has a constant repair problem. The machines break down on an average of three per hour, and the breakdowns are distributed in a Poisson manner. Downtime costs the company $25/hour per machine, and each maintenance worker gets $4 per hour. One worker can service machines at an average rate of five per hour, distributed exponentially; two workers working together can service seven per hour, distributed exponentially; and a team of three workers can do eight per hour, distributed exponentially.

What is the optimal maintenance crew size for servicing the machines?

Solution

Case I—One worker:

$\lambda = 3$/hour Poisson, $\mu = 5$/hour exponential

There is an average number of machines in the system of

$$L_s = \frac{\lambda}{\mu - \lambda} = \frac{3}{5-3} = \frac{3}{2} = 1\tfrac{1}{2} \text{ machines}$$

Downtime cost is \$25 × 1.5 = \$37.50 per hour; repair cost is \$4.00 per hour; and total cost per hour for 1 worker is \$37.50 + \$4.00 = \$41.50.

$$\begin{array}{rr} \text{Downtime } (1.5 \times \$25) = & \$37.50 \\ \text{Labor } (1 \text{ worker} \times \$4) = & \underline{4.00} \\ & \underline{\underline{\$41.50}} \end{array}$$

Case II—Two workers:

$\lambda = 3, \mu = 7$

$$L_s = \frac{\lambda}{\mu - \lambda} = \frac{3}{7-3} = .75 \text{ machine}$$

$$\begin{array}{lr} \text{Downtime } (.75 \times \$25) & = \$18.75 \\ \text{Labor } (2 \text{ workers} \times \$4.00) = & \underline{8.00} \\ & \underline{\underline{\$26.75}} \end{array}$$

Case III—Three workers:

$\lambda = 3, \mu = 8$

$$L_s = \frac{\lambda}{\mu - \lambda} = \frac{3}{8-3} = \frac{3}{5} = .60 \text{ machine}$$

$$\begin{array}{lr} \text{Downtime } (.60 \times \$25) & = \$15.75 \\ \text{Labor } (3 \text{ workers} \times \$4) = & \underline{12.00} \\ & \underline{\underline{\$27.00}} \end{array}$$

Comparing the costs for one, two, or three workers, we see that Case II with two workers is the optimal decision.

SOLVED PROBLEM 3

American Bank has a single automated teller machine (ATM) located in a shopping center. Data were collected during a period of peak usage on Saturday afternoon and it was found that the average time between customer arrivals is 2.1 minutes with a standard deviation of .8 minute. It also was found it takes an average of 1.9 minutes for a customer to complete a transaction with a standard deviation of 2 minutes. Approximately how long will customers need to wait in line during the peak usage period?

Solution

Step 1. Calculate expected customer arrival rate (λ), service rate per server (μ), coefficient of variation for the arrival distribution (C_a), and service distribution (C_s).

$$\lambda = \frac{1}{\overline{X}_a} = \frac{1}{2.1} = .47619 \text{ customer per minute}$$

$$\mu = \frac{1}{\overline{X}_s} = \frac{1}{1.9} = .526316 \text{ customer per minute}$$

$$C_a = \frac{\sigma_a}{\overline{X}_a} = \frac{.8}{2.1} = .380952$$

$$C_s = \frac{\sigma_s}{\overline{X}_s} = \frac{2}{1.9} = 1.052632$$

Step 2. Calculate the expected server utilization (σ).

$$\rho = \frac{\lambda}{S\mu} = \frac{.47619}{1 \times .526316} = .904762$$

(Operators are expected to be busy 90.5 percent of the time.)

Step 3. Calculate the expected number of people waiting (L_q) and the length of the wait (W_q).

$$L_q = \frac{\rho^{\sqrt{2(S+1)}}}{1-\rho} \times \frac{C_a^2 + C_s^2}{2} = \frac{.904762^{\sqrt{2(1+1)}}}{1 - .904762} \times \frac{.380952^2 + 1.052632^2}{2}$$

$$= 5.385596 \text{ customers}$$

(This is the number of customers that we expect to be waiting on hold.)

$$W_q = \frac{L_q}{\lambda} = \frac{5.385596}{.47619} = 11.30975 \text{ minutes}$$

On average we expect customers to wait 11 minutes and 19 seconds (.30975 × 60) before having access to the ATM.

Review and Discussion Questions

1 Cultural factors affect waiting lines. For example, fast checkout lines (e.g., 10 items or less) are uncommon in Japan. Why do you think this is so?

2 How many waiting lines did you encounter during your last airline flight?

3 Distinguish between a *channel* and a *phase.*

4 What is the major cost trade-off that must be made in managing waiting line situations?

5 Which assumptions are necessary to employ the formulas given for Model 1?

6 In what way might the first-come, first-served rule be unfair to the customer waiting for service in a bank or hospital?

7 Define, in a practical sense, what is meant by an *exponential service time.*

8 Would you expect the exponential distribution to be a good approximation of service times for
a. Buying an airline ticket at the airport?
b. Riding a merry-go-round at a carnival?
c. Checking out of a hotel?
d. Completing a midterm exam in your OM class?

9 Would you expect the Poisson distribution to be a good approximation of
a. Runners crossing the finish line in the Boston Marathon?
b. Arrival times of the students in your OM class?
c. Arrival times of the bus to your stop at school?

Problems

1 Students arrive at the Administrative Services Office at an average of one every 15 minutes, and their requests take on average 10 minutes to be processed. The service counter is staffed by only one clerk, Judy Gumshoes, who works eight hours per day. Assume Poisson arrivals and exponential service times.
a. What percentage of time is Judy idle?
b. How much time, on average, does a student spend waiting in line?
c. How long is the (waiting) line on average?
d. What is the probability that an arriving student (just before entering the Administrative Services Office) will find at least one other student waiting in line?

1. *a.* 33.33%.
b. 20 minutes.
c. 1.33 students.
d. 44.44%.

2 The managers of the Administrative Services Office estimate that the time a student spends waiting in line costs them (due to goodwill loss and so on) $10 per hour. To reduce the time a student spends waiting, they know that they need to improve Judy's processing time (see Problem 1). They are currently considering the following two options:
a. Install a computer system, with which Judy expects to be able to complete a student request 40 percent faster (from 2 minutes per request to 1 minute and 12 seconds, for example).
b. Hire another temporary clerk, who will work at the same rate as Judy.
If the computer costs $99.50 to operate per day, while the temporary clerk gets paid $75 per day, is Judy right to prefer the hired help? Assume Poisson arrivals and exponential service times.

2. Yes.

3. *a.* 33.33%.
   b. 44.44%.
   c. 11.11%.
   d. 1.33 customers.
   e. .20 hour or 12 minutes.

3 Sharp Discounts Wholesale Club has two service desks, one at each entrance of the store. Customers arrive at each service desk at an average of one every six minutes. The service rate at each service desk is four minutes per customer.
a. How often (what percentage of time) is each service desk idle?
b. What is the probability that both service clerks are busy?
c. What is the probability that both service clerks are idle?
d. How many customers, on average, are waiting in line in front of each service desk?
e. How much time does a customer spend at the service desk (waiting plus service time)?

4. *a.* 55%.
   b. 1.1 customers.
   c. .12 hour or 7.2 minutes.
   d. Yes.

4 Sharp Discounts Wholesale Club is considering consolidating its two service desks (see Problem 3) into one location, staffed by two clerks. The clerks will continue to work at the same individual speed of four minutes per customer.
a. What is the probability of waiting in line?
b. How many customers, on average, are waiting in line?
c. How much time does a customer spend at the service desk (waiting plus service time)?
d. Do you think the Sharp Discounts Wholesale Club should consolidate the service desks?

5. $W_s = 4.125$ minutes.
   $L_q = 4.05$ cars.
   $L_s = 4.95$ cars.

5 Burrito King (a new fast-food franchise opening up nationwide) has successfully automated burrito production for its drive-up fast-food establishments. The Burro-Master 9000 requires a constant 45 seconds to produce a batch of burritos. It has been estimated that customers will arrive at the drive-up window according to a Poisson distribution at an average of one every 50 seconds. To help determine the amount of space needed for the line at the drive-up window, Burrito King would like to know the expected average time in the system, the average line length (in cars), and the average number of cars in the system (both in line and at the window).

6. *a.* 3 minutes.
   b. 45 seconds.
   c. Yes, about 36 seconds.

6 The Bijou Theater in Hermosa Beach, California, shows vintage movies. Customers arrive at the theater line at the rate of 100 per hour. The ticket seller averages 30 seconds per customer, which includes placing validation stamps on customers' parking lot receipts and punching their frequent watcher cards. (Because of these added services, many customers don't get in until after the feature has started.)
a. What is the average customer waiting time in the system?
b. What would be the effect on system waiting time of having a second ticket taker doing nothing but validations and card punching, thereby cutting the average service time to 20 seconds?
c. Would system waiting time be less than you found in *b* if a second window was opened with each server doing all three tasks?

7. *a.* 2.08 people.
   b. 2.91 people.
   c. 0.208 hour.
   d. 0.291 hour.
   e. Infinity.

7 To support National Heart Week, the Heart Association plans to install a free blood pressure testing booth in El Con Mall for the week. Previous experience indicates that, on the average, 10 persons per hour request a test. Assume arrivals are Poisson from an infinite population. Blood pressure measurements can be made at a constant time of five minutes each. Assume the queue length can be infinite with FCFS discipline.
a. What average number in line can be expected?
b. What average number of persons can be expected to be in the system?
c. What is the average amount of time that a person can expect to spend in line?
d. On the average, how much time will it take to measure a person's blood pressure, including waiting time?
e. On weekends, the arrival rate can be expected to increase to over 12 per hour. What effect will this have on the number in the waiting line?

8. *a.* 3 people.
   b. 1 minute.
   c. 75%.
   d. .4219.
   Time reduced by 22 seconds: avg. no. of people reduced by 1.125.

8 A cafeteria serving line has a coffee urn from which customers serve themselves. Arrivals at the urn follow a Poisson distribution at the rate of three per minute. In serving themselves, customers take about 15 seconds, exponentially distributed.
a. How many customers would you expect to see on the average at the coffee urn?
b. How long would you expect it to take to get a cup of coffee?
c. What percentage of time is the urn being used?
d. What is the probability that three or more people are in the cafeteria?

If the cafeteria installs an automatic vendor that dispenses a cup of coffee at a constant time of 15 seconds, how does this change your answers to *a* and *b*?

9. *a.* $L = 4(.055) = .22$ waiting.
   b. $W = .466$ hour.
   c. $D = .362$.

9 An engineering firm retains a technical specialist to assist four design engineers working on a project. The help that the specialist gives engineers ranges widely in time consumption. The specialist has some answers available in memory; others require computation, and still others require significant search time. On the average, each request for assistance takes the specialist one hour.

The engineers require help from the specialist on the average of once each day. Because each assistance takes about an hour, each engineer can work for seven hours, on the average, without assistance. One further point: Engineers needing help do not interrupt if the specialist is already involved with another problem.

Treat this as a finite queuing problem and answer the following questions:

a. How many engineers, on average, are waiting for the technical specialist for help?
b. What is the average time that an engineer has to wait for the specialist?
c. What is the probability that an engineer will have to wait in line for the specialist?

10 L. Winston Martin (an allergist in Tucson) has an excellent system for handling his regular patients who come in just for allergy injections. Patients arrive for an injection and fill out a name slip, which is then placed in an open slot that passes into another room staffed by one or two nurses. The specific injections for a patient are prepared, and the patient is called through a speaker system into the room to receive the injection. At certain times during the day, patient load drops and only one nurse is needed to administer the injections.

Let's focus on the simpler case of the two—namely, when there is one nurse. Also assume that patients arrive in a Poisson fashion and the service rate of the nurse is exponentially distributed. During this slower period, patients arrive with an interarrival time of approximately three minutes. It takes the nurse an average of two minutes to prepare the patients' serum and administer the injection.

a. What is the average number you would expect to see in Dr. Martin's facilities?
b. How long would it take for a patient to arrive, get an injection, and leave?
c. What is the probability that there will be three or more patients on the premises?
d. What is the utilization of the nurse?
e. Assume three nurses are available. Each takes an average of two minutes to prepare the patients' serum and administer the injection. What is the average total time of a patient in the system?

10. *a.* 2 people.
b. 6 minutes.
c. .2964.
d. 67%.
e. 0.03375 hour.

11 The Judy Gray Income Tax Service is analyzing its customer service operations during the month prior to the April filing deadline. On the basis of past data it has been estimated that customers arrive according to a Poisson process with an average interarrival time of 12 minutes. The time to complete a return for a customer is exponentially distributed with a mean of 10 minutes. Based on this information, answer the following questions:

a. If you went to Judy, how much time would you allow for getting your return done?
b. On average, how much room should be allowed for the waiting area?
c. If Judy stayed in the office 12 hours per day, how many hours on average, per day, would she be busy?
d. What is the probability that the system is idle?
e. If the arrival rate remained unchanged but the average time in system must be 45 minutes or less, what would need to be changed?

11. *a.* 1 hour.
b. 4.17.
c. 10.
d. .167.
e. $\mu \geq 6.33$

12 A graphics reproduction firm has four units of equipment that are automatic but occasionally become inoperative because of the need for supplies, maintenance, or repair. Each unit requires service roughly twice each hour, or, more precisely, each unit of equipment runs an average of 30 minutes before needing service. Service times vary widely, ranging from a simple service (such as pressing a restart switch or repositioning paper) to more involved equipment disassembly. The average service time, however, is five minutes.

Equipment downtime results in a loss of $20 per hour. The one equipment attendant is paid $6 per hour.

Using finite queuing analysis, answer the following questions:

a. What is the average number of units in line?
b. What is the average number of units still in operation?
c. What is the average number of units being serviced?
d. The firm is considering adding another attendant at the same $6 rate. Should the firm do it?

12. *a.* Units in line = .288 machine.
b. Units in operation = 3.18.
c. Units being serviced = .531.
d. No. Added cost would be $1.42 per hour.

13 Benny the Barber owns a one-chair shop. At barber college, they told Benny that his customers would exhibit a Poisson arrival distribution and that he would provide an exponential service distribution. His market survey data indicate that customers arrive at a rate of two per hour. It will take Benny an average of 20 minutes to give a haircut. Based on these figures, find the following:

a. The average number of customers waiting.
b. The average time a customer waits.
c. The average time a customer is in the shop.
d. The average utilization of Benny's time.

13. *a.* $1\frac{1}{3}$.
b. $\frac{2}{3}$ hour.
c. 1 hour.
d. 67%.

14 Benny the Barber (see Problem 13) is considering the addition of a second chair. Customers would be selected for a haircut on a FCFS basis from those waiting. Benny has assumed that both barbers would take an average of 20 minutes to give a haircut, and that business would remain unchanged with customers arriving at a rate of two per hour. Find the following information to help Benny decide if a second chair should be added:

a. The average number of customers waiting.

14. *a.* .0837.
b. 0.0418 hour or 2.51 minutes.
c. .3751 hour or 22.51 minutes.

b. The average time a customer waits.
c. The average time a customer is in the shop.

15. *a.* 1.5 people; 15 minutes.
b. 60%.
c. 1 − .784 = .216.
d. 0.1099 hour.

15 Customers enter the camera department of a store at the average rate of six per hour. The department is staffed by one employee, who takes an average of six minutes to serve each arrival. Assume this is a simple Poisson arrival exponentially distributed service time situation.
a. As a casual observer, how many people would you expect to see in the camera department (excluding the clerk)? How long would a customer expect to spend in the camera department (total time)?
b. What is the utilization of the clerk?
c. What is the probability that there are more than two people in the camera department (excluding the clerk)?
d. Another clerk has been hired for the camera department who also takes an average of six minutes to serve each arrival. How long would a customer expect to spend in the department now?

16. *a.* 9.167 minutes.
b. 9.091 people, or 10 people.
c. 0.7513.
d. 0.9091; 90.9% of the time.
e. .1099 hour or 4.65 minutes.

16 Cathy Livingston, bartender at the Tucson Racquet Club, can serve drinks at the rate of one every 50 seconds. During a hot evening recently, the bar was particularly busy and every 55 seconds someone was at the bar asking for a drink.
a. Assuming that everyone in the bar drank at the same rate and that Cathy served people on a first-come, first-served basis, how long would you expect to have to wait for a drink?
b. How many people would you expect to be waiting for drinks?
c. What is the probability that three or more people are waiting for drinks?
d. What is the utilization of the bartender (how busy is she)?
e. If the bartender is replaced with an automatic drink dispensing machine, how would this change your answer in part *a*?

17. *a.* 0.833.
b. 5 documents.
c. 0.2 hour.
d. 0.4822.
e. L_q tends to infinity.

17 An office employs several clerks who originate documents and one operator who enters the document information in a word processor. The group originates documents at a rate of 25 per hour. The operator can enter the information with average exponentially distributed time of two minutes. Assume the population is infinite, arrivals are Poisson, and queue length is infinite with FCFS discipline.
a. Calculate the percentage utilization of the operator.
b. Calculate the average number of documents in the system.
c. Calculate the average time in the system.
d. Calculate the probability of four or more documents being in the system.
e. If another clerk were added, the document origination rate would increase to 30 per hour. What would this do to the word processor workload? Show why.

18. *a.* 0.667.
b. 2 students.
c. 0.5 hour.
d. 0.1976.
e. $L_q \to \infty$.

18 A study-aid desk staffed by a graduate student has been established to answer students' questions and help in working problems in your OM course. The desk is staffed eight hours per day. The dean wants to know how the facility is working. Statistics show that students arrive at a rate of four per hour, and the distribution is approximately Poisson. Assistance time averages 10 minutes, distributed exponentially. Assume population and line length can be infinite and queue discipline is FCFS.
a. Calculate the percentage utilization of the graduate student.
b. Calculate the average number of students in the system.
c. Calculate the average time in the system.
d. Calculate the probability of four or more students being in line or being served.
e. Before a test, the arrival of students increases to six per hour on the average. What does this do to the average length of the line?

19. *a.* 4.17 vehicles.
b. $\frac{1}{2}$ minute.
c. 83.3%.
d. 1 − .423 = .5787.

19 At the California border inspection station, vehicles arrive at the rate of 10 per minute in a Poisson distribution. For simplicity in this problem, assume that there is only one lane and one inspector, who can inspect vehicles at the rate of 12 per minute in an exponentially distributed fashion.
a. What is the average length of the waiting line?
b. What is the average time that a vehicle must wait to get through the system?
c. What is the utilization of the inspector?
d. What is the probability that when you arrive there will be three or more vehicles ahead of you?

20. *a.* .175 vehicles.
b. .101 minute or 6.06 seconds.
c. .596.
d. .143 minute or 8.58 seconds.

20 The California border inspection station (see Problem 19) is considering the addition of a second inspector. The vehicles would wait in one lane and then be directed to the first available inspector. Arrival rates would remain the same (10 per minute) and the new inspector would process vehicles at the same rate as the first inspector (12 per minute).
a. What would be the average length of the waiting line?
b. What would be the average time that a vehicle must wait to get through the system?

If a second lane was added (one lane for each inspector):

c. What would be the average length of the waiting line?

d. What would be the average time that a vehicle must wait to get through the system?

21 During the campus Spring Fling, the bumper car amusement attraction has a problem of cars becoming disabled and in need of repair. Repair personnel can be hired at the rate of $20 per hour, but they only work as one team. Thus, if one person is hired, he or she works alone; two or three people work together on the same repair.

One repairer can fix cars in an average time of 30 minutes. Two repairers take 20 minutes, and three take 15 minutes. While these cars are down, lost income is $40 per hour. Cars tend to break down at the rate of two per hour.

How many repairers should be hired?

21. 1 repair person $∞/hr.
2 repair persons $120/hr.
3 repair persons $100/hr.
Use 3 repair persons

22 A toll tunnel has decided to experiment with the use of a debit card for the collection of tolls. Initially, only one lane will be used. Cars are estimated to arrive at this experimental lane at the rate of 750 per hour. It will take exactly four seconds to verify the debit card.

a. In how much time would you expect the customer to wait in line, pay with the debit card, and leave?

b. How many cars would you expect to see in the system?

22. *a.* 14 seconds.
b. 2.92 cars.

23 You are planning a bank. You plan for six tellers. Tellers take 15 minutes per customer with a standard deviation of 7 minutes. Customers arrive one every three minutes according to an exponential distribution (recall that the standard deviation is equal to the mean). Every customer that arrives eventually gets serviced.

a. On average how many customers would be waiting in line?

b. On average how long would a customer spend in the bank?

c. If a customer arrived, saw the line, and decided not to get in line, that customer has ________.

d. A customer who enters the line but decides to leave the line before getting service is said to have ________.

23. *a.* 1.846801
b. 20.540404 minutes.
c. Balked.
d. Reneged.

24 You are planning the new layout for the local branch of the Sixth Ninth Bank. You are considering separate cashier windows for the three different classes of service. Each class of service would be separate with its own cashiers and customers. Oddly enough, each class of service, while different, has exactly the same demand and service times. People for one class of service arrive every four minutes and arrival times are exponentially distributed (the standard deviation is equal to the mean). It takes seven minutes to service each customer and the standard deviation of the service times is three minutes. You assign two cashiers to each type of service.

a. On average, how long will each line be at each of the cashier windows?

b. On average how long will a customer spend in the bank (assume they enter, go directly to one line, and leave as soon as service is complete).

You decide to consolidate all the cashiers so they can handle all types of customers without increasing the service times.

c. What will happen to the amount of time each cashier spends idle? (increase, decrease, stay the same, depends on ______)

d. What will happen to the average amount of time a customer spends in the bank? (increase, decrease, stay the same, depends on ______)

24. *a.* 3.413825 customers.
b. 20.6553 minutes.
c. Stay the same.
d. Decrease.

SELECTED BIBLIOGRAPHY

Davis, M. M., and M. J. Maggard, "An Analysis of Customer Satisfaction with Waiting Times in a Two-Stage Service Process." *Journal of Operations Management* 9, no. 3 (August 1990), pp. 324–34.

Fitzsimmons, J. A., and M. J. Fitzsimmons. *Service Management.* New York: Irwin/McGraw-Hill, 1998, pp. 318–39.

Gross, D., and C. M. Harris. *Fundamentals of Queuing Theory*. New York: Wiley, 1997.

Hillier, F. S., et al. *Queuing Tables and Graphs.* New York: Elsevier–North Holland, 1981.

Katz, K. L.; B. M. Larson; and R. C. Larson. "Prescription for the Waiting-in-Line Blues: Entertain, Enlighten, and Engage." *Sloan Management Review* 32, no. 2 (Winter 1991), pp. 44–53.

Kleinrock, L., and R. Gail. *Queuing Systems: Problems and Solutions.* New York: Wiley, 1996.

Winston, W. L., and S. C. Albright. *Practical Management Science: Spreadsheet Modeling and Application.* New York: Duxbury, 1997, pp. 537–79.

FOOTNOTES

[1] $n!$ is defined as $n(n-1)(n-2)\cdots(2)(1)$.

[2] We are indebted to Gilvan Souza of the Robert H. Smith School of Business, University of Maryland, for his help with this section.