CHAPTER OUTLINE AND LEARNING OBJECTIVES

1. Understand the Relationships Between Services and Tangibles.
   - Identify How Services and Nonservices Differ

2. Identify and Apply the Key Elements of Service Design
   - Describe how managers design for service quality.
   - Explain how managers and designers design products for service recovery.
   - Explain the differences between business-to-consumer (B2C) and business-to-business (B2B) services.
   - Describe customer-interactive services and the components of the customer experience.

3. Understand and Apply the Process Chain Network (PCN) Tool for Service Design
   - Understand process chain networks.
   - Explain process positioning.
   - List the three process principles.
   - List and use the steps in developing a PCN diagram.

4. Describe and Use the Planning Service Capacity for Uncertain Demand
   - Understand the components of capacity planning.
   - Describe the tools that managers use to plan and manage capacity.

5. Applying Queuing Theory
Customer Inputs to Mobile Phone Service

The customer plays a key role in services. Consider mobile phone service. You may not think that you have much influence on your phone service, but you do. You provide your location and your signal to your phone service provider; both are fundamental elements of mobile phone service. You can also change your carrier based on the quality of its service.

Every mobile phone user knows about those dreaded dead spots, locations where your carrier’s signal fails, calls drop, and data slows to a crawl. When you sign up for service, you’ve got to hope that the signal is decent at your primary locations, such as where you live and where you work.

Wouldn’t it be great if there were a way to instantly determine which cell phone carrier is best so that you could better evaluate the service you might receive? In this chapter, we will discuss service design and the role of the customer in design. At the end of the chapter, we will discuss one way to evaluate your cell provider.

**services operations**
Production processes wherein each customer is a supplier of process inputs.

**services**
The result of services operations.

**tangibles**
Products, technology, and other outputs and inputs associated with services.

Dentistry is an example of a service. While they use resources in their work, they cannot complete their job without a patient’s teeth to clean and fix.

Services are everywhere. Airplanes are flying around the world right now. Restaurants are feeding customers, hotels are providing lodging, therapists are counseling people, surgeons are healing, and on and on. The university or college that you are attending is also an example of a service. The mantra for business today is to provide better and better service to satisfy customers in order to increase market share and profitability.

In Chapter 3, we discussed the design of products and processes. In this chapter, we discuss service design. Service design is different from product design in that in a service system, the product is a process. There are many definitions of service. Some of them are as simple as “anything that isn’t manufacturing,” but we adopt a more specific definition. Services operations are production processes wherein each customer is a supplier of process inputs. Services are the result of services operations. Implicit in this definition is the concept that customers are intimately involved in services. Our definition of services refers to what we would more precisely call a service operation, or an operation that processes customer inputs. This definition differs from the often-used (in academia) yet inaccurate assumption that services are intangible products. Services are not only quite tangible, but can involve many tangible elements.
Although this dental example may seem humorous, it illustrates the contrasts between service processes and a typical make-to-stock manufacturing operation. For example, a factory that makes dental tools can produce those tools even at times of low demand, and it can keep the tools in inventory until customers purchase them. The factory can wait for dentists to order the tools, but dentists cannot clean teeth in times of low demand; they can only clean teeth simultaneously with demand.

*Table 4.1* provides examples of customer resources that are processed in service operations. In each case, the provider can prepare for production without customers, but requires customer resources to produce in a substantive, revenue-generating way.

Because production in service operations is dependent on customer resources, the design of services leads to unique challenges. We begin with the elements of service design.

### Managing Across Majors 4.1
Marketing majors, remember that service operations need information from marketing to help identify and understand customer needs.

---

**The Key Elements of Service Design**

An element of service design alluded to in the dentistry example is known as *simultaneity*, which means that the production of services occurs at roughly the same time as customer demand. Instead of producing ahead of demand, service production largely occurs after demand. Restaurants can prepare food products and set tables before customers arrive, but the revenue-generating activity of serving customers can only occur when customers arrive and present their demand (and their input resources).

In manufacturing, inventory is a great resource for dealing with fluctuations in demand. Manufacturing firms can often produce relatively constant quantities of goods across times of high demand and low demand. At times of low demand, manufacturers produce goods and store them in inventory, which can be used during times of high demand.

Because of simultaneity, service production usually cannot be inventoried. Instead, managers must rely on extra capacity to meet fluctuations in demand. If the service provider believes that it is important to meet demand even during times of high demand, it needs to set the capacity to produce to match the high demand levels. During periods of low demand, it may leave much of the capacity idle.

---

### Table 4.1

<table>
<thead>
<tr>
<th>Service Operation</th>
<th>Customer Input Resources</th>
<th>Service Provider Outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accounting</td>
<td>Financial records</td>
<td>Financial statements</td>
</tr>
<tr>
<td>Air transportation</td>
<td>Selves and baggage</td>
<td>Transported passengers</td>
</tr>
<tr>
<td>Auto repair</td>
<td>Broken car</td>
<td>Repaired car</td>
</tr>
<tr>
<td>Consulting</td>
<td>Business problems</td>
<td>Business solutions</td>
</tr>
<tr>
<td>Dentistry</td>
<td>Patient’s teeth</td>
<td>Mended teeth</td>
</tr>
<tr>
<td>Education</td>
<td>Student's mind</td>
<td>Educated mind</td>
</tr>
<tr>
<td>Legal services</td>
<td>Legal problems</td>
<td>Legal remedies</td>
</tr>
<tr>
<td>Healthcare</td>
<td>Illness and injuries</td>
<td>Healthier patients</td>
</tr>
</tbody>
</table>
Chapter 4  Service Design

It is difficult and can be costly to have variable capacity. Because many elements of capacity cannot easily change on the same short-term basis in which demand fluctuates, managers often make capacity decisions over the long term. For example, demand at many restaurants is higher on weekends than on weekdays, but most restaurants cannot add more seating space only for weekends; they have to add space for long-term use. As shown in Table 4.2, the choices relative to service demand depend on the planning horizon.

For example, demand for airline flights fluctuates. Some flights are more popular on weekends than on weekdays, and flights at convenient times are more popular than flights that leave or arrive late at night. Besides passengers and their baggage, one of the primary resources airline companies use is jet airplanes. These jets represent a large fixed cost, a cost that is incurred whether or not passengers are flying on the jet. If an airline has too few jets for times of high demand, it will lose sales. That same airline, however, may also have too many jets for times of low demand and have many idle seats. To adapt to this variable need, some airlines have begun to share planes. Airline companies cannot “save” the high demand seats for planes flying below capacity. The inability to “inventory” unfilled seats to use during times of high demand is an unfortunate condition of service operations. The attribute of service operations known as time-perishable capacity means that unused capacity (at times of low demand) is lost forever and cannot be used to meet later demand.

So, how do airlines deal with time-perishable capacity? One way is to attempt to manage demand by shifting passengers from periods of high demand to periods of low demand by offering lower prices for flights with low demand and charging more for popular flights. This pricing practice encourages people who are flexible with their travel plans to fill planes, tempering problems with time-perishable capacity.

### Designing for Service Quality

Providing service quality can be difficult because of the variation that comes from customer interaction. Customers may not always fulfill their responsibilities in the service process, causing the service to fail. For example, if a customer writes the wrong address on a package, the package delivery service is not going to go as desired.

Service firms should actively seek to identify failure points, places in the service process that are prone to fail. Then, they should identify ways of preventing the failure, or at least reducing the likelihood of its occurrence. One way to do so is by designing [poka yoke](#), which is the Japanese term for failsafing.

For example, FedEx realizes that passengers might give incorrect addresses, so it equips its employees with portable terminals that check the accuracy of postal codes listed on package addresses when mailing. This practice reduces the likelihood that a package will enter the system with an invalid postal code.

### Table 4.2  

<table>
<thead>
<tr>
<th>Planning Horizon</th>
<th>Increasing Capacity</th>
<th>Decreasing Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short term</td>
<td>Work overtime</td>
<td>Underutilize employees</td>
</tr>
<tr>
<td></td>
<td>Hire temporary workers</td>
<td>Assign workers to other areas</td>
</tr>
<tr>
<td></td>
<td>Lease more space</td>
<td>Reduce service offerings</td>
</tr>
<tr>
<td>Long term</td>
<td>Hire more employees</td>
<td>Lay off employees</td>
</tr>
<tr>
<td></td>
<td>Add more space</td>
<td>Sell facilities</td>
</tr>
<tr>
<td></td>
<td>Build new facilities</td>
<td>Close unprofitable sites</td>
</tr>
</tbody>
</table>

*time-perishable capacity* 
Unused capacity (at times of low demand) is lost forever, and cannot be used to meet later demand.

*poka yoke* 
The Japanese term for failsafing.
Service poka yokes can be used in many businesses. Some fast-food restaurants make the opening of their garbage cans slightly smaller than the size of their food trays, preventing customers from throwing away the trays. ATMs require bank customers to retrieve their bank card before giving them cash withdrawals, preventing customers from leaving without their card. Some dry cleaners might require customers to turn the pockets of clothing inside out, thus reducing the likelihood that customers will leave items in the pockets.

**Designing for Service Recovery**

Despite poka yokes and good intentions, service processes sometimes fail to do what they are designed to do. One response to this problem is service recovery, which is an attempt to rectify the problem that the customer experienced with the service. Service recovery is particularly challenging because the failure is not only in the process, but in the way it affects the customer’s attitude about the service provider.

For example, a guest may stay at a hotel that experiences unforeseen mechanical problems, such as an air-conditioning system that breaks in the middle of the night. Moving the customer to another room would only solve part of the problem and may not make up for the customer’s lost sleep. Even providing the guest with a free night’s stay might be of little value if the guest is only in that town for one night or if the room is being paid for by an employer.

Although there is no single formula for effective service recovery, the best firms recognize that the need for service recovery is inevitable. An effective recovery system should specify what types of problems are recovered at what levels of the organization, with simple recoveries handled by the front-line employees and more difficult problems handled by managers. Even more important is assuring that service recovery is coupled with a process to avoid similar service failures in the future.

**B2B versus B2C Services**

The examples we have considered thus far in this chapter—dentistry, restaurants, and airlines—are service operations that directly meet the needs of consumers who are the beneficiaries of the service. They are business-to-consumer, or B2C, operations.

Some service businesses help other businesses meet the needs of their customers. Such arrangements are called business-to-business, or B2B, services. An example of a B2B service is management consulting. Management consultants help client businesses meet the needs of their customers more effectively. For example, information technology consultants help businesses develop information systems that serve clients.

B2B and B2C services are similar in that they process resources that come from customers. Even management consulting requires information about the client’s operational needs before the consulting can proceed. Consulting firms can independently acquire expertise and develop other intellectual resources before serving clients, but the actual consulting service begins when a client engages the consultant to solve specific problems. **Table 4.3** gives examples of B2B and B2C services with varying levels of customer interaction.

**Customer-Interactive Processes**

Service operations require customer resources, which implies that, for all services, there is some degree of interaction between the service provider and the customers. This interaction can enable the service provider to understand each customer’s specific needs and perhaps develop a productive relationship with customers.

No matter how constructive it can be, customer interaction comes with a tremendous cost in terms of process efficiency. Customers can vary dramatically in terms of their ability to navigate a service system, the timing of their resources, their ability or willingness to interact...
Chapter 4
Service Design

This customer variation can disrupt the service process and keep it from achieving high levels of productivity and efficiency. As a result of these disruptions, service design usually needs to be robust enough to handle customer variation in a way that is economically and operationally acceptable.

For example, express delivery services need senders to provide accurate addresses. If senders provide incomplete or inaccurate addresses, the delivery service provider may be unable to meet customer expectations for delivery time. That is one reason FedEx, a leading express delivery service provider, provides customers with standard mailing labels.

Disneyland has a service offering that includes theme-park rides, live entertainment, food services, and souvenir retail. Disney can develop ride equipment, entertainment programs, food items, and other goods as part of the offering, but the service offering is more than that. The service offering also includes designing the processes that customers will go through in receiving the service. People who visit the park are also affected by crowds and lines, and Disney has to focus on making sure that these other elements do not affect their experience. One example of service design is figuring out how customers approach a food-service station given that customers will likely have to wait in line. Disney park planners can design waiting lines, or queues, in many different ways. One food-service station may have three counters, allowing three customers to place orders and be served at the same time. Another queue configuration has a separate queue for each counter, which helps lines stay relatively short but may make it difficult for customers to determine which line to enter. A third queue configuration has one queue for all three counters, which can appear longer but which may eliminate the need for customers to choose a line. Disney has done a number of things to improve waiting times, including issuing fast passes at their parks.

Service design is largely about designing the customer experience as it interacts with the firm operations. Doing so

<table>
<thead>
<tr>
<th>Examples of B2B and B2C Services</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Higher Customer Interaction</strong></td>
</tr>
<tr>
<td>Consulting</td>
</tr>
<tr>
<td>Training</td>
</tr>
<tr>
<td><strong>Lower Customer Interaction</strong></td>
</tr>
<tr>
<td>Market research</td>
</tr>
<tr>
<td>Contract field repair</td>
</tr>
<tr>
<td>Facilities maintenance</td>
</tr>
<tr>
<td><strong>B2B</strong></td>
</tr>
<tr>
<td>Doctors</td>
</tr>
<tr>
<td>Attorneys</td>
</tr>
<tr>
<td><strong>B2C</strong></td>
</tr>
<tr>
<td>Mail catalog sales</td>
</tr>
<tr>
<td>Call centers</td>
</tr>
<tr>
<td>Auto repair</td>
</tr>
</tbody>
</table>

**Table 4.3**

**service offering**
A process in which providers ask customers to participate.

**customer experience**
The result of a service process in terms of customer emotions.

**queue**
A waiting line.

Disney has made an art out of managing queues.
Process Chain Network Tool for Service Design

Process chain network (PCN) diagram is a tool that managers can use when designing service delivery systems. This kind of flowchart evaluates the interactions between service providers and customers.

As we discussed in Chapter 3, flowcharts are graphical diagrams with boxes that represent process steps. The steps are connected by arrows that represent process dependency; an arrow between steps suggests that accomplishing one step is dependent on accomplishing another step.

A PCN diagram categorizes the flowchart steps according to whether or not they involve interaction between entities such as providers and customers. In PCN diagrams, an entity is a service provider, a customer, or a supplier. Entities can also be further up or down the supply chain, such as suppliers’ suppliers or customers’ customers.

Service Design Changes at Rogers

As managers, you need to understand what is the state of the art in designing the customer experience. One such example is Rogers. Rogers Communications, a wireless telecommunications company, has designed a new retail store concept that includes a new design and a new service strategy.

Rogers introduced the new retail concept in Toronto, Canada. The store includes a high-touch approach that overcomes the coldness of telecommunications technology. In the Rogers store, customers receive tailored services that serve their specific needs.

“Service and community are at the heart of our innovative new stores,” said Sian Doyle, vice president of retail at Rogers Communications. “This store opening is part of a larger retail transformation to enhance how we service and sell to our customers. With interactive learning sessions and seasoned tech experts available onsite, consumers can learn more and get the latest technology and services in-store.”

The following are some of the elements of the Rogers retail experience:

- **New modern design**: The store uses an open concept that focuses the customer on Rogers’ innovativeness.
- **Personalized service**: Every customer is met at the entry by a host who assesses the customer’s needs and sets the customer up with an appropriate customer service representative. Account questions and hardware needs can both be handled seamlessly in-store.
- **Latest products and experiences**: Exciting product stations have been created where customers can access new products and participate in product demos.
- **Access to community and education**: At the center of the Rogers store is a community table where customers can be trained in new hardware and software.
- **In-store goodies**: The stores provide charging stations, free wifi, a children’s play area, refreshments, and magazines to read.

According to management, “Our new retail stores answer our customers’ need for innovative products and services that consistently deliver great value, where and when they want them, in a seamless, reliable, and easy to use way.”


can be challenging because different customer segments may have different needs that require different processes. In **GLOBAL CONNECTIONS BOX 4.1**, we travel to Canada to see how one telecommunications company has tailored its retail experience.
**FIGURE 4.1** Pizza Restaurant PCN Diagram

1. The **direct interaction** region includes process steps that involve the pizza restaurant's interacting with another entity, such as a local cheese supplier or a customer.
2. The **surrogate interaction** region includes process steps in which the restaurant is acting on the resources coming from another entity without direct interaction. For example, the restaurant may order napkins online.
3. The **independent processing** region includes process steps that the restaurant performs without interacting with other entities, such as developing recipes and turning on the ovens.

The angled “roof” of Figure 4.1 reminds us that organizations have more control over steps they perform independently than they do over steps that require direct interaction with another entity.

**Process Chain Networks**

In the single-process entity shown in Figure 4.1, the key element of service is the interaction between providers and customers. **FIGURE 4.2**, however, shows how the restaurant process incorporates interaction between itself and a customer. The customer has responsibility for controlling specific process steps, including traveling to the restaurant and waiting to be seated. The provider has responsibility for developing recipes and maintaining supplies. The pizza ordering process involves coproduction, meaning that the provider and the customer...
work together to accomplish the task. Note that both customers and service providers have roles in the process. The customer eats the pizza, the provider prepares the check, and the customer pays. The actor in each step is under the “roof” area that specifies the restaurant’s process domain and the customer’s process domain.

The N of PCN stands for network, which means that operating processes typically involve a network of entities. PCN diagrams allow us to visualize service supply chains, which are interactive and interdependent networks of entities.

The lower part of Figure 4.2 shows labels for specific types of process steps. The surrogate and direct interaction steps constitute the service operations. Service steps considered back office are performed outside of the view of the customer, and those considered front office are in view of the customer. Self-service is when customers use provider resources to perform steps. When customers perform steps independently, we often call it do-it-yourself (DIY).

Managing Across Majors 4.3 Information systems majors, recognize that the interactions between entities are often facilitated by an information system architecture.

Process Positioning

The three processing regions of a PCN diagram each have different operating characteristics. Regions of independent processing are more efficient than regions of direct interaction. Because providers have focused competencies and resources, they have greater economies of
scale than their customers. Customers, however, have a greater ability to customize the goods according to their own specific needs. Designers can position process steps in a PCN diagram to achieve improved operating characteristics.

Figure 4.2 showed the step of assembling the pizza in the provider’s region of surrogate interaction, but it is just one process option. Figure 4.3 shows other process options. Option 1 means that the pizza company assembles the pizza independently from customers, such as in a remotely located factory. Option 2 is the make-to-order pizza that we have already discussed, where the pizza is produced in the kitchen, away from direct customer interaction. Option 3 is an interactive assembly process, similar to what Subway sandwiches does with sandwich assembly, with customers choosing from a set number of ingredients. Option 4 is in the customer’s process domain, with the customer assembling the pizza using the provider’s resources. Option 5 has the customer assembling the pizza using ingredients and utensils that they own.

**Three Process Principles**

The bottom part of Figure 4.3 highlights three important process principles:

*Principle 1: Process inefficiency.* The least efficient steps will tend to be those that involve direct interaction. For example, pizza assembly time in option 3 will take much longer than in option 1.

*Principle 2: Economies of scale.* Specialized providers are most likely to have economies of scale, meaning that the fixed costs of obtaining skills and equipment are spread across more units of production. For example, if the pizza restaurant assembles pizzas independently, it can achieve high use of its production equipment.

*Principle 3: Customization.* The more we allow customers to influence or control the process, the greater their potential for customization. Option 5’s pizzas will be the most varied, because the customers are in control of almost everything, from pizza size and shape to cooking time.

How do we know which process positioning is the best? The answer depends on the competency requirements of the provider in relation to the need requirements of the given customer group (called a customer segment). If customers require a large amount of customization, the process step should be closer to the customer (option 5). If the process requires specialized skills and resources, the process step should be closer to the provider (option 1). If both are required, the process may need to be shared, which may come with a cost of inefficiency.

For example, the making of LED light fixtures requires specialized skills but little customization and therefore is positioned in the manufacturer’s independent processing region. Solving customers’ health problems requires specialized skills of physicians but also tremendous customization, so healthcare tends to involve costly direct interaction. For most people, lawn maintenance requires appropriate competencies but a high amount of customization (each lawn is unique), so customers often buy lawn-mowing equipment and maintain their own lawns (independently from lawn-mower manufacturers).

Steps in Developing a PCN Diagram

The following are basic steps for creating a PCN diagram.

1. Identify a process to analyze. As suggested above, the appropriate unit of analysis is a process or process segment, not a firm. PCN analysis takes place at the process level.
2. Identify the process entities that participate in the given process segment. Usually included are a focal firm and an immediate customer or customer segment. The diagram might also include suppliers, partners, and others involved in the value network.
3. Record the steps that mark the start and end of the chosen process segment. Process segments often start with an identified customer need and end with the fulfillment of that need.
4. Fill in intermediate steps, showing in which process domain and region each step occurs. Included here are steps in the process domains of the focal firm, customers of the focal firm, suppliers of the focal firm, and other entities in the process chain network. As mentioned, the arrows between process steps indicate state dependencies (which may or may not involve product flows).

PCN analysis can also identify the value proposition and elements contributing to that value proposition. The following steps may be included at this stage.

5. Identify the steps where the customer receives benefits (i.e., need-filling value that provides motivation to compensate a focal firm) and where the customer incurs nonmonetary costs (such as inconvenience). We tag customer benefits with ☺ and nonmonetary costs with ☹. These tags identify the process’s value proposition to the customer.
6. Identify the steps where the provider firm(s) incurs costs (tagged −$) or receives monetary compensation (tagged +$). Cost steps may include labor costs, component costs, and facility capacity costs. By depicting cost and compensation steps, we get an idea of the profit effect of that given process segment as currently configured.
7. Consider how moving steps across and between process regions would affect the value proposition, either by increasing customer benefits (☺) or by decreasing provider costs ($).

PCN Diagrams in Action

Problem: You are assigned to develop a PCN diagram to show interactions between a patient, health clinic, insurance company, and pharmacy.

Solution: A PCN diagram is shown in Figure 4.4, where the health clinic is entity 1, the patient is entity 2, the insurance company is entity 3, and the pharmacy is entity 4. Steps are connected across the diagram with circles, which show a connecting letter and the entity number where the connection is found.
As we saw in the airline example, one challenge in service operations is time-perishable capacity. Overcoming this challenge requires the supply chain and operations (SC&O) manager to plan for uncertain demand and to manage cost trade-offs. A fundamental planning issue is **service capacity planning**, or deciding how much capacity to provide.

**Capacity Components**

There are many components of capacity, including physical facilities, equipment, and personnel. The capacity of a fast-food restaurant, for example, is limited by the number of cash registers, number of tables, size of the grill, and number of employees. Capacity decisions are largely about managing the trade-offs associated with costs to the provider and costs to the customer. Capacity costs are usually a direct cost to the provider and are usually passed on to the customer in terms of higher prices. For example, you might have been to a restaurant that did not have enough servers. If management were to hire more servers, the costs of doing business could increase, leading to an increase in costs to the customers. Managers must consider these indirect costs when planning the service experience.

If capacity is below demand at a given instance, customers will need to wait. Customer waiting is a direct cost to the customer. If customers refuse to wait and instead leave, the restaurant loses that business. As such, waiting costs are an indirect cost to the firm.

![An Example of a Healthcare PCN Diagram](image)

**Figure 4.4**

**Planning Service Capacity for Uncertain Demand**


---

**Service Capacity Planning**

The process of determining the productive capability needed in a service firm.
Service demand is variable and uncertain. Think of when you go to a retail store. The need for checkout cashiers can vary greatly throughout the day and week. A mistake such as not properly pricing products can lead to delays in cashier service time that can result in longer lines at the checkout as well as customer frustration. As a result, operational mistakes can use up service capacity because of rework. Therefore, firms must decide how much capacity to provide. Often, employees are somewhat flexible because they can be cross-trained to do different functions. Sometimes, part-time labor can be used, but because demand is uncertain, extra labor might need to be available.

Physical facilities and equipment are somewhat less flexible than other components and represent fixed costs. For example, gas stations must decide how many pumps to install. Because fixed costs dominate the cost structures of most services, using services efficiently becomes key to profitability.

Elements of capacity demand planning are found in the story of the new smaller urban Walmart stores discussed in SC&O Current Events 4.1.

Some other approaches for managing service capacity include (1) identifying the cost of service capacity and the cost of lost sales and (2) identifying the appropriate capacity level to meet demand. The tool that we use for this service capacity management issue is called the newsvendor model. Two other approaches for managing service capacity are coproduction and queuing theory.

Coproduction means that customers participate in the service delivery process, and therefore provide their own labor capacity. A common example is self-service retail checkout. The checkout process in retail involves people operating checkout terminals. Retailers need to decide how many checkout terminals and how many cashiers they need at different time intervals. If, however, the retailer provides self-serve checkouts in which customers operate the terminals themselves, the cashier capacity will exactly match demand because the customers are the cashiers!

Another way is to start with a target customer waiting cost and then calculate the amount of capacity needed to meet that target. The tool that managers use for these calculations is queuing theory. Queuing theory (or waiting line theory) is a method for determining customer service system performance and wait times for customers who must wait in lines. We discuss queuing theory in greater depth later in the chapter.

Capacity Planning Tools

There are some basic tools for capacity planning. We will begin with a forecast of demand that can then be used to identify an appropriate capacity level. The costs of excess capacity
are traded off against the costs of insufficient capacity through newsvendor analysis. Finally, queuing theory will show us the effect of capacity decisions on the amount of customer waiting.

**FORECASTING** The basis for service capacity planning is understanding the nature of demand, which is primarily accomplished through demand forecasting. Demand forecasting helps us predict what future demand might be. Forecasting techniques will be covered in Chapter 8. Demand management approaches can also be used to level demand. They are also discussed in Chapter 8.

We must remember that future demand is uncertain, meaning that the forecast represents a probability distribution. The probability distribution of the demand forecast will be used in planning capacity.

**NEWSVENDOR ANALYSIS** Newsvendor analysis is another way to make capacity decisions under a different set of assumptions. In particular, you may want to trade off the costs of not enough capacity ($C_U$, or the cost of understocking) against the costs of excess capacity ($C_O$, or the cost of overstocking). With newsvendor analysis, we make the following assumptions:

1. Excess capacity costs $= C_O$ dollars per time period.
2. If customers have to wait, the cost is $C_U$ per time period. This cost could be lost goodwill or lost sales.
3. Capacity is perishable. In other words, you cannot use today’s extra capacity to meet tomorrow’s demand.

If these conditions are met, the optimal capacity is the point in the cumulative distribution of demand known as the **critical fractile**, $CF$:

$$CF = \frac{C_U}{C_O + C_U}$$

(4.1)

The newsvendor problem is also called the perishable inventory problem because it is used to calculate the optimal amount of inventory of a perishable item. It is called the newsvendor problem because newspapers are an example of a perishable item. Daily newspapers have a certain value on the day they come out, but little or no value if they are not sold that day.

A newsstand in New York City can purchase newspapers for $0.10 each and sell them for $0.25 each. Daily demand is normally distributed with a mean of 100 and a standard deviation
of 20. Newspapers that are not sold at the end of the day are worthless. So, the cost having too many ($C_U$) is $0.10 per paper (the purchase cost), and the cost of having too few ($C_D$) is $0.15 per paper (the lost profit, which is $0.25 revenue minus $0.10 cost). To find the critical fractile using the newsvendor model use Equation 4.1. So, here the critical fractile is $15/(10 + 15) = 0.60$. The 0.60 point on a cumulative normal distribution (see Appendix Table A-2) is $z = 0.255$. So, the optimal number of newspapers is $100 + (0.255 \times 20) = 105.1$, or 105 newspapers. A more complete problem is given in Solved Problem 4.2.

The Newsvendor Problem in Action

Problem: In service environments, an “inventory” issue is capacity. Service operations have a capacity for meeting customer demand according to how they are designed. If there is not enough capacity, customer demand may not turn into sales. Excess capacity comes with a cost as well.

For example, a restaurant chain is opening a new location in a business district. The question is how many tables to design in the restaurant. The key revenue period is the weekday lunch seating, so the restaurant desires to plan capacity for that demand. Lunchtime demand is forecast to be normally distributed with a mean of 25 parties and a standard deviation of 5 parties. (Assume one party per table.) How many tables should the restaurant have?

A naive view would have 25 tables, but that ignores the asymmetric cost structure. The average party spends $40 for lunch, with ingredient costs being 25 percent of that amount.

Solution: Therefore, the cost of insufficient tables ($C_U$) is $40 - (40 \times 0.25) = $30 per table, which is the average profit contribution per party. Each table takes up 100 square feet of space, and space costs $3 per square foot per month (approximately $0.10 per square foot per day).

Therefore, the cost of an extra table is $100 \times 0.10 = $10 per table per day. This cost suggests that an optimal number of tables would have $CF = 30/(10 + 30) = 0.75$, or 75 percent of the cumulative distribution. Looking up 0.75 in the normal cumulative distribution table (see Appendix A-2) shows that $z = 0.675$. Therefore, the optimal number of tables is $25 + 0.675 \times 5 = 28.375$, or approximately 28 tables. See Figure 4.5 for the formulas and a sample layout for a newsvendor problem.

You can also use Excel to apply newsvendor analysis to Solved Problem 4.2.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Example 4-1 Newsvendor Analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>=C3/(C4+C3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Cost of understocking ($)</td>
<td>30</td>
<td>Critical Fractile:</td>
<td>0.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Cost of overstocking ($)</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Average demand (tables):</td>
<td>25</td>
<td>Optimal Number of Tables:</td>
<td>28.375</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Std dev of demand (tables):</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>z value:</td>
<td>0.675</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>(z value is from the table in the appendix, or use =norm.s.inv(B3) command)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

FIGURE 4.5
Newsvendor Excel Spreadsheet
WAITING LINES  Queuing theory shows the relationship between capacity levels and expected wait costs. A queue is a waiting line. Queues can take many forms, such as fast-food drive-through lanes, will-call windows at concerts, and process production lines with variation in processing times. Queuing theory requires that we have information about the distribution of demand (which we get from forecasting) and about the behavior of the queue system. The remainder of this chapter discusses details of queues and queuing theory.

QUEUING THEORY

Managers and designers use queuing theory when planning service capacity if they are uncertain about demand. This section will show you how managers approach queuing theory with respect to psychology, queue configuration, and wait times.

Queuing Psychology

Queuing psychology recognizes that the customer’s cost of waiting is not just about the time customers spend waiting in line, but includes what customers think about the waiting. Uncomfortable waits can seem much longer than they really are. Entertaining waits can seem much shorter. Queuing psychology identifies ways that service operations managers can improve waits by improving the perception of those waits. David Maister provided the following eight points of queuing psychology:

1. Unoccupied waits seem longer than occupied waits.
2. Preprocess waits seem longer than in-process waits.
3. Anxiety makes waits seem longer.
4. Uncertain waits seem longer than waits of a known duration.
5. Unexplained waits seem longer than explained waits.
6. Unfair waits seem longer than equitable waits.
7. The more valuable the service, the longer people will be willing to wait.
8. Waiting alone seems longer than waiting with a group.

Queue Systems and Service Stations

There are many options for queue systems. A system may have one queue or multiple queues. Each queue leads up to one or more servers or service stations. For example, at a fast-food restaurant, customers may wait to place their order and then wait again to pay for it. When there are multiple queues and service stations, one after another, we say that it is a multiphase queue system. The fast-food restaurant example is a two-phase system. The service system includes the waiting line(s) and the service station(s).

There may be one station at each phase, or there may be multiple stations in parallel. Some fast-food restaurants have three or more order-processing stations. The queue configuration tells how the queue is organized for a given phase of a system. If a service system has one server, there will probably be one queue. If there are multiple servers, there could still be one queue leading to the next available server, or each station could have its own queue. Sometimes, multiple stations each have their own queue. Customers entering the system need to decide which queue to enter. In some cases, customers may accidentally choose a queue that moves more slowly than other queues. One way to prevent this selection dilemma is to provide a single queue that is shared by all stations, or what is sometimes called a “snake.” A single queue helps overcome the customers’ dilemma of selecting a queue, but snake queues can appear longer.

Some situations use different queues for different classes of customers. For example, at an airline check-in counter, first-class passengers may have a different queue than coach-class customers.

Even with a single queue, there are different ways to determine which customer is next to be served. The queue discipline is the method for determining who goes next. One common option is first-come, first-served. Another option is customers with the most urgent needs going first, as often happens in a hospital emergency room.

**Wait Times**

How much time customers will spend in a queue can be estimated by queuing theory equations, which are provided in Table 4.4. These equations are used for three different types of queuing problems that will be introduced sequentially. The equations for simple queues are quite simple, but the equations for more complex queues are beyond the scope of this book.

**SINGLE-SERVER QUEUE SYSTEM WITH EXPONENTIAL SERVICE TIMES** We have already talked about single-server systems. They are simply queues with a single line and a single server. Exponential service times are those that are distributed with a long tail. Think about your experiences in lines at grocery stores. Most customers are fairly quick, but some take much more time because they may have items with price checks or they are very slow. A simple single-server queue is based on the following assumptions:

- A single server
- A single queue (only one waiting line)
- Random arrival of customers (Poisson arrival process with mean of \( \lambda \))
- Exponential service times with mean (or average) of \( 1/\mu \)
- First-come, first-served
- No balking (customers will enter the queue regardless of the queue length)
- No reneging (once you enter the system, you cannot leave)
- An infinite calling population that may enter the queue
- No limits on line length

**SOLVED PROBLEM 4.3** shows how to apply these formulas for a single-server/single-queue system.
### Queuing Models and Variables

<table>
<thead>
<tr>
<th>Model</th>
<th>Service Times</th>
<th>Average Line Length (Length of the Queue)</th>
<th>Average Time Spent in Line (Waiting in the Queue)</th>
<th>Average Utilization of the Server</th>
<th>Average Number of Items in System</th>
<th>Average Time Spent in the System</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model I</strong></td>
<td>Single phase/exponential service times</td>
<td>[ L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} ]</td>
<td>[ W_q = \frac{L_q}{\lambda} = \frac{\lambda}{\mu(\mu - \lambda)} ]</td>
<td>[ \rho = \frac{\lambda}{\mu} ]</td>
<td>[ L_s = L_q + \frac{\lambda}{\mu} ]</td>
<td>[ W_s = \frac{L_s}{\lambda} = W_q + \frac{1}{\mu} = \frac{1}{\mu - \lambda} ]</td>
</tr>
<tr>
<td><strong>Model II</strong></td>
<td>Single phase/constant service times</td>
<td>[ L_q = \frac{\lambda^2}{2\mu(\mu - \lambda)} ]</td>
<td>[ W_q = \frac{L_q}{\lambda} = \frac{\lambda}{2\mu(\mu - \lambda)} ]</td>
<td>[ \rho = \frac{\lambda}{\mu} ]</td>
<td>[ L_s = L_q + \frac{\lambda}{\mu} ]</td>
<td>[ W_s = \frac{L_s}{\lambda} = ]</td>
</tr>
<tr>
<td><strong>Model III</strong></td>
<td>Multiphase/exponential service times</td>
<td>[ L_q = \frac{\lambda \mu(\lambda/\mu)^M}{(M - 1)!}\left(\frac{\lambda}{M\mu}\right)^{M-1} P_0 ]</td>
<td>[ W_q = \frac{L_q}{\lambda} ]</td>
<td>[ \rho = \frac{\lambda}{M\mu} ]</td>
<td>[ L_s = L_q + \frac{\lambda}{\mu} ]</td>
<td>[ W_s = \frac{L_s}{\lambda} ]</td>
</tr>
</tbody>
</table>

- \( \lambda \) = arrival rate
- \( \mu \) = service rate
- \( \rho = \frac{\lambda}{\mu} \) = ratio of arrival rate to service rate
- \( P_0 \) = probability of 0 customers in the system
- \( \frac{1}{\mu} \) = average service time
- \( \frac{1}{\lambda} \) = average time between arrivals
- \( M \) = number of servers
Model I in Action for a Single-Phase Queue with a Single Server and Exponential Service Times

Problem: The Curtis Fast Food Service is considering installing a drive-through lane (or channel) for its customers, which consists of one lane and one serving window. It is expected that 20 customers will arrive every hour. There is an average 2-minute service time that is exponentially distributed so that an average of 30 customers can be served per hour. Find the following:

a. average line length
b. average time spent in line
c. average server utilization
d. average number of customers in the system
e. average time spent in the system

Solution: Using the model I formulas, we compute the following:

a. average line length (length of the queue) = \( L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{20^2}{30(30 - 20)} = 1.33 \) customers
b. average time spent in line (waiting in the queue) = \( W_q = \frac{L_q}{\lambda} = \frac{1.33}{20} = 0.067 \) hour = 3.99 minutes
c. average utilization of the server = \( \frac{\lambda}{\mu} = \frac{20}{30} = 0.666 \)
d. average number of customers in system = \( L_s = L_q + \frac{\lambda}{\mu} = 1.33 + 0.666 = 2 \) customers
e. average time spent in the system = \( W_s = \frac{L_s}{\lambda} = \frac{2}{20} = 0.10 \) hour = 6 minutes

SINGLE-SERVER QUEUE SYSTEMS WITH CONSTANT SERVICE TIMES The assumptions for the single-server queue system with constant service times are the same as above except that service times are always the same. Consider a drive-through automated car wash where the time for the wash is constant from one car to the next. To compute the necessary statistics, we will use model II in Table 4.4. You will notice that the length of the queue is cut in half. We demonstrate this in SOLVED PROBLEM 4.4.

Model II in Action for a Single-Phase Queue with a Single Server and Constant Service Times

Problem: The RoboWorld Car Wash services cars in a constant time of 5 minutes. During this time, the cars are washed, waxed, and dried. The service time is constant, with a throughput of 12 per hour. On average, 8 customers arrive per hour. Compute the following:

a. Average line length
b. Average time spent in line
c. Average server utilization
d. Average number of customers in the system
e. Average time spent in the system

Solution:

a. average line length (length of the queue) = \( L_q = \frac{\lambda^2}{2\mu(\mu - \lambda)} = \frac{8^2}{2(12)(12 - 8)} = 0.67 \) customer
b. average time spent in line (waiting in the queue) = \( W_q = \frac{L_q}{\lambda} = \frac{0.67}{8} = 0.0833 \text{ hour} = 5 \text{ minutes} \)

c. average utilization of the server = \( \frac{\lambda}{\mu} = \frac{8}{12} = 0.67 \)

d. average number of customers in system = \( L_s = L_q + \frac{\lambda}{\mu} = 0.67 + 0.67 = 1.33 \) customers

e. average time spent in the system = \( W_s = \frac{L_s}{\lambda} = \frac{1.33}{8} = 0.166 \text{ hour} = 10 \text{ minutes} \)

MULTISERVER QUEUE SYSTEM WITH EXPONENTIAL SERVICE TIMES What happens when we consider using two servers instead of just one server? To analyze this scenario, we will use model III in Table 4.4. The equations for this model are quite complex, so \( L_q \) values are provided in Table 4.5. To use Table 4.5, you need to calculate \( \frac{\lambda}{\mu} \), which specifies a particular row of the table (round as necessary). The columns of the table are for different numbers of servers \( (M) \). Note that we again assume exponential service times. Use of this model is demonstrated in SOLVED PROBLEM 4.5.

SOLVED PROBLEM 4.5 > MyOMLab Video

Problem: A bank is trying to figure out how many tellers to staff during the lunch time from 11:30 a.m. to 1:30 p.m. During that time, customers arrive at an average rate of 18 per hour, distributed according to a Poisson distribution. On average, a customer spends 3 minutes with a teller to complete a transaction. Assume that all assumptions of model III are met. The bank manager does not want the average customer wait time to be more than 3 minutes.

a. What is the average customer wait time with one teller?
b. What would the average customer wait time be if the bank had two tellers?
c. How long would customers wait on average with three tellers?

Solution:
\( \lambda = 18 \) customers per hour (arrival rate)
\( \frac{1}{\mu} = 3 \) minutes per customer (service time)

So,\( \mu = 20 \) customers per hour (service rate)
\( \frac{\lambda}{\mu} = \frac{18}{20} = 0.9 \)

a. For \( M = 1 \), \( L_q = 8.100 \) customers (from Table 4.5), so \( W_q = L_q/\lambda = 8.1/18 = 0.45 \) hour, which is 27 minutes waiting on average \((0.45 \times 60 = 27 \text{ minutes})\).
b. For \( M = 2 \), \( L_q = 0.2285 \) customer, so \( W_q = L_q/\lambda = 0.2285/18 = 0.0127 \) hour, which is 0.76 minute waiting on average. This meets the manager's goal.
c. For \( M = 3 \), \( L_q = 0.03 \) customer, so \( W_q = L_q/\lambda = 0.03/18 = 0.0017 \) hour, or less than 0.1 minute waiting on average, which is almost no wait at all.

In summary, queuing theory is useful because it allows us to estimate system performance as long as we are able to reasonably meet assumptions. As you can see, queuing theory gets somewhat complex when we consider other than just the simplest assumptions. In situations in which the simple assumptions are not reasonable, we need to consider alternate methods of analysis, such as mathematical simulation.
<table>
<thead>
<tr>
<th>( \lambda/\mu )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.0111</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>0.0264</td>
<td>0.0008</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>0.0500</td>
<td>0.0020</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>0.0833</td>
<td>0.0039</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.30</td>
<td>0.1285</td>
<td>0.0069</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.35</td>
<td>0.1884</td>
<td>0.0110</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.40</td>
<td>0.2666</td>
<td>0.0166</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.45</td>
<td>0.3681</td>
<td>0.0239</td>
<td>0.0019</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.50</td>
<td>0.5000</td>
<td>0.0333</td>
<td>0.0030</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.55</td>
<td>0.6722</td>
<td>0.0449</td>
<td>0.0043</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.60</td>
<td>0.9000</td>
<td>0.0593</td>
<td>0.0061</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.65</td>
<td>1.2071</td>
<td>0.0767</td>
<td>0.0084</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.70</td>
<td>1.6333</td>
<td>0.0976</td>
<td>0.0112</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.75</td>
<td>2.2500</td>
<td>0.1227</td>
<td>0.0147</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.80</td>
<td>3.2000</td>
<td>0.1523</td>
<td>0.0189</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.85</td>
<td>4.8166</td>
<td>0.1873</td>
<td>0.0239</td>
<td>0.0031</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td>8.1000</td>
<td>0.2285</td>
<td>0.0300</td>
<td>0.0041</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.95</td>
<td>18.0500</td>
<td>0.2767</td>
<td>0.0371</td>
<td>0.0053</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>3.3333</td>
<td>0.0454</td>
<td>0.0067</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.50</td>
<td>1.9286</td>
<td>0.2368</td>
<td>0.0448</td>
<td>0.0086</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.00</td>
<td>0.0888</td>
<td>0.1730</td>
<td>0.3980</td>
<td>0.0090</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.50</td>
<td>3.5112</td>
<td>0.5331</td>
<td>0.1304</td>
<td>0.0339</td>
<td>0.0086</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.00</td>
<td>1.5282</td>
<td>0.3541</td>
<td>0.0991</td>
<td>0.0282</td>
<td>0.0077</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.50</td>
<td>5.1650</td>
<td>0.8816</td>
<td>0.2485</td>
<td>0.0762</td>
<td>0.0232</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.00</td>
<td>2.2164</td>
<td>0.5694</td>
<td>0.1801</td>
<td>0.0590</td>
<td>0.0189</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.50</td>
<td>6.8624</td>
<td>1.2650</td>
<td>0.3910</td>
<td>0.1336</td>
<td>0.0460</td>
<td>0.0155</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.00</td>
<td>2.9375</td>
<td>0.8102</td>
<td>0.2785</td>
<td>0.1006</td>
<td>0.0361</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.50</td>
<td>8.5902</td>
<td>1.6736</td>
<td>0.5527</td>
<td>0.2039</td>
<td>0.0767</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.00</td>
<td>3.6878</td>
<td>1.0707</td>
<td>0.3918</td>
<td>0.1518</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.50</td>
<td>10.3406</td>
<td>2.1019</td>
<td>0.7298</td>
<td>0.2855</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.00</td>
<td>4.4471</td>
<td>1.3471</td>
<td>0.5172</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.50</td>
<td>12.1088</td>
<td>2.5457</td>
<td>0.9198</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.00</td>
<td>5.2264</td>
<td>1.6364</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.50</td>
<td>13.8914</td>
<td>3.0025</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6.0183</td>
</tr>
<tr>
<td>9.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15.6861</td>
</tr>
<tr>
<td>10.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chapter 4  Service Design

1. The chapter began by reviewing the importance of tangibles in services.
   a. All services involve a bundling of tangibles and intangibles.
   b. Both the tangibles and intangibles have to be correct.

2. We then discussed the key elements of a service design.
   a. An important aspect of service design is that customers provide inputs to all services.
      This aspect is a major distinction between services and manufacturing.
   b. B2B and B2C services differ in that the entities are different.

---

Summary

1. The chapter began by reviewing the importance of tangibles in services.
   a. All services involve a bundling of tangibles and intangibles.
   b. Both the tangibles and intangibles have to be correct.

2. We then discussed the key elements of a service design.
   a. An important aspect of service design is that customers provide inputs to all services.
      This aspect is a major distinction between services and manufacturing.
   b. B2B and B2C services differ in that the entities are different.
3. We followed our service design discussion with details of the process chain network (PCN) diagram.
   a. PCN diagrams provide a basis for process positioning, that is, determining best how to interact with customers and suppliers.
   b. PCN diagrams place service processes in the context of supply chains.
4. We then provided tools for managing services with uncertain capacity.
   a. The first tool was the newsvendor model, which is useful for perishable capacity.
   b. Next, we discussed queuing models as a means for optimizing customer service and waiting lines.

**Key Terms**

back office 105  
coproduction 109  
critical fractile 110  
customer experience 102  
customer interaction 101  
direct interaction 104  
do-it-yourself (DIY) 105  
etity 103  
front office 105  
independent processing 104  
multiphase queue system 112  
newsvendor analysis 110  
poka yoke 100  
process chain network (PCN) diagram 103  
process domain 104  
queue 102  
queue configuration 112  
queue discipline 113  
queuing psychology 112  
queuing theory 109  
service capacity planning 108  
service offering 102  
service operations 98  
service recovery 101  
service stations 112  
service system 112  
simultaneity 99  
surrogate interaction 104  
tangibles 98  
time-perishable capacity 100  
two-phase system 112

**Integrative Learning Exercise**

Identify an organization, company, or business that provides a service. Use a process chain network (PCN) diagram to evaluate the interactions between service providers and customers for the organization. Be sure to identify the process domain region, the direct interaction region, the surrogate interaction region, and the independent processing region of the service design. If applicable, recommend improvements to the service delivery system that you evaluated.

**Integrative Experiential Exercise**

Together with a student group, visit a business or organization that provides a service. Identify a process or process segment in the business or organization that can be analyzed using a PCN diagram. Identify the process level, the process entities, and the beginning and ending steps of the process segment. Be sure to identify the points where the customer receives benefits and the provider incurs costs. Finally, comment on how moving and rearranging steps across and between process regions might affect the value proposition by increasing customer benefits, decreasing provider costs, or both.

**Discussion Questions**

1. Briefly describe service operations and service.
2. In what ways do services involve tangible elements?
3. Identify the customer input resources and the service provider outputs for the following service operations: accounting, education, computer repair, and healthcare.
4. What is meant by simultaneity in services? What is a major consequence of simultaneity?
5. What are some long-term responses for increasing and decreasing service capacity?
6. What is meant by the term time-perishable capacity as it relates to service operations? Provide an example.
7. Customers are generally involved in the service delivery process. What are some negative consequences associated with customer interaction in the service operation?
8. Briefly define and describe how a process chain network (PCN) diagram can be used in designing service delivery systems.
9. How can you shift the focus of your operations using a PCN diagram?
10. What trade-offs are generally made when making capacity decisions?
11. How do capacity choices vary in the near and long terms?
12. How can queuing theory be used to help evaluate capacity decisions for service providers?
13. Queuing psychology identifies ways that service operations managers can improve waits by improving the perception of those customers who do wait. What are some of the fundamental points related to queuing psychology?
14. In waiting lines, sometimes technological advances cannot make it easier to manage queues. How can psychology help with this problem?
15. How does the newsvendor model allow service firms to evaluate capacity decisions?

Solved Problems

Planning Service Capacity for Uncertain Demand

CAPACITY PLANNING TOOLS

SOLVED PROBLEM 4.2
1. A bookstore must decide how many copies of a special release of a political thriller to order. The demand for the book is assumed to be normally distributed with a mean of 2500 and a standard deviation of 150. The bookstore will sell the book for $25. It costs the bookstore $15 for each copy it stocks. There is no market for the book once the next book in the series is released; therefore, the book has no salvage value for unsold copies. How many copies of the book should the bookstore stock (order) if it wants to maximize its expected profit?

Answer:
\[ C_U = \text{price} - \text{cost} = 25 - 15 = $10 \]
\[ C_O = \text{cost} - \text{salvage} = 15 - 0 = $15 \]
\[ CF = \frac{C_U}{C_O + C_U} = \frac{10}{10 + 15} = 0.4 \]
\[ z = -0.25 \]
\[ Q = \mu + z(\sigma) = 2500 + (-0.25)(150) = 2462.5, \text{or 2463 books} \]

SOLVED PROBLEM 4.3
2. A small video store has a single checkout aisle staffed by one cashier. Customers arrive at the checkout aisle at the rate of 25 every hour. The cashier is able to process 40 customers per hour. Her service time is estimated to be exponentially distributed. Find the following:
   a. Average use of the machine
   b. Average line length
   c. Average time spent in line
   d. Average number of customers in the system (in the video store)
   e. Average time spent in the video store

Answer:
\[ \lambda = 25 \text{ per hour} \]
\[ \mu = 40 \text{ per hour} \]
\[ a. \text{average server utilization} = \frac{\lambda}{\mu} = \frac{25}{40} = 0.625 \]
\[ b. \text{average line length} = L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{25^2}{40(40 - 25)} = 1.042 \text{ customers} \]
\[ c. \text{average time spent in line} = W_q = \frac{L_q}{\lambda} = \frac{1.042}{25} = 0.0417 \text{ hour} = 0.0417 \times 60 = 2.5 \text{ minutes} \]
\[ d. \text{average number of customers in the system (in the video store)} = L_s = L_q + \frac{\lambda}{\mu} = 0.0142 + \frac{25}{40} = 1.6667 \]
\[ e. \text{average time spent in the video store} = W_s = \frac{L_s}{\lambda} = \frac{1.6667}{25} = 0.0667 \text{ hour} = 0.0667 \times 60 = 4 \text{ minutes} \]

SOLVED PROBLEM 4.4
3. A clothing store has a single machine that screens logos onto shirts. The time to screen on the logo is a constant 3 minutes. On average, there is a request for 10 screened shirts per hour. Compute the following:
   a. Average use of the machine
   b. Average number of shirts waiting to be screened
c. Average time a shirt waits to be screened
d. Average number of shirts waiting in the system
e. Average time spent in the shirt-screening process

Answer:
\[ \lambda = 10 \text{ per hour} \]
\[ \mu = \frac{60 \text{ min}}{3} = 20 \text{ per hour} \]
a. average utilization of the machine = average utilization of the server
\[ \frac{\lambda}{\mu} = \frac{10}{20} = 0.5 \]
b. average number of shirts waiting to be screened
\[ L_q = \frac{\lambda^2}{2\mu(\mu - \lambda)} = \frac{10^2}{2(20)(20 - 10)} = 0.25 \text{ shirt} \]
c. average time a shirt waits to be screened
\[ W_q = \frac{L_q}{\lambda} = \frac{0.25}{10} = 0.025 \text{ hour} = 0.025 \times 60 = 1.5 \text{ minutes} \]
d. average number of shirts waiting in the system
\[ L_s = L_q + \frac{\lambda}{\mu} = 0.25 + \frac{10}{20} = 0.75 \text{ shirt} \]
e. average time spent in the shirt-screening process
\[ W_s = \frac{L_s}{\lambda} = \frac{0.75}{10} = 0.075 \text{ hours} = 0.075 \times 60 = 4.5 \text{ minutes} \]

SOLVED PROBLEM 4.5

4. Calls arrive at a help desk at the rate of three every 2 minutes. The help desk has four service representatives who are able to handle incoming calls. If, on average, a representative can process one call in 2 minutes, compute the following:
   a. The average utilization of the representatives
   b. Average time a caller must wait to have his or her call answered
   c. Average amount of time a caller spends in the system.

Answer:
\[ \lambda = 90 \text{ per hour} = \frac{60 \text{ min}}{2} \times 3 \]
\[ \mu = 30 \text{ per hour} = \frac{60 \text{ min}}{2} \]
\[ M = 4 \]
a. average utilization of the representatives
\[ \frac{\lambda}{M\mu} = \frac{90}{4 \times 30} = 0.75. \]
b. average time a caller must wait to have call answered
\[ W_q = \frac{L_q}{\lambda} \]
Using Table 4.5:
\[ \frac{\lambda}{\mu} = 3 \]
\[ M = 4 \]
\[ L_q = 1.5282 \]
\[ W_q = \frac{1.5282}{90} = 0.017 \text{ hour} = 0.017 \times 60 = 1.0188 \text{ minutes.} \]
c. average amount of time a caller spends in the system
\[ W_s = \frac{L_s}{\lambda} \]
\[ L_s = L_q + \frac{\lambda}{\mu} = 1.5282 + \frac{90}{30} = 4.5283 \]
\[ W_s = \frac{4.5283}{90} = 0.0503 \text{ hour} = 0.0503 \times 60 = 3.0189 \text{ minutes} \]

Problems

Planning Service Capacity for Uncertain Demand

NEWSVENDOR PROBLEMS

1. A local bookstand believes that the demand for the Olympic edition of a sports magazine is normally distributed with a mean of 1200 and a standard deviation of 200. Each copy of the magazine costs the bookstand $1.50 per copy, and the bookstand will sell the issue for $5.00. Following the Olympic Games, there will be no demand for the magazine, and all leftover copies will be recycled because they will have no salvage value. What is the optimal number of copies of the Olympic edition that the bookstand should order?

2. The demand for next year’s wildlife calendar at a bookstore is assumed to be normally distributed with a mean of 500 and a standard deviation of 75. Each calendar costs the bookstore $5.50 each and will be sold for $12.50 each. Any calendars remaining for sale after Christmas will be discounted and sold for $1.00 each. The bookstore believes that any calendar remaining to be sold after Christmas can be cleared at the $1.00 price. How many wildlife calendars should the bookstore stock if it wants to maximize its expected profit from wildlife calendars?

3. A retail store must decide how many Mother’s Day cards to have in stock for this year’s Mother’s Day. Cards must be ordered months in advance, and there is only an opportunity to order one time. The store believes that the demand for cards will be normally distributed with a mean of 750 and a standard deviation of 50. The cards cost the shop $2.00 each and will be sold for $3.00 each. Any cards remaining after Mother’s Day will be destroyed. How many cards should the retail store order for Mother’s Day?
Queueing Theory

SINGLE SERVER QUEUING

4. A small coffee shop has a single barista. Because the shop is small, there are no tables or chairs. Consequently, customers wait in a single line to order and receive their coffee and leave the shop as soon as their order is received. Customers arrive at the shop at the rate of 15 per hour. It is estimated that the barista needs, on average, 120 seconds (exponentially distributed) to serve each customer. Determine the following service operating characteristics for the coffee shop:
   a. Average server utilization
   b. Average line length in the coffee shop
   c. Average time spent in line
   d. Average number of customers in the coffee shop
   e. Average time spent by customers in the coffee shop (includes waiting time and service time)

5. An ATM located on a college campus is able to process, on average, one customer request every 2 minutes. (Assume that the processing times are exponentially distributed and that each person using the ATM makes only one request per visit.) If students arrive at the ATM at the rate of 10 per hour throughout the day, find the following:
   a. Percent of time the ATM is busy
   b. Probability the ATM is not in use when a student arrives
   c. Average number of students waiting to use the ATM
   d. Average time a student spends waiting in line to use the ATM
   e. Average time spent waiting and using the ATM
   f. Average number of students waiting and using the ATM (i.e., the average number of customers in the system)

6. A quick-service restaurant has a single drive-through lane with one worker at the window. It is assumed that the worker can process an order every 4 minutes on average and that the processing (service) times are exponentially distributed. If customers arrive at the drive-through at the rate of 10 per hour, calculate the following service operation performance characteristics:
   a. Percentage of time the worker at the drive-through is busy
   b. Average number of cars waiting at the drive-through
   c. Number of cars, on average, in the drive-through system
   d. Average time a car spends waiting in the drive-through
   e. Time, on average, spent in the drive-through system

SINGLE PHASE WITH SINGLE SERVER AND CONSTANT SERVICE TIMES

7. Tourists arrive at a security checkpoint at the rate of 25 per hour. The checkpoint has a single machine that scans all tourists wanting to enter the secure area. Tourists are scanned one at a time according to a first-come, first-served discipline. A constant 90 seconds are required to scan a tourist. Compute the following:
   a. Percent of time the scanning machine is busy
   b. Average number of tourists waiting in line to be scanned
   c. Average time spent in the line waiting to be scanned
   d. Average time spent in the system (waiting time and scanning time)
   e. Average number of tourists in the system

8. A vanilla milkshake machine requires a constant 3 minutes to make the special shake served at the Burger Shack. The machine can process one vanilla shake at a time. Requests for the special shake occur every 4 minutes, or 15 per hour on average. Compute the following:
   a. Probability that the shake machine is busy
   b. Probability that the shake machine is not busy
   c. Average amount of time spent waiting to use the shake machine
   d. Average amount of time spent in the shake process (waiting and processing time)
   e. Average number of shakes ordered but not yet started processing (i.e., waiting for machine)
   f. Average number of shakes in process (ordered but not yet delivered to customers)

9. A cleaning tool requires a constant 5 minutes to polish a part. Parts arrive at the tool for processing at the rate of one part every 10 minutes.
   a. On average, what percent of the time is the cleaning tool busy?
   b. On average, how many parts are waiting to be cleaned?
   c. On average, how many parts are in the cleaning process?
   d. On average, how long does a part spend waiting before beginning processing?
   e. On average, how long does a part spend in the cleaning process?

MULTISERVER SYSTEM WITH EXPONENTIAL SERVICE TIMES

10. Customers arrive at teller windows at a bank at the rate of 12 per hour. A teller is able to process 15 customers per hour on average. If the bank has a goal that customer wait time should be no more than 30 seconds to see a teller, how many tellers should the bank have available to process customer requests?

11. A video store has four checkout lanes. The clerks in each lane have the same abilities, with each clerk being able to check out a customer in 3 minutes on average. At its busiest times, customers to the video store arrive at the checkout counters at the rate of 50 per hour.
   a. What is the average waiting time if all four checkout lanes are being used by the video store?
   b. The video store is contemplating closing one of the checkout lanes. If it reduces the number of lanes from four to three, how will customer wait time be affected during the busiest store hours?
   c. If the video store reduces the number of checkout lanes from four to three, how will the average processing time (service time) change?

12. A company can easily expand its service operation by adding more servers to its process. If each server is able to process 10 customers per hour and customers arrive at the company at the rate of 25 per hour, how many servers should be employed to achieve the standard that no customer should have to wait more than 30 seconds to be served? What is the average utilization of the servers for the number of servers required to meet this service goal?
Tommy Hernandez had recently been assigned to the service design team at XLG Enterprises. Tommy had been with XLG for a little over two years when the opportunity to join the service design team became available. The service design team performs a variety of roles, one of which is to analyze and recommend improvements for existing customer service operations performed at XLG.

The design team is now analyzing a new customer service process. The process would handle a variety of customer requests, including billing disputes, shipping and product delivery issues, and product returns. These activities would take place at a newly designed service facility close to the XLG headquarters. Most of XLG’s customers are small to medium-sized businesses located in the same city as the proposed customer service facility. It is the hope of XLG management that the new central location for customer service will be a way to facilitate and expedite customer requests related to product billing, shipping, and returns. Customer orders would still be placed mostly over the telephone or the Internet. A sizable number of XLG customers would come to the customer service facility to pick up deliveries or to make returns. The facility would also handle customer-related issues concerning service and billing.

A stated goal of XLG management is that the facility should ensure that customers rarely have to wait more than 15 minutes before speaking to a service representative, even during the busiest of times. XLG anticipated that it would staff the new facility with two service representatives at all times. During the busiest times of the day, however, management recognizes that they might have to increase staffing to as many as six service representatives to meet their stated objectives.

Tommy has been asked to join the team that was designing the new facility. As part of his role, he is to conduct analysis of customer waiting times. Tommy has been given information related to expected customer arrival rates during the busiest service periods throughout the day, average service times, and costs related to both resource staffing and customer waiting. Here is a summary of the information given to Tommy:

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Arrival Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 a.m.–1 p.m.</td>
<td>10 per hour</td>
</tr>
<tr>
<td>1 p.m.–5 p.m.</td>
<td>15 per hour</td>
</tr>
<tr>
<td>5 p.m.–10 p.m.</td>
<td>6 per hour</td>
</tr>
</tbody>
</table>

Average service time per customer representative (all servers are assumed to have the same capabilities) 15 minutes

Cost of waiting (estimated hourly cost for customers in the XLG system, waiting and being served) $20

Hourly cost of service representative (including benefits) $25

Average waiting time goal 3 minutes

Tommy remembers studying waiting line analysis in his SC&O management course. A quick review of his old notes helps him recall that the total hourly costs of providing service
are comprised of two components: the cost of providing the service and the cost of having customers wait for service. He found the following equation:

\[
\text{Total cost} = \text{number of servers} \times \text{cost per server per unit of time} + \text{expected number of customers in the system} \times \text{cost per customer to wait per unit of time}
\]

Tommy realizes that management wants an analysis that would help them determine how to staff the facility throughout the day as well as an estimate related to the overall estimated cost of the staffing plan, including customer waiting costs.

Management is eagerly waiting for the results of his analysis, and Tommy has promised them a report early Monday.

**Questions**

1. Could the facility operate during the busiest period with only a single service representative?
2. If four service representatives are used during the busiest period, what are the facility's operating characteristics as they relate to waiting lines?
3. If four service representatives are employed, can the company attain its stated goal related to waiting time during the busiest period?
4. How many service representatives should be employed to meet the company’s stated service goal for each time segment provided?
5. What are the total costs associated with waiting for the cases in which four, five, or six representatives are employed during the busy period?