CHAPTER 14

Waiting Time Management

LEARNING OBJECTIVES *The material in this chapter prepares students to:*

- Know why waiting lines are often mismanaged.
- Conceptually understand the nonintuitive nature of waiting lines.
- Find numerical solutions to simple waiting line problems.
- Understand the cost tradeoffs involved in the strategic decision of centralizing versus decentralizing service providers.
- Provide strategies to reduce customers' perceived waiting time or reduce the psychological cost of waiting.
- Additional material on the Student CD prepares students to find numerical solutions to more complex waiting line problems.[1]

Three primary reasons explain why waiting time management is worth studying:

1. The pervasiveness of waiting lines
2. The importance of the problem
3. The lack of managerial intuition surrounding waiting lines

Pervasiveness of Waiting Lines

Waiting lines are ubiquitous in services. Certainly, waiting lines are a common enough problem in services that require a high degree of customer contact. We experience those lines as consumers constantly in restaurants, retail stores, and banks, among many other services. Waiting lines are also an important factor in services without face-to-face contact. A large purchaser of waiting line software is the call center industry. An estimated 4 to 6.5 million people in the United States work in a call center where the largest expense is personnel, at $100 to $200 billion annually. The techniques in this chapter enable you to determine how many people to hire and when to schedule their work time to provide the lowest customer waiting times for the least cost.

[1] The subject of waiting time management poses both qualitative and quantitative problems. Although the chapter contains quantitative content, the focus will be on a qualitative understanding of the problems and some simple math to provide managers with a reasonable approximation of the results of their decisions. More quantitative material is available on the Student CD.

Waiting lines don't occur only in high-contact situations, however. Banks use the information in this chapter to determine how many ATMs to put in place. Back offices of banks, insurance agencies, package delivery firms, payment processing centers, and other services generally do not have customers physically walking in the door. Still, they must complete their work in a timely manner, and the amount of work that comes in daily can be highly variable. Consequently, even back-office services need the material in this chapter.

This chapter is also applicable to the "waiting lines" of e-mails to be answered, phone calls to return, or tasks in an in-box to be done. Regardless of what form the work may come in—face-to-face contact, information on a screen, or paper on the desk—anyone who cannot say to her customers, "I'll get to you when I feel like it," can benefit by knowing the principles behind waiting lines.

Importance of the Problem

Frequently, the amount of waiting time a customer endures is THE customer service standard. Waiting time often comes at the beginning of the service process and can have a "halo" or "pitchfork" effect on how customers view the rest of the service encounter; that is, whether customers view the rest of the service with either a favorable or highly critical eye depends on how they view their up-front wait. Especially in professional services, a customer often cannot tell whether his lawyer, doctor, or dentist did a good job. Results may be evident, but even the best doctors doing their best work may not be able to restore perfect health to a patient. Likewise, great lawyers lose some cases. However, one thing that customers can determine for themselves is whether they have been waiting an unreasonable amount of time, and that aspect can color the customers' perception of the entire service.

This material can be important because of the strong link to personnel requests. Analysis of this kind must be performed to obtain accurate numbers on how many people should be working in a service. As an example in the next few pages will show, if staffing needs are projected based on the total workload, without including waiting line math, a service will be chronically understaffed.

In addition to deciding how many people to hire, the material here presents a strong link to what those people should do. As will be shown, the level of service desired interacts with job descriptions and affects actual job content.

Lack of Managerial Intuition Surrounding Waiting Lines

Lastly, this material is important to study because it is not obvious. It is one of those topics where diligent, intelligent managers who work hard will arrive at drastically wrong answers if they fail to consider this material.

Qualitative Understanding of Waiting Lines

Many people misunderstand why waiting lines form. It is often assumed that waiting lines form only because there's too much work for employees to do in an aggregate sense—for example, giving someone 15 hours of work to do in an eight-hour workday will result in seven hours of the work waiting at the end of the day. However, waiting lines also form when there appear to be more than enough people to handle the tasks in aggregate. This brings us to basic rules of waiting lines.

Rule 1: Waiting Lines Form Even When Total Workload Is Less Than Capacity

If only six hours of work is given to someone working an eight-hour day, waiting lines will still occur, which is an exceedingly important lesson to learn when deciding how many people should staff a service. The reason is variance: Customers do not arrive in uniform time intervals, and the time it takes to serve them is often highly variable.

To put some numbers to this idea, take the *waiting line pop quiz.*

> *Question: How long is the waiting line if a customer arrives exactly every 15 seconds and can be served in exactly 14 seconds?*
>
> *Answer: No waiting line forms at all. This scenario is like an assembly line in a manufacturing plant, where the worker has one second every 15 to relax, put his or her feet up, and read the newspaper.*
>
> *Question: How long is the waiting line if a customer arrives not exactly every 15 seconds, but 15 seconds on average, and can be served in, on average, every 14 seconds (and the arrivals and service times correspond to probability distributions discussed later)?*
>
> *Answer: The length of the waiting line will average 13 people. The number of people in line will bounce up and down, with no one utilizing the system 7% of the time, but 13 will be the average number waiting. (The calculations needed to determine this answer will be shown later.) Even though a customer may arrive on average every 15 seconds, on some occasions five customers may arrive in a 15-second span while at other times no customers may arrive for a few minutes. Further, on those occasions when five customers arrive in a short span, the first customer waited on sometimes takes 90 seconds to satisfy, causing a great deal of waiting for everyone else, even though enough capacity appears to be available, in aggregate.*

Rule 2: Waiting Lines Are Not Linearly Related to Capacity

This second rule of waiting lines defies managerial common sense, because nearly everyone not trained in the mathematics of waiting lines assumes a linear relationship exists between capacity and waiting lines. To visualize this concept, consider the next *waiting line pop quiz:*

> *Question: If one bank teller is working and the average number of customers in line is 12 (point A on Figure 14.1), what would be the average number of customers in line when a second bank teller is hired?*
>
> *Answer: In dozens of presentations to undergraduates, MBA students, and executives, the overwhelming response is that, of course, the line would be cut in half when capacity is doubled. Increase from one to two employees, and the line will drop from 12 to six (point B on Figure 14.1). Unfortunately, this linear "natural" response is dead wrong. As will be demonstrated numerically later, doubling capacity usually causes waiting lines to decrease by more than 90%, and the line will drop from 12 to about one (point C on Figure 14.1). The good news about Figure 14.1 is that adding just a bit of capacity can solve many waiting line situations. The bad news about Figure 14.1 is that this relationship is not as well known in the business world as it should be.*

Top management, however, is generally more interested in profit and loss than waiting time statistics. The relationships in Figure 14.1 drive the ultimate dollar relationships in Figure 14.2. The cost of adding capacity is easy for managers to see and tends to be linear: If one employee is $25/hour, two employees are $50/hour, and so

FIGURE 14.1: *Waiting Line Math*

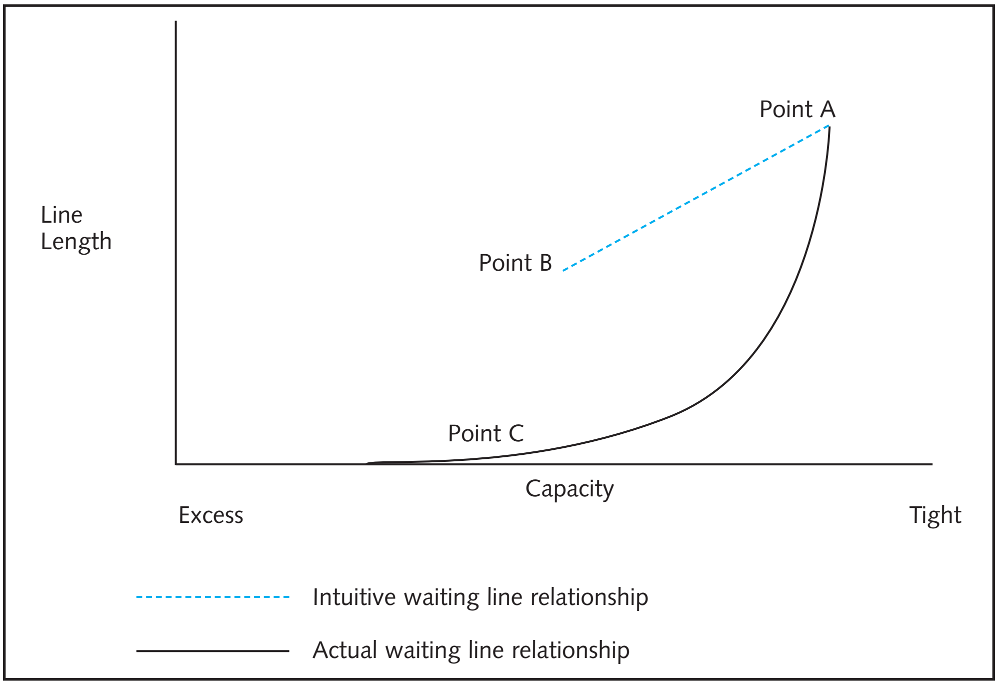
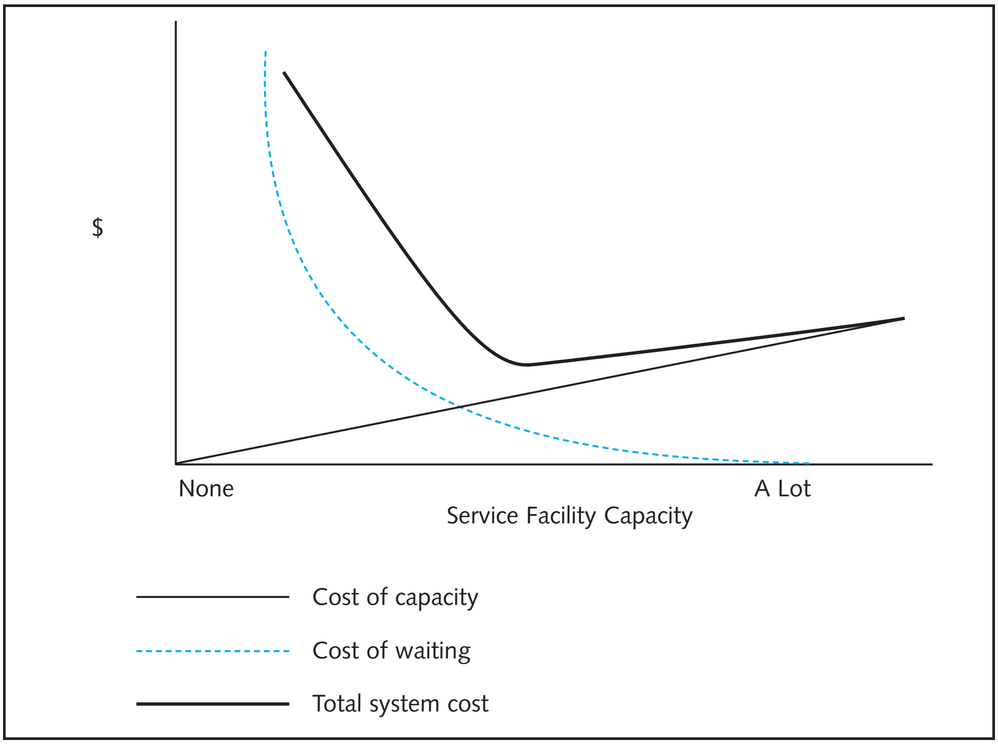


FIGURE 14.2: *Economics of Waiting Lines*



SERVICE OPERATIONS MANAGEMENT PRACTICES

Answering the Phone at L.L.Bean

"I've been on hold forever—think I'll try Lands' End instead."

Most of L.L.Bean's business comes through a telephone call center, and a large portion of that business comes in the concentrated time period six weeks prior to Christmas. If customers get busy signals, or stay on hold too long, many of them will take their business elsewhere. Because of their staffing plans, in some time slots 80% of the incoming calls received a busy signal, and customers who got through were on hold 10 minutes waiting for an agent. Lost sales were approaching $500,000 on some days, and because they were calling on "toll-free" numbers, L.L.Bean was paying $25,000 per day in telephone charges to keep their customers on hold.

Of course, it is not profitable to employ enough staff to answer every call. As the numbers in this chapter show, to do so would mean lots of employee idle time when calls are running low. L.L.Bean implemented a queuing and staffing model based on the economics of lost sales, telephone charges, and salaries to employees. This model shifted their staffing schedules to give them the right number of people at the right time to reach the level of service they were seeking.

The results: The percentage of callers giving up dropped 81% and the average time on hold decreased 83%. L.L.Bean increased profits by $9 million a year due to this study, and the study cost only $40,000.

Source: Condensed from Quinn, Andrews, and Parsons (1991).

on. The cost of waiting can be internal or external: Your own employees waiting in line—while collecting pay—or customers sitting on hold and eventually hanging up. This cost is nonlinear and rises in the same manner as the line length in Figure 14.1. For an example of a firm that went through the process of finding the best balance between the profits lost from customers waiting and the costs of hiring more people, see the Service Operations Management Practices: Answering the Phone at L.L.Bean.

To see how Figures 14.1 and 14.2 can possibly be right, and to demonstrate the two rules of waiting lines, we introduce Example 14.1.

EXAMPLE 14.1: *Teller Staffing at Feehappy Bank and Trust*

As the new manager of the Wilmington branch of Jones B&T, Katrina must decide how many tellers she should staff. The target market for Wilmington is the high net worth customer. Jones charges high fees but promises superior service, so waiting times must be kept short. To make the problem simpler to solve, assume that the branch is open from 8:00 A.M. to 4 P.M. weekdays, no work occurs before opening or after closing, and that the workers refuse to take any lunch breaks. (To see how normal employees may change this analysis, see the case study on teller staffing at the end of the chapter.)

To figure out how many tellers to staff, a study was conducted to determine what amount of work needed to be performed. Table 14.1 shows the various transactions performed by the tellers, the amount of time they take, and what percentage of total transactions this category represents. The expected transaction time, calculated from Table 14.1, is 10(0.05) + 25(0.05) + . . . + 3(0.35) = 5 minutes. Consequently, a teller could be expected to perform an average of 12 transactions per hour.

Table 14.2 shows how many customers enter the branch by time of day. In this table three particular days are surveyed and the average number of customers is in the far right column. We make the simplifying assumptions that every day is the same in terms of incoming customers and that each customer makes one transaction.

In aggregate, Table 14.1 shows that each worker should be able to help 12 customers per hour, and Table 14.2 shows that an average of 180 customers per day come by: This means that there are 180/12 = 15 hours of work to do in an eight-hour workday. If only all the bank customers would drop their work off in the morning and kindly say to the tellers, "Please handle this whenever you can find the time," the work could be inventoried, only two tellers would need to be hired, and each teller would even be able to relax for a half-hour per day.

Unfortunately, as we will show, two tellers would result in quite poor service. Two different sources of variance cause the "two teller" or manufacturing-based solution to be a bad one.

TABLE 14.1: *Work Content at Jones*

Work Content for the Average Customer

Transaction	Average Minutes	Percentage of Transactions
Cashiers' check	10	5%
Open checking account	25	5%
Deposit/cash back	2	25%
Straight deposit	1	10%
Corporate deposit	8	10%
Balance inquiry	1	5%
Dispute	15	5%
Other	3	35%

Average transaction: 5 minutes

Transactions performed in an hour by one teller: 60/ 5 = 12

TABLE 14.2: *Customer Arrivals at Jones*

Time	May 1	May 8	May 15	Average Number of Transactions
8:00-9:00 A.M.	6	12	9	9
9:00-10:00 A.M.	4	11	12	9
10:00-11:00 A.M.	18	24	39	27
11:00-Noon	52	28	28	36
Noon-1:00 P.M.	40	60	35	45
1:00-2:00 P.M.	31	25	25	27
2:00-3:00 P.M.	25	10	19	18
3:00-4:00 P.M.	5	7	15	9
Total:	181	177	182	180

180 transactions × 5 minutes/transaction × 1 hour/60 minutes = 15 hours of work/day

15 hours of work = 1.875 workers

TABLE 14.3: *Variance of Customer Arrivals During the Day*

Workers Handle 12 Transactions per hour

Time	Number of Transactions	Workers Needed
8:00-9:00 A.M.	9	1
9:00-10:00 A.M.	9	1
10:00-11:00 A.M.	27	3
11:00-Noon	36	3
Noon-1:00 P.M.	45	4
1:00-2:00 P.M.	27	3
2:00-3:00 P.M.	18	2
3:00-4:00 P.M.	9	1

Source of Variance 1: Within Day Variance

Table 14.3 shows how many tellers are needed at different times of the day. As with many other retail firms, lighter traffic characteristically occurs early in the morning and toward the end of the day, with heavy traffic at traditional lunch hours. Although only one worker might be able to handle the early customers, the noontime rush requires at least four tellers. Given this typical intraday pattern, the best solution for Feehappy would be to hire only one full-time employee and a host of part-time workers, one of whom works only from noon to 1:00 P.M. every day. Unfortunately, such a plan is not realistic, and intraday variance often requires that more workers be available than are needed for that time slot's average.

Source of Variance 2: Service Time and Customer Arrival Time Variance

Table 14.4 shows an extreme case of service time and customer arrival variance. On an average day, between 11:00 A.M. and noon, 36 transactions would occur, taking five minutes each, requiring 36 × 5/60 = 3 workers. This number of transactions establishes only an average. Consider the day May 1 in Table 14.2: 52 transactions took place on that day and time. What if those transactions included more time commitment transactions, such as more account openings, than average? For Table 14.4, we assume that the average transaction time took seven minutes rather than the usual five. In that case we would need 52 × 7/60 = 6 workers. Therefore, we would need six workers to do the work that could be done by only two workers if we could inventory the work.

However, the crush of customers in aggregate may not be the sole cause of a need for more employees. Table 14.5 gives some specific numbers to the example in

TABLE 14.4: *Variance of Transaction Times and Number of Customers*

Average day, 11:00 A.M. to Noon
36 transactions × 5 minutes/transaction = 180 minutes of work
180 minutes of work = 3 workers

May 1, 11:00 A.M. to Noon
52 transactions: 6 accounts opened, 4 disputes . . .
(higher than average transaction time)
52 transactions × 7 minutes/transaction = 364 minutes of work
364 minutes of work = 6 workers

TABLE 14.5: *A Tale of Two Tellers*

One Teller Scenario

Arrival Time	Transaction	Transaction Time	Waiting Time	Leaves Teller
08:00	Balance inquiry	1	0	8:01
08:04	Deposit/cash back	2	0	8:06
08:08	Open account	25	0	8:33
08:19	Cashier's check	10	14	8:43
08:25	Other	3	18	8:46
08:29	Deposit/cash back	2	17	8:48
08:46	Straight deposit	1	2	8:49
08:52	Other	3	0	8:55
08:54	Other	3	1	8:58
Total		50	52	

Two Teller Scenario

Arrival Time	Transaction	Transaction Time	Waiting Time	Leaves Teller 1	Leaves Teller 2
08:00	Balance inquiry	1	0	8:01	
08:04	Deposit/cash back	2	0		8:06
08:08	Open account	25	0	8:33	
08:19	Cashier's check	10	0		8:29
08:25	Other	3	4		8:32
08:29	Deposit/cash back	2	3		8:34
08:46	Straight deposit	1	0	8:47	
08:52	Other	3	0		8:55
08:54	Other	3	0	8:57	
Total		50	7		

Figure 14.1 (the example of doubling capacity from one to two tellers). The first series in Table 14.5 demonstrates *Rule 1: Waiting lines form even when total workload is less than capacity*. Table 14.5 shows that even though only 50 total minutes of work is being done in an hour, horrendous lines can still form. Consider an average 8:00–9:00 A.M. time period, where the average number of customers show up with average transactions, or nine customers with a little more than five minutes per transaction. One would think that with only 50 minutes of actual work to do in an hour, a single teller should be able to handle the job, but Table 14.5 shows that the "one teller scenario" results in a total of 52 minutes of waiting for customers. The second series on Table 14.5 demonstrates *Rule 2: Waiting lines are not linearly related to capacity.* As promised by Figure 14.1, doubling the number of tellers to two doesn't just cut the waiting time in half, but cuts waiting time from 52 minutes to just 7. However, the trade-off for better customer service means a steep price in productivity. The one teller scenario pays for 60 minutes and gets 50 minutes' worth of work, or a productivity rate of 83%, while the two teller scenario operates at only 42% productivity.

Quantitative Methods: Single Server Model

Although the preceding tables and figures contribute to an intuitive understanding of waiting lines, they do not solve the basic question of Example 14.1: How many tellers should be hired? How often will customers be spaced the way they are in Table 14.5, and how frequently will the customer arrivals be like Table 14.4?

To start, we make some simplifying assumptions about the system we are facing.[2] Given these assumptions, only two basic quantities must be known to calculate how many tellers we need:

$$\lambda \text{ (lambda)} = \text{Arrival rate (example: people per hour)}$$

$$\mu \text{ (mu)} = \text{Service rate (example: people per hour)}$$

Once these basic quantities are calculated, all the basic service information such as the average time in line, average line length, and so on, can be calculated for this system according to the calculation in Table 14.6.

For Jones B&T at 8:00–9:00 A.M., $\lambda = 9$ people per hour, and $\mu = 12$ people per hour, so with a single teller the average time in line would be $\lambda/[\mu(\mu - \lambda)] = 9/36 = {}^1\!/\!_4$ hour or 15 minutes. Adding a second teller requires some mathematics that are a bit more complicated, but consider a similar idea that is roughly equivalent: Hire a teller that is twice as fast as a regular teller. With this "super teller" the average time in line would drop to $9/[24(24 - 9)] = 9/360$ hour, or 1.5 minutes, a 90% decrease from the original solution.

As for the general number of tellers to staff, it depends on the desired service level. If the four-teller equivalent were hired, average wait times would be 18.8 minutes for the lunch rush, which is probably unacceptable for catering to high net worth individuals. The five-teller equivalent results in an average 3-minute wait at lunch. If service is to be really stellar, perhaps the 1½-minute wait of the six-teller equivalent would be

TABLE 14.6: *Basic Waiting Line Model*

Assumptions: 1 server, customer arrivals Poisson distributed, service time exponentially distributed

λ = Arrival rate (example: people per hour)
μ = Service rate (example: people per hour)

$1/\lambda$ = Average time between arrivals (example: minutes per person)
$1/\mu$ = Average service time (example: minutes per person)

Steady state calculations of managerial interest

ρ = Utilization = λ/μ (percentage of time the server is busy)

n_L = Average number in line = $\lambda^2/[\mu(\mu - \lambda)]$
n_S = Average number in the system = $\lambda/(\mu - \lambda)$

t_L = Average time in line = $\lambda/[\mu(\mu - \lambda)]$
t_S = Average time in the system = $1/(\mu - \lambda)$

P_n = Probability of n people in the system = $(1 - \lambda/\mu)(\lambda/\mu)^n$

Service rate necessary given a specific time in line goal:

$$\mu = \frac{\lambda t_L + \sqrt{(\lambda t_L)^2 + 4\lambda t_L}}{2t_L}$$

The formulas above are already programmed in an Excel spreadsheet on the Student CD called "queue.xls" under the "infinite queues" worksheet. This spreadsheet was written by John McClain, Johnson Graduate School of Management, Cornell University.

Access your Student CD now for the Queue.xls spreadsheet containing the formulas.

[2] Technical considerations: The waiting line is from an infinite source of customers with an infinite potential line length, the number of arrivals per unit time is Poisson distributed, there is no balking or reneging, and service times are exponentially distributed. These distributions are used both because they are found to represent many business situations well, and they result in the simple equations found in Table 14.6. The results provided by the formulas used here are "steady rate" results; that is, the results that would accrue if these systems were run at the indicated levels indefinitely.

more appropriate. The important idea here is that using these methods provides managers with information about their potential choices. Table 14.7 assesses the potential choices.

Note that such solutions would result in enormous amounts of idle time throughout the day. Consequently, if high service solutions are desired, it is usually valuable to include in those job descriptions numerous duties that are not time dependent and can be done at the employee's leisure during the inevitable downtime between customers.

Centralizing Waiting Lines: Multiple Servers

An important strategic decision for many services is the level of centralization. For example, should an information systems group within a company be a separate, centralized unit that serves the whole company, or should each company division form its own information systems group? As an example that many consumers can relate to, should an airline operate a few large call centers or dozens of call centers located throughout the country? A separate call center in each major city would reduce telephone charges, but airlines find that with a few centrally located, large centers, often with 1,000 or more employees, the increased telephone charges are more than offset by the personnel cost benefits of centralization.

To see how this strategy might work intuitively, consider the following "social" situation: The "party problem." Consider a party with two bartenders. Should both bartenders be in the same room, perhaps next to each other, or should they be in separate rooms where the line at one cannot be seen by patrons in another? If the goal is to require guests to stand in line as little as possible, then the two bartenders' lines should be within eyesight of each other. If the two bartenders are together, it's not possible for one to have a long line while the other is idle. If they are in separate rooms, however, different line lengths could easily be the case. As is discussed in the additional quantitative material on the Student CD, putting two separated bartenders together can cut waiting in line by 30%.

The "party problem" seems to be somewhat trivial, but the role of queue centralization is a serious business issue. Table 14.8 shows a potential set of choices faced by those who wish to set up a telephone call center system. Consider a system that

Access your Student CD now for Table 14.7 as an Excel worksheet.

TABLE 14.7: *Average Minutes Waiting in Line at Jones B&T*

(single server calculations based on Table 14.6)

	Number of Tellers*						
Time of Day	1	2	3	4	5	6	7
8:00-9:00 A.M.	15.0	1.5	0.6	0.3	0.2	0.1	0.1
9:00-10:00 A.M.	15.0	1.5	0.6	0.3	0.2	0.1	0.1
10:00-11:00 A.M.	**	**	5.0	1.6	0.8	0.5	0.3
11:00-Noon	**	**	**	3.8	1.5	0.8	0.5
Noon-1:00 P.M.	**	**	**	18.8	3.0	1.4	0.8
1:00-2:00 P.M.	**	**	5.0	1.6	0.8	0.5	0.3
2:00-3:00 P.M.	**	7.5	1.7	0.8	0.4	0.3	0.2
3:00-4:00 P.M.	15.0	1.5	0.6	0.3	0.2	0.1	0.1

*Assumes a single teller with the speed of the number of tellers shown.

**"Steady state" averages cannot be reached. Workload greater than capacity. As time continues, waiting lines would continue to increase without end.

TABLE 14.8: *Centralization of Waiting Lines*

Example: Telephone Call Center
Average handle time per call = 3 minutes
Service level desired: Average seconds to answer = 10
Call volume = 4,000 calls per hour

Facilities	Call Volume per Facility	Workload Hours per Facility	Staff Required per Facility*	Total Staff Required
20	200	10	14	280
8	500	25	30	240
4	1,000	50	56	224
2	2,000	100	107	214
1	4,000	200	209	209

*Staffing numbers based on the Erlang-C probability distribution, which are not shown in this text, but is the most common distribution used in practice for determining call center capacity.

receives 4,000 calls per hour with an average handle time of three minutes per call, or 4,000 × 3/60 = 200 hours worth of work to do in an hour. Given a service objective of taking an average of 10 seconds to answer a call, some choices on facilities are given in Table 14.8. At the extremes, one could have one large call center employing 209 people, or one could have 20 smaller centers that each employ 14 people. The mathematics of queue centralization cause the choice between hiring 209 or 14 × 20 = 280 people to provide the same level of service.

Advanced Queuing Models

Quantitative material on advanced queuing modeling techniques can be found on the Student CD. The material includes:

- formulas for multiple channel queues
- adapting the formulas on Table 14.6 to constant service times
- the general system approximation
- priority queues

Access your Student CD now for quantitative material on advanced queuing models.

The Psychology of Queuing

The famous philosopher Berkeley claimed that "perception is essence," which is clearly the case in waiting lines. How long customers wait in line matters far less than how long they *believe* they wait or whether they perceive the wait to be fair or unfair. The perception of waiting times can be drastically different from the waiting times that actually occurred. When people are asked how long they waited for a service, it is not unusual for their answer to be either half as much or twice as much time as actually passed. In one research project, a customer timed at waiting 90 seconds in line claimed to be waiting more than 11 minutes.

Researchers developed several rules concerning the management of the psychology of queues (e.g., Katz, Larson, and Larson, 1991).

- **Perception is more important than reality.**

  Researchers found that the overall opinion of a customer correlates more highly with how long the customer thinks he waited than how long he really has been waiting. Consequently, being attuned to the psychological aspect of waiting lines can be vital.

- **Unoccupied time feels longer than occupied time.**

  Operational Action: Distract and entertain with related or unrelated activity.

  Time waiting with nothing to do feels longer. The hands of the clock appear to move more slowly when a customer is not occupied. In reaction, businesses should attempt to distract the customer.

  One story in operations' lore tells of a large Boston hotel that received complaints about long waiting times for the elevators. Instead of installing more elevators, management changed the wall covering in the elevator lobby to mirrors, presumably to allow the guests to check out their hairdos while waiting. As the story goes, the number of complaints plummeted.

  Many businesses adopt similar strategies. A cottage industry centers around songwriters who focus on small ditties played to customers on telephone hold. Southwest Airlines is legendary for its humorous approach to telephone holds. One time it ran a recording that asked customers a series of questions, only to tell the customer at the end of the session that the survey served no purpose other than to help her pass the time. A number of banks and hospitals attempt to deal with the psychology of waiting by installing televisions or newswires for waiting customers to view, but the results have been mixed, with some efforts actually decreasing customer satisfaction.

- **Preprocess waits feel longer than in-process waits.**

  Operational Action: Communicate as soon as possible and get customers "in process."

  Amusement park waiting lines can be long, hot, and dull. Disney, however, makes customers feel "in process" while waiting by having entertainment to view and listen to in the line. Restaurants can provide menus or drinks to waiting customers to make them feel in process.

  Medical offices also practice this technique. Customers waiting in a large holding area do not know whether they will ever be helped, so the wait seems longer. However, when moved to a different waiting area—an examination room—patients realized that they were in process, so a 20-minute wait in the lounge and a 10-minute wait in the exam room seemed less than a 30-minute wait in the lounge.

- **Uncertain or unexplained waits feel longer than known waits.**

  Operational Action: Communicate frequently.

  Amusement parks are also adept at setting customer expectations as to how long their wait will be, thereby reducing anxiety. Seeing a line for the "Killer Koaster" stretch into the distant horizon is dispiriting not only for the certainty of a long wait, but the uncertainty of whether the wait may be 30 minutes, 90 minutes, or four hours. However, when one sees the sign stating, "Your wait will be 3 days, 9 minutes from this point," one finds some comfort from removal of the uncertainty. Further yet, many such signs deliberately overestimate the time required, so that when the front of the line is reached in only 2 days, the customers are actually happy that they beat expectations.

This psychological aspect of waiting is seen frequently in appointment situations. If a customer arrives at 1:45 for a 2:00 appointment, the first 15 minutes of waiting pass quickly psychologically because the customer is anchored to a 2:00 time frame. However, any unexplained wait after 2:00 is viewed as the work of an international conspiracy aimed at the customer. This is because the customer is no longer mentally anchored to a time. The customer lacks an expectation of when his or her wait will end to replace the original expectation of 2:00. The way to avoid this feeling is to continue to acknowledge the customer, provide a reason for lateness, and anchor the customer to another time that will be met.

Customers also form expectations due to a physical setup. For example, a West Texas bank experienced a rush of activity every Friday as a number of customers received weekly paychecks. To be sensitive to customer needs, the banker physically installed a dozen teller windows—up from the usual three—to be used solely on Fridays to accommodate the rush. Unfortunately, he reported that customer satisfaction declined the rest of the week when customers entering the branch saw a dozen teller windows available, but only three tellers behind them. This situation constituted an unexplained wait for these customers.

- **Unfair waits feel longer**.

  Operational Action: Physically segment different markets.

  Especially within the United States, any line that is not first-come, first-served is viewed as unfair. Unfortunately, to be successful, businesses must use the most-important-customer-first rule instead. This rule can be applied more easily in non-face-to-face situations such as telephone call centers. Software now available can recognize incoming telephone calls from important customers and push those calls to the head of the line. In face-to-face situations, the best method is to physically segment customers so that different customer classes no longer can view the service differential. This practice is prominent in banking, where high-end customers are often served outside the traditional bank branch system.

  Some examples of "unfair"—or not first-come, first-served—systems that are tolerated by customers include the separate check-in lines for first-class airline passengers, the "12 items or less" lines at grocers, and the "commercial accounts" lines at banks. The more businesses can move to these segmented line strategies, the more profitable they can become.

Summary

The waiting line problem is an important one for many services. It often forms the basis of the customer service quality judgment. Further, it is a problem that is inherently nonlinear and, therefore, difficult for managers to understand. Thoughtful, hard-working managers who are responsible for hiring decisions but do not understand this material can find themselves chronically short staffed and not know why.

Waiting line systems are not simply numbers, though. Marketing and operations must jointly attack the problem both through number crunching and by attacking the psychology components of waiting. Appropriately setting customer expectations and responding to unspoken customer needs for reassurance can be just as effective as adding expensive capacity.

Access your Student CD now for information on advanced queuing models.

Many waiting line situations are too complex for either the formulas discussed in this chapter. For those systems, simulation, rather than plugging in formulas, is the best way to discover answers to managerial questions. The simulation of complex systems is discussed in Chapter 9.

Review Questions

1. Why do waiting lines form even when more than enough capacity is available to handle customers?
2. What are the major cost trade-offs for managers to consider in waiting line situations?
3. What is the basic reason that centralizing waiting lines reduces wait times?
4. What are some techniques that can be used to manage customers' perceptions of wait times?

Problems

14.1. Mary Jane Smith, sole employee of The Office of Student Complaints About Operations Courses, has an irate visitor every 20 minutes on average, and the visitor takes an average of 15 minutes to handle. Assuming Poisson arrivals and exponential service times:
   a. What percentage of time is Mary Jane idle?
   b. How long is the line, and how much time do students spend waiting, on average?

14.2. If a server takes precisely 15 seconds to serve a customer and customers arrive exactly every 20 seconds, what is the average waiting line length?
   a. −5.3
   b. 0
   c. 2.25
   d. Infinite
   e. None of the above

14.3. If the customer arrival and service time numbers in problem 14.2 are not exact, but only averages conforming to the exponential distribution, what would be the average line length?

14.4. Due to its size, Incredibly Big Discount Store positions a service desk at both ends of the store, about a mile apart. John staffs the north entrance, where 18 customers arrive per hour. Marsha staffs the south entrance, where 12 customers arrive per hour. Both can serve a customer in three minutes. Assuming Poisson arrivals and exponential service times, how long is the line and how much time do customers spend waiting, on average?

14.5. (Appendix material) John and Marsha from problem 14.4 fell in love and petitioned management to move their service desks next to each other. Aside from the humanitarian benefits, what will be the effect on customer waiting?

14.6. Your summer internship takes you to a Southwest Airlines reservation center. The 1,000 telephone service representatives answer calls concerning flight times, book reservations, handle customer complaints, and so on. The following memo requesting additional service reps is on your desk.

"Worker utilization is getting too low. Our workers now average about 20% idle time. The average waiting time for our customers is now only five seconds. That is, they spend an average of only five seconds on hold before service reps can answer their calls. In keeping with our low cost philosophy, I recommend that we cut the work force by 10%. The increase in customer hold time to 5.5 seconds is a trade-off that I believe is warranted by the cost savings."

Is this analysis appropriate? Why or why not?

14.7. Eric Johnson started the Johnson Grocery Company after years of frustration in not being able to get good, fresh baked goods in Nashville. Johnson Grocery originally specialized in cakes, tarts, breads, and doughnuts sold from Eric's home. His success among the affluent, gourmet food crowd allowed him to expand to a second store within six months and to a half-dozen outlets within the first year. Due to customer requests, he expanded into bagels, ready-made salads, and other manufacturers' products such as gourmet ice cream and gourmet pet food.

Once Eric got beyond a dozen stores, the expense and administrative burden of overseeing all those kitchens caused Eric to maintain a central kitchen/supply location downtown that restocked all the stores. Currently, the distribution center is causing problems in the form of underutilized employees and long lines for truck loading.

The loading dock at the distribution center will accommodate only one truck for loading or unloading at a time. Company-owned trucks arrive according to a Poisson distribution with a mean rate of three trucks per day. At present the company employs a crew of three to load and unload the trucks, and the unloading/loading rate is Poisson distributed with a mean rate of five trucks per day. The company can employ additional (or fewer) persons in the loading crew and increase the average loading rate by one truck per day for each additional employee up to a maximum of six persons who can be utilized effectively in the process (for example, four workers could load/unload six per day, or two workers could load/unload four per day). The company estimates that the cost of an idle truck and driver is $40 per hour and the company pays $12 per hour (including benefits) for each employee in the loading crew.

Eric called in Luke Froeb, foreman of the loading crew, to his office to discuss the problem. Luke contends that they are overstaffed. He dislikes seeing idle workers, and is roundly feared by the loading crew. "We can be more profitable if we fire one of these so-called workers," Luke said. "They are always just sitting around waiting for trucks to come in."

Luke's comments seemed true enough, but the trucking supervisor, Bruce Barry, had earlier told Eric that the truckers' feelings were hurt by the long lines they had to sit in waiting to be unloaded.

a. Advise Eric on the proper number of loading dock personnel to employ to minimize costs under the current system.

b. Comment on his business plan. Also, note any operational changes that could be made to improve the situation.

14.8. The words of Big Boss echo in your head: "I don't want to see any of our customers on hold more than 20 seconds, but don't go crazy with the number of employees either."

One of your responsibilities is a small telephone call center. You receive 1,000 calls per hour, every hour, from 8 A.M. to 5 P.M. If all operators are busy the call is automatically placed on hold. Each call lasts an average of two minutes. Employees go to lunch at either 11:00 or 12:00 for one hour; no other lunch choices and no other breaks are offered. Use the adapted single server queuing models (Table 14.6) to determine how many employees you should employ.

14.9. What is better from a waiting line perspective: One employee who is twice as fast as a normal employee, or two employees?

14.10. After graduating with a degree in operations from Big State U., Joan found her job of a lifetime at Yukon Savings of Nome, Alaska, and she was put in charge of operations. A memo from her assistant, Dave, a classmate of hers at BSU, dealt with waiting lines at ATMs. "Recently, I heard customer complaints of intolerable waiting times at our ATMs. I suggest we study the matter further. I would like your approval to hire 50 data collectors to view waiting lines around the state. This entire process should be completed in three weeks at a cost of less than $100,000."

Joan sadly shook her head and thought, "If only Dave had paid more attention in operations class. We can get the information we need directly from the data we already have."

Each ATM logs a time entry for the beginning and end of each transaction. Analyze the following data for 1:00 P.M. to 3:00 P.M. and determine the extent of the problem.

Transaction Begins	Transaction Ends
1:00	1:01
1:01	1:03
1:03	1:10
1:10	1:14
1:14	1:22
1:22	1:23
1:30	1:31
1:31	1:33
1:33	1:44
1:44	1:47
1:57	2:02
2:06	2:08

Transaction Begins	Transaction Ends
2:08	2:18
2:18	2:21
2:21	2:22
2:34	2:38
2:38	2:39
2:39	2:41
2:49	2:50
2:55	2:58

14.11. (Appendix material) Your current facility employs six servers and has an average line length of 6.66. How long would the line be if a seventh server is added?

14.12. You landed a choice job as operations manager of a new fast-food chain, Continental Cuisine. Continental Cuisine plans to serve a large array of non-traditional fast foods to go, such as veal picata and steak diane, in a bit more time than is required to serve traditional fast foods. The operational key to success is the plan to have a number of mini-kitchens available to your servers.

As operations chief you must choose between four alternative service designs suggested:

- In design #1, customers enter the system, wait in one line for the first server to place an order, then the second server completes the service. At that point the next customer places an order.
- In design #2, separate waiting lines form in front of each server, where each server takes the orders from customers and processes them.
- Design #3 is similar to design #2, except that customers join only one queue.
- In design #4, customers form a single waiting line and are served by a service team consisting of two servers. One server takes the customer's order and then both servers work together to prepare the order. Due to the tasks involved in preparation, work can easily be split in half between the team.

Preliminary studies indicate that arrival and service times are appropriate for the use of queuing formulas. The average ordering time is small, averaging 30 seconds. Preparation time, on the other hand, is long, averaging 12 minutes.

Which service design do you choose and why?

Selected Bibliography

Katz, K., B. Larson, and R. Larson. 1991. Prescription for the Waiting-in-Line Blues: Entertain, Enlighten, and Engage. *Sloan Management Review*, 32(2), 44–53.

Quinn, P., B. Andrews, and H. Parsons. 1991. Allocating Telecommunications Resources at L.L. Bean, Inc. *Interfaces*, 21(1), 75–91.

Randhawa, S., A. Mechling, and R. Joerger. 1989. A Simulation-Based Resource-Planning System for the Oregon Motor Vehicles Division. *Interfaces*, 19(6), 40–51.

CASE STUDY

Queuing Psychology and the Oregon Department of Motor Vehicles[3]

The news was shocking and, as usual, work ground to a halt in the Oregon Department of Motor Vehicles North Eugene branch. A memo with the governor's signature stated that their new mission was, "to serve the Oregon public." This new mission statement immediately replaced their prior mission, which seemed to many as, "to be personally amused by annoying the Oregon public."

It was clear that such a radical change in mission required a change in operational strategy. Currently, the DMV, staffed by five counter operators, one of whom also functioned as the manager, processed approximately 40 different types of transactions, each with separate forms and procedures. The Oregon public could enter one of four different lines, depending on the transaction type they needed.

Lines averaged a steady 10–12 people at each station. Whenever lines exceeded 20, the employee at that window closed the window and spent the next 15 minutes on a coffee break. The coffee pot and break room were positioned at the center of the office so that the DMV employees on break could get a good look at how angry their customers would become. Most amusing of all, 9% of the customers were in the wrong line and had to wait in a second line to complete their transactions. A loudly stifled guffaw from the employees alerted one another to such a customer. Another 18% had incomplete paperwork and were required to return another day. Combined, these groups of customers took virtually no time to wait on and, consequently, the productivity of the office (measured by customers served per employee) was one of the best in the state.

Questions:

1. What queuing psychology rules are being broken at the Oregon DMV and what should be done to fix the problems?
2. Although many of the behaviors in the Oregon DMV case are not true, these particular numbers are true and need attention: 9% of the customers were in the wrong line and had to wait in a second line to complete their transactions, and another 18% had incomplete paperwork and were required to return another day. What can be done to address this situation?

[3] *Source*: Adapted in part from Randhawa, Mechling, and Joerger (1989).

CASE STUDY

Staffing and Scheduling Bank Branch Tellers

Banking hours for your branch are from 9:00 A.M. to 5:00 P.M. Monday through Thursday, and 9:00 A.M. to 7:00 P.M. on Friday. The bank lobby area contains six teller spaces, and a drive-through facility offers three spaces. Due to the bank configuration, tellers cannot handle customers in both the lobby and drive-through.

Data were collected for both customer arrival patterns and service times. The average numbers of customer arrivals for both the drive-through and lobby are in Table 14.9. The average time it takes to serve a customer in both the lobby and drive-through averaged three minutes. An overall service goal for the bank was that customers should not have to wait more than an average of two minutes in line. Use the formulas provided in the text to determine staff requirements during the week.[4]

Translating staff requirements to an actual employee schedule can be a difficult task. Employees cannot be scheduled for more than 40 hours/week. Part-time employees are to be avoided, if possible, due to their higher cost. If part-time employees are used, they must be scheduled in at least four-hour blocks. Tellers are entitled to a paid 15-minute break each day, an unpaid one-hour lunch break, and an additional 30-minute break on Friday for supper. Regulations state that workers cannot work more than five-hour blocks without a meal break. Tellers were scheduled to work 8:30 A.M. to 5:15 P.M. Monday through Thursday and until 7:15 P.M. on Friday due to necessary prep and closing work.

The head teller at each branch takes on extra duties that require an hour per day away from the teller line, but these duties do not require a one-hour block.

Questions:

1. Determine the overall amount of work required in a week. If inventory could be used, how many tellers would be needed?
2. Determine the personnel requirements for each hour of the day using queuing formulas and taking into consideration the service goal.
3. If the purpose is to meet the requirements from the queuing model with the fewest number of staff, fill out Table 14.10 with your solution. Due to the different work hours required on Fridays, separate tables will be needed for at least Monday–Thursday and Friday. Table 14.10 contains room for eight personnel, although it is unlikely that you will need that many. Each employee schedule should contain a designation for the 15-minute break and lunch breaks. The head teller also requires administrative time off.

[4] The one-server formulas in Table 14.6 are not technically correct for this multiple server problem. However, they may still be used and will provide an upper bound on the number of tellers needed. Alternatively, the formulas and tables in the appendix may be used to increase accuracy.

CASE STUDY

4. What would be the savings in personnel of rearranging the branch so that tellers would be able to handle both drive-through and lobby traffic?
5. Consider the effect of absent employees due to two weeks of vacation time per year, and a reasonable number of sick days per employee.

Access your Student CD now for Table 14.9 as an Excel worksheet.

TABLE 14.9: *Average Number of Customers*

	Lobby		Drive-Through	
Time	**Monday–Thursday**	**Friday**	**Monday–Thursday**	**Friday**
9:00-10:00 A.M.	6	15	5	10
10:00-11:00 A.M.	15	21	7	15
11:00-Noon	24	45	15	24
Noon-1:00 P.M.	36	63	19	30
1:00-2:00 P.M.	38	70	19	32
2:00-3:00 P.M.	31	52	8	18
3:00-4:00 P.M.	20	44	6	9
4:00-5:00 P.M.	13	36	9	10
5:00-6:00 P.M.		69		28
6:00-7:00 P.M.		75		25

TABLE 14.10: *Employee Schedule*

	Head Teller	Teller 2	Teller 3	Teller 4	Teller 5	Teller 6	Teller 7	Teller 8
8:30 A.M.								
8:45								
9:00								
9:15								
9:30								
9:45								
10:00								
•								
•								
•								
6:00 P.M.								
6:15								
6:30								
6:45								
7:00								
7:15								