

# **Service Marketing and Management: Capacity as a Strategic Marketing Variable**

By

Steven M. Shugan

University of Florida

## ***Overview of this Chapter: Marketing With Capacity Constraints***

We start this chapter with the classical discussion of whether service marketing is different from the marketing of manufactured and extractive (e.g., agriculture, fishing, mining) goods. We conclude that although philosophical debates are inconclusive, certain problems are prevalent for service providers. One of those problems, faced by many but not all service providers, is the problem of coordinating marketing and operations. This problem is often made difficult by the presence of capacity constraints. We devote this chapter to the topic of marketing with capacity constraints.

After that general discussion, we argue that both service quality and service strategy are integrally related to capacity decisions. For it is capacity (i.e., available employee hours, available physical facilities, lengths of queues, etc.) that ultimately determines whether the service provider can satisfy and retain buyers. Capacity decisions also impact costs and profitability. There are basic compromises between creating additional capacity to better serve customers and increasing costs.

After discussing the relationship between capacity strategy and service strategy, we discuss capacity-constrained strategies in two settings. In the first setting, we examine the case when demand is predictable. This case arises when factors such as predictable seasonality allow us to accurately forecast peak and off-peak demand. Seasonality may be related to the time of day, the day of week, the month of the year or particular holidays.

The section on capacity-constrained strategies considers strategies such as demand shifting. Demand shifting occurs when service providers attempt to shift demand from peak to off-peak periods. We discuss both the social welfare as well as the profitability implications associated with these strategies. Beyond demand shifting, we also consider other strategies including non-price rationing, offering different levels of service at times of peak demand and the bundling of services.

One interesting part of peak pricing strategies occurs when the opportunity cost of a resource changes over time. For example, consider a restaurant where table space is severely limited but only during peak hours. In this situation, the price of foods (e.g., coffee) that take more time to consume should have much higher prices during peak hours because these foods consume more capacity. Off-peak, however, excess capacity lowers the opportunity cost of consuming capacity to zero.

The last section of this chapter, before the conclusions, considers the case of capacity-constrained strategies with unpredictable demand. This case often occurs when exogenous and unpredictable events impact the arrival of buyers. A change in interest rates, a change in weather

conditions, a special event or just randomness could all cause a sudden increase or decrease in demand.

This section discusses how unpredictable demand usually decreases service quality and causes longer waiting times. The section also discusses the impact of unpredictable demand on the capacity decision and whether to build excess capacity into the system. We discuss marketing strategies with unpredictable demand when prices are flexible and when they are not. We also discuss rationing and different methods for allocating capacity when service providers are unable to satisfy all customers.

The chapter ends with conclusions and directions for future research.

## ***Is Service Marketing Different?***

Most discussions of service marketing start with a justification for service marketing (Zeithaml and Bitner 2000). Some authors argue that service marketing is vastly different from the marketing of manufactured goods (Berry 1980, Shostack 1977, Onkvisit and Shaw 1991). However, it is obvious that when services include such diverse activities as banking, retailing, wholesaling, consulting, litigation and surgery, then there are very few characteristics shared by all services (Lovelock, Patterson and Walker 1998, Rust and Zahorik 1996). Therefore, it is counter-productive to argue that manufactured goods and services each have unique characteristics not shared by each other

Instead, we should merely state the obvious. Service providers find some problems more prevalent than manufacturers (Folland, Ziegenfuss, and Chao 1988; Sirdeshmukh, Singh, and Sabol 2002). Service providers face problems not frequently encountered in other sectors, i.e., manufacturing and agriculture (Spicer and Bernhardt 1997). Consider, for example, the product-line manager for soup. This manager is responsible for developing and implementing a marketing-plan for soup. The plan probably includes specific marketing instruments such as trade promotions that offer discounts to retailers. The plan also includes coupons, package design and shelf facings. The plan could also include some ideas for a new eight-ounce size. This manager needs to understand marketing instruments such as trade promotions, coupons, package design, shelf facings and package size.

Now consider the marketing manager for a major hospital. The manager of the major hospital has different concerns about the implementation and dissimilar marketing instruments. Legal constraints force the hospital to offer some services and not offer others. More business comes from physician and other referrals than customer choices. Marketing plans must consider the reaction of patients, physicians, insurance companies, government regulators, hospital staff and private agencies. Prices are only flexible for elective procedures. Promotional efforts must consider capacity and location constraints. Hours of operation and allocation of employee time are primary considerations (Hirschberg 2000).

Although this argument suggests service providers face different problems than manufacturing firms, this argument also applies within the service sector. Given the broad definition of the service sector (Shugan 1994), the sector embraces very different service providers. In fact, the service section is the miscellaneous sector of the economy for businesses not classified elsewhere. It is expected, therefore, that many service organizations are not only different from organizations in manufacturing and agriculture, but also different from each other. Services such as museums and railroads appear to have less in common than agricultural organizations such as cattle ranches and cash-crop farms. Similarly, banks and electric utilities appear to have far less in common than automobile manufacturers and electronics manufacturers. Hence, although it is important to understand problems that are prevalent among service providers, not all service providers will face all these problems.

We might wonder whether service businesses are unique in any way. At one level, everything is unique. Industry terminology, for example, is often specific. Different industries may have different terms for the same concept. Bridge-buying, forward-buying, stockpiling,

stacking, cross-purchasing, inventorying are different terms used in different industries for exactly the same concept. At this level of analysis, only experience in the specific industry applies.

However, it is wrong to conclude that service providers are entirely different and require completely new and unique approaches. Many marketing principles and marketing research tools apply equally well to both service providers and manufacturers. We should never conclude a business is unique. That conclusion is very dangerous and encourages each manager to ignore knowledge accumulated in other industries. It also encourages an inward focus so that managers ignore valuable innovations just because they do not occur in the same industry.

We should only conclude that most service providers face many common problems such as capacity constraints (Desiraju and Shugan 1999) and intensive use of human resources (Szymanski and Henard 2001). Given these common problems, there is a set of common principles and methods useful to these service providers. All service providers should be aware of existing tools for capacity management and development of human resources. Most marketing principles apply to every context including the marketing of agricultural products, durable goods, industrial products, services and even ideas. Let us not use superficial differences between service sector firms and other firms as an excuse to forget basic principles. We should not claim every industry is unique merely to avoid the work associated with learning advanced methods developed elsewhere. Many service firms do exactly that. They waste time reinventing the wheel and, in many cases, inventing an inferior wheel.

There are several good marketing textbooks (Zeithaml and Bitner 1999, Hoffman and Bateson 1997, Kurtz and Clow 1997) dealing with the important consumer behavior issues in services marketing (Zeithaml 2000). There are also several good service operations textbooks (Hart, Heskett, and Sasser 1990) dealing with the critical issues in service delivery. However, very few articles or textbooks deal with the integration of managing demand with operational considerations. This chapter focuses on that topic.

This chapter argues one key problem faced by many services is the management of demand given capacity constraints. We argue that capacity constraints are a primary motivator of marketing activities in a wide variety of services. We also argue the capacity decision is linked to the service quality decision. In fact, in many industries, capacity decisions may be the most important aspect of service quality.

## **Service Capacity as a Competitive Advantage**

### ***Service Capacity and Service Quality***

#### ***Service Capacity as Service Quality***

We can think of capacity as a very general term. In many ways, service capacity is an important measure of our ability to delivery service quality. Capacity represents our capabilities. Greater capacity implies a greater capability to supply service quality. Considering all the aspects of service quality (Parasuraman, Zeithaml, and Berry 1985), we realize that it is capacity that allows us to deliver them.

Our capacity strategy, therefore, directly impacts service strategy. For example, we can increase capacity by expanding the number of servers. As we increase the number of servers, our customers may enjoy shorter waiting times in the service queue or benefit from a higher probability of service. With greater capacity, we also have the capability of providing each customer with more service. Increasing capacity, therefore, actually improves the quality of the service output. We expect that capacity should be directly related to customer satisfaction (Anderson 1995; Anderson and Fornell 1999, 2000; Hauser, Simester and Wernerfelt 1994).

In many service industries, we can equate superior service to higher capacity. In retailing, for example, increasing the number of trained floor employees helps customers find assistance more quickly. In medical services, increasing the number of employees can decrease

waiting times and allow health-care workers to spend more time with each patient. In the airline industry, increasing the number of flights might allow passengers to choose more convenient flight times. For many home repair services, having more employees increases the probability the home repair provider can make a repair on the same day that the customer calls. We see that customer satisfaction (Anderson, Fornell, and Lehmann 1994, Anderson, Fornell, Rust 1997) is directly related to available capacity.

We also see that selecting capacity often indirectly determines the level of service that we provide. Stated differently, a superior service strategy requires a greater investment in capacity. More capacity in the form of more employees, larger facilities or just a faster processing speed, all translate into higher service levels. The result is more ability to provide customer service and a greater reliability when demand is uncertain (Herk 1993)

Finally, setting capacity can sometimes deter or prevent competitive entry (Perrakism and Warskett 1983, Spence 1977) and, other times encourage collusive behavior by competitors (Compte, Jenny and Rey 2002; Brock and Scheinkman 1985). Similar to increasing quality, increasing capacity and allowing advance sales can also create a dominant position in the market (Lee 2001). Of course, advance selling alone can increase profits when capacity constraints are binding or not (Shugan and Xie 2000, 2002, Xie and Shugan 2001).

### *Capacity as Strategic Advantage*

In this section, we will explore the ability of a firm to either compete on service capacity or use service capacity as a competitive advantage. Capacity may be the seller's most important resource (Wernerfelt 1995). It can lead to customer satisfaction on a multitude of dimensions (Hauser 2001).

Here, capacity means differentiation (Bergen, Dutta and Shugan 1996). We ask whether we can be different on capacity and, in doing so, insulate ourselves to some degree from competition. As we have seen, choosing capacity is similar to choosing the level of service because, when we choose a lower capacity, we also choose a lower level of service quality. Less capacity often implies longer waiting times in service queues and, in industries such as airlines, the probability of any service may decrease. Here, the ability to compete on capacity is similar to the ability to compete on service. Hence, we are exploring whether firms can compete on service itself.

In more service industries than previously believed (Allen and Liu 1995), scale economies are critical to achieving profitability. The need for these scale economies may limit a service provider's ability to choose capacity and compete on capacity. Large benefits from scale economies, for example, may require all service producers to invest in relatively large capacity. Without these economies, small service providers may face an overwhelming cost disadvantage. Consequently, competition may create a market with only a very small number of very large capacity providers. For these few producers, capacity is seldom the key differentiating factor. In office-products retailing, for example, nearly all retailers are relatively large and often of similar size. Similarly, many wholesale distributors in various device industries are small, but of similar size because of a lack of economies-of-scale. Here, capacity fails to provide higher quality service or a benefit for the customer.

There is another situation when capacity may provide very little advantage. In some industries, all users may want very similar levels of service. For example, when developing film, most users may want 1 hour processing and are not willing to pay much more for ½ hour processing. If the cost of providing ½ hour processing is much greater than the cost of providing 1 hour processing, all service providers would be forced into charging providing the same level of service, i.e., 1 hour processing.

We see that in some industries, cost considerations, such as economies-of-scale may require all service providers to make specific investments in capacity. We also see a lack of heterogeneity in user wants also limits our ability to increase capacity as a source of competitive advantage. Later in this section, we discuss other factors influencing capacity decisions. At this

point, however, we further explore the topic of when capacity can be a source of competitive advantage or, at least, differentiation.

To differentiate on capacity, both low-capacity and high-capacity strategies must be viable. In other words, if some service providers are already profitable at some level of capacity, we want to know whether we can enter the market with a different level of capacity and still be profitable. We want to know whether a corporate strategy using a different level of capacity is viable. For simplicity, we refer to this different strategy as a high-capacity strategy. In the following section, we examine the viability of a high capacity strategy when profitable low-capacity providers are already in the market. The concepts in this section are equally applicable to adopting a low-capacity strategy in a market already containing profitable high-capacity service providers. Our interest here is in whether service providers can compete by adopting different levels of capacity and, thereby, offering different levels of service quality.

## ***The Impact of Service Capacity Constraints***

### ***Interpreting Capacity Constraints***

The meaning of capacity varies from industry to industry. We will refer to capacity constraints as being either hard or soft. A hard capacity constraint strictly constrains or limits the service provider's ability to serve customers. As the service provider reaches capacity, the service provider becomes unable to serve additional customers.

When a hard capacity constraint becomes binding, the service provider can no longer serve additional customers. A restaurant, for example, may have insufficient seating capacity to serve additional customers during the peak dinner hour. When all tables at a restaurant become occupied, the capacity constraint is binding and the restaurant can no longer seat additional customers. At capacity, an accounting firm may have an insufficient number of accountants to complete additional tax returns during the tax season. Its capacity constraint becomes binding.

Note that when hard capacity constraints are binding, unused capacity may be lost forever. Although granting refunds may provide customer benefits, these refunds must be coupled with the opportunity cost associated with unused capacity. These opportunity costs are irrelevant when surplus capacity exists (Chu, Gerstner and Hess 1998).

In addition to hard capacity constraints, there are also soft capacity constraints. These constraints are not strictly binding, and it may be possible to serve additional customers. Serving these customers, however, may be very costly. For example, consider an electric utility. During the peak summer season, the utility may be unable to generate sufficient power to feed the many hungry air conditioners. Despite that constraint, the electric utility can purchase power from other utilities and use that power to meet the increased demand from its customers. Buying electric power on the open market, however, may be much more costly for the electric utility than generating the electric power itself.

Another feature of soft capacity constraints is the ability to expand capacity by decreasing service levels. Here, the firm may decrease service times or service rates as it approaches capacity. A health care provider, for example, may spend less time helping each patient as the medical facility approaches capacity. An airline may give passengers less seating choices as a plane approaches capacity. An amusement park may close particular attractions as the park approaches capacity or as demand for those attractions exceeds available seating. In each case, the service provider temporarily expands capacity to serve more users, but each user receives some qualitatively lower level of service.

Note that, in some industries, such as entertainment (Sawhney and Eliashberg 1996) service providers often offer a line of services. In that event, insufficient capacity at one service may encourage buyers to switch to another service that may be as profitable as the unavailable service. The consequence of this switching is a much lower opportunity cost associated with capacity constraints than for service providers with only one service.

We see that capacity constraints are sometimes not constraints but rather abrupt changes in the cost of serving customers. Hence, we can view capacity constraints as high costs. As the

service provider approaches capacity, the cost of serving additional customers dramatically increases. As an aircraft reaches capacity, the airline must substitute a larger plane or put passengers on a competitor's plane. As a professional service approaches capacity, the service must employ workers over-time at higher wages or employ temporary workers at still higher wages. As a telephone network approaches capacity, it becomes necessary to purchase additional capacity on the spot market.

Another factor influencing capacity constraints is the length of the planning horizon. A binding capacity constraint, in the short-term, can be relaxed in the long-term. Consider an airline. In the very short-term, airlines are unable to add an additional seat to a full airplane. Given more time, however, the airline could use a larger plane on a particular route to accommodate additional passengers. Given a still longer planning horizon, the airline could purchase additional slots and fly additional flights. Given the very long-term, the airline could buy additional airplanes. Hence, short-term capacity constraints could become less binding as the planning horizon increases.

To some extent, all capacity constraints can be soft given a sufficiently long planning horizon. Restaurants, for example, can build additional rooms. Telephone companies can lay additional lines. Accounting firms can hire additional accountants. It is, therefore, necessary to consider both the degree to which the capacity constraint is binding and the length of the planning horizon.

### ***Long-term versus Short-term Constraints***

In the short-term, a capacity constraint is not a managerial decision. It is a binding constraint faced by management. Rather than deciding on capacity, the short-term decision is how to ration capacity in times of high demand and how to forecast when rationing will be necessary.

At this point, we focus our attention on long-term capacity constraints. This distinction is important because we will view the capacity decision as an integral part of the overall corporate strategy. For example, when designing the Washington-New York shuttle, Eastern Airlines decided to commit to full capacity by allocating sufficient planes to meet maximum demand. Subsequent airlines that entered the same market often took other strategies. New York Air, for example, allocated only sufficient capacity for average demand. When the number of potential passengers exceeded the number of available seats, New York Air turned those additional potential passengers away.

We see that, in the long-term, a firm can always decide to purchase sufficient capacity to accommodate maximum or peak demand. Purchasing that additional capacity, however, can be expensive and increase the costs of the service provider. An airline, for example, that has many empty unused planes has higher costs than an airline with fewer planes.

In a competitive market, service providers who choose to have more capacity than other service providers will have higher expenses than service providers who choose a lower level of capacity. An airline, which desires more capacity, for example, would have the additional cost of purchasing more planes. Additional revenues caused by servicing additional customers will offset some of these expenses. The airline with additional capacity, for example, has more planes to fly additional flights. The airline, however, will only achieve additional revenues during periods when peak demand exceeded available seating. Moreover, the additional revenues are limited to the number of passengers who wanted to fly but were unable to fly because of limited capacity.

We see that having additional capacity alone may not generate sufficient revenue to justify the additional capacity. In many cases, additional revenues will fail to offset all the costs associated with additional capacity. There are at least three reasons for this situation. First, additional revenue made possible by more capacity is limited to times of peak demand. Second, additional revenue from added capacity is limited to the amount of excess demand during times of peak demand. Third, a better pricing strategy may be able to decrease demand during peak periods and generate additional revenue without the cost of additional capacity. Finally, having

additional capacity may fail to attract more customers when that capacity provides no tangible benefit for customers.

Now consider two service providers who are very similar, but one provider chooses to have more capacity than the other provider. The low-capacity provider acquires sufficient capacity to serve all customers in times of off-peak demand, but far less capacity than necessary to serve customers during times of peak demand. The high-capacity provider acquires sufficient capacity to serve most, if not all, customers during peak demand.

We see that the high-capacity provider will have the ability to serve more customers during times of peak demand, than the low-capacity provider. The high-capacity service provider will also have a disadvantage. During times of low demand, the high-capacity provider will suffer from more excess capacity. The high-capacity provider will more often be in a position of excess capacity than the low-capacity service provider. Those situations will cause the high-capacity service provider to have a lower capacity utilization rate than the low-capacity provider. In other words, the high-capacity provider will have less revenue per unit of capacity. In strategic terms, the high-capacity provider becomes the high-cost provider.

Note that, the excess capacity is not what makes the high-capacity provider into a high cost provider. It is the absolute amount of capacity owned by the high-capacity provider that inflicts the higher costs. The high-capacity provider would be the higher cost provider regardless of whether the capacity is fully exploited or goes idle. This distinction is important because a better understanding of the problem provides us with the proper focus. Our focus should not be on capacity utilization but, instead, on maximizing long-term profits. When we become a high-capacity provider, we expect additional revenue during periods of peak-demand. During those periods, we expect greater profits than our low-capacity competitors. During off-peak periods, the excess capacity is irrelevant to our decisions.

With higher costs, the high-capacity service provider must either be able to charge a higher rate (that is, price) than the low-capacity service provider or save costs elsewhere. At this point, let us consider only the case where the high-capacity producer gets a higher price. This higher price compensates the high-capacity service provider for times of excess capacity. The additional revenue from the high price enables the service provider to support the excess capacity during off-peak periods and satisfy additional customers during the peak periods. In the next section, we explore situations when a high-capacity high-price strategy is viable.

## **Capacity-Constrained Strategies with Predictable Demand**

### ***Peak and Off-peak Marketing Objectives***

The section discussed marketing strategies when the capacity constraint was non-binding or there was a threat of it becoming non-binding. This section discusses marketing strategies when the capacity constraint becomes binding. The constraint may be binding for only a short, and sometimes predictable, time each year. That predictability often comes from predictable seasonality. For services such as retailing, a peak season may correspond to only 16 to 33 days a year. During these days, the retailer encounters peak customer traffic. Moreover, these days may generate the critical sales volume that determines whether the retailer is profitable or not. The rest of the year may only generate sufficient profit to cover overhead expenses. Every dollar during the season, however, may represent the retailer's profit for the year (Greenidge 1983).

When the capacity constraint is binding, the service provider may be unable to serve all available customers. The capacity constraint becomes binding when the service provider decides to have less capacity than needed to meet peak-demand. As a service provider with insufficient capacity to meet peak-demand, we must consider implementing two strategies. These strategies are not mutually exclusive. Most service providers attempt to employ both strategies.

The first strategy is to reduce demand during the peak period by extracting greater user-surplus during the peak period. Either charging a higher price during the peak period or offering a less costly service during the time of peak-demand often accomplishes this strategy. A hotel,

for example, may raise rates while a retailer may offer fewer specials or promotions. The result is less demand during the peak-period. A very large price increase during peak-demand may reduce demand to the actual level of capacity. There are also non-pricing strategies which we discuss later.

The second strategy is to shift demand from the off-peak season to the peak-season. Here, the service provider gives users incentives to use the service during the off-peak season rather than during the peak-season. A restaurant, for example, may offer an “early-dinner special” which offers a lower price for dinner if the patron arrives at the restaurant before the peak-time. The restaurant hopes that some patrons will choose to eat earlier. If some patrons do, the restaurant will get better utilization of capacity before the peak-dining period, and still fill all tables during the peak-dinning period.

Note that more strategies are viable during the peak-period because potential competitors are already at capacity. At capacity, they pose less of a threat. For example, Haskel and Martin (1994), using survey data on capacity constraints merged into a panel industry data set, show a positive relationship between profits and binding capacity constraints.

## ***Marketing Strategies at the Peak***

### ***Pricing Strategies with One Capacity Constraint***

#### **Different Objectives for Private and Public Sectors**

When managing public services, reducing peak demand may itself be an objective (Crew, Fernando, Kleindorfer 1995). Within a larger context, it may be socially desirable to encourage citizens to consume less during times of peak demand. The benefits may be associated with the conservation of scarce resources. Public services may attempt to avoid, for example, socially undesirable consequences of peak demand including congestion, undue burden of scarce public resources, destruction of public goods and regressive taxation. These events occur in public services including the national parks, electrical utilities, the public highway system and the environment.

In private services, however, the objective is seldom as simple as merely decreasing peak demand. In private services, the reduction of peak demand is only important when it leads to additional revenues. These revenues can occur during the peak period, during the off-peak period or both. For example, increasing the rate charged for accounting services during the peak season may, first, increase the revenue from billed hours during the peak period, and, second, shift some demand to the off-peak period, generating more revenue during the off-peak period.

In this section we discuss how to increase revenues during the peak season. We refer to this type of revenue generation as price discrimination because users pay different prices during different time periods. A telephone caller, for example, who makes a call during a weekday at noon, pays more per minute than when making a call during a weekend at noon. Prices explicitly depend on demand.

New competition, unfortunately, sometimes limits ability to fully exploit the peak by raising prices. During times of peak demand, temporary entry of competitors may occur. For example, when the Christmas season nears and toy retailers approach capacity, other retailers, who usually fail to carry toys, do carry toys for the short peak period of Christmas. Hence, as a toy retailer, we may be unable to significantly raise prices during the Christmas season. Moreover, having more capacity does not necessarily make one the dominant firm in the marketing (Van Cayseele and Furth 2001).

## Pricing to Maximize Profits

### *The Theory*

In the last section, we discussed the optimal off-peak price when capacity is non-binding. We said that the capacity constraint is irrelevant and that it should not distract us from setting the best price. We should not attempt to decrease our price to fill any excess capacity. The excess capacity is irrelevant to our pricing decision.

To set prices, we should start by determining the best price assuming we face no capacity constraints. We call this price the best-unconstrained price. It is the price we would set were there no capacity constraints.

If the best-unconstrained price results in excess capacity, we are finished. We should use that optimal price and ignore the capacity. This situation should occur in situations of off-peak demand because we should set off-peak capacity to be non-binding. Hence, setting the off-peak price can be equivalent to setting a price without capacity constraints. The size of capacity has no effect on the off-peak price.

If, however, the best-unconstrained price results in a binding capacity constraint, then capacity does have an impact on the optimal price. The best-unconstrained price is not optimal and we need to take another step.

We need to increase the price to diminish demand to the point when the capacity is just sufficient to meet demand. Let us explore the reason for this action.

Once we fill capacity, we have little opportunity to extract any revenue from customers who are denied service. Increasing the price accomplishes this goal. Conceptually, we want to continue to increase our price until the peak-demand exactly equals our available capacity. At that point, we obtain the maximum revenues allowable given our current capacity. If our price is any lower, we fail to extract revenues from customers denied service. If our price is any higher, we no longer have a binding capacity constraint. At the optimal price, peak-demand exactly equals current capacity.

For example, consider electric utility industry. In this industry, peak price could be many times off-peak prices. Surveys of customers suggest that, on average, customers would be willing to pay at least 100 times the off-peak price rather than forego peak service for any period of time (Rose and Mann 1995). However, peak prices fail to reach those levels for three reasons. First, given that customers' willingness-to-pay is very high, electrical utilities maintain large amounts of excess capacity to insure against blackouts during peak periods. Some plants may only operate at the time of peak demand. Second, because electricity can easily move long distances, competition provides some downward pressure on peak prices. Finally, the public sentiment combined with extensive regulation provides some pressure for peak and off-peak prices to equate. Of course, the consequence may be higher off-peak prices.

### *Implementing The Theory*

In concept, the theory is clear. We should increase our price to diminish demand to the point that we have exactly sufficient capacity to meet demand. In practice, however, service providers may be reluctant to increase prices beyond a certain level. There are several reasons for this reluctance.

The primary reason is our inability to predict demand. With uncertainty, it may be inadvisable to raise prices to diminish demand. We discuss this issue in the next section that further discusses unpredictable demand.

Another reason is the difficulty of continuously changing prices over time. Many service providers must announce prices and these providers become committed to their announced prices. They cannot renege on those prices and insist on higher prices just because it is temporarily profitable to do so. Of course, whenever an over-booked airplane fills, the airline probably wishes it could increase the fare to everyone on the plane.

Finally, it is administratively difficult to set complex pricing schemes. A restaurant might prefer to charge a different price for every item for different hours of the day and different

days of the week. Complexity prevents them from doing so. Restaurants, instead, have different menus for lunch and dinner, have early-bird specials, have happy hours and adopt other simpler mechanisms for approximating optimal prices.

Despite the fact that pricing schemes must be simpler than theoretically desirable, it is still important to understand the optimal price. We must know the optimal price before we choose among, possibly, simpler schemes. Knowing the optimal prices helps us better choose among available options when we are unable to increase the price to the point of maximizing profits.

We know, for example, we want to raise the prices of all capacity consuming services that provide little contribution to profits. In the restaurant example, after-dinner coffee may cause patrons to occupy the table longer and cause the restaurant to forego profits from another patron occupying the table. The other patron might order a complete dinner and the coffee price should reflect that fact. When possible, the restaurant should charge a very high price for coffee.

## Framing Peak Prices

Psychological theory tells us that presentation or framing is important. Customers may like a concept when presented in one-way but not another. Considerable research suggests that price is no exception. How we present prices influence how customers respond to them.

It is often useful to present peak prices as regular prices rather than premium prices. In that way, off-peak prices become discounts from the regular price. Hence, the matinee price for a movie is considered to be a discount off the regular price. A weekend and night telephone rates are discounts from the daytime rate. In each case, the consumer who uses the service during the peak pays a regular price, but the regular price is higher than the off-peak price. Other examples include family nights at sporting events, two-for-one drink specials, free off-peak upgrades on car rentals or hotel rooms, coupons good only during off-peak hours, early-bird dinner specials, special residence off-peak admission to amusement parks and so on.

## Whether to Announce Peak Prices

In most cases, the off-peak price is lower than the peak price. Moreover, in most cases, the service provider announces both the off-peak and peak prices. Telephone companies, resort hotels, car rentals and accounting services, for example, announce higher rates during their peak periods of demand. Here, the user knows the price during the peak period. This knowledge gives many users a greater ability to shift demand to off-peak periods.

The primary reason for announcing peak-prices is to shift some demand to off-peak periods. As a service provider, we hope to shift the demand that we are unable to satisfy, because of capacity constraints, to a time when capacity is non-binding. That leaves only the most price-insensitive users who will pay the highest price for our limited capacity.

When demand is unpredictable, there is another reason for announcing peak-prices. With unpredictable demand, more capacity implies a higher probability of service or a shorter waiting time. When we announce a high price, we signal users that our capacity is sufficient to meet demand. The high price only works with a competitive market because, as a high-priced provider, we could not survive unless we offered some service associated with our higher price. This reason for announcing price is only valid when demand uncertainty exists.

We see there are several reasons to announce peak-prices. There are also reasons not to announce peak-price. Rather than announcing prices, we can vary prices as capacity fills. We can link rates directly to demand conditions. Rates begin to change as capacity fills. Here, neither the service provider, nor the user knows tomorrow's rate. Consider, for example, a hotel booking space for a date in the future. On that date, the hotel plans to offer the first 20% of its rooms at a discount rate, the next 50% of its rooms at a regular rate, the next 20% at a high rate and the last 10% at a very high premium rate. As the hotel data approaches, the hotel implements the plan and the rate rises as the hotel fills. When the date is about to arrive, however, the hotel risks that any remaining rooms may remain empty. At that point, the hotel may no longer have a

binding capacity constraint and we face the conditions described in the previous section. Without binding capacity, in some cases, there may be a sudden decrease in rates.

Another reason for not announcing peak-prices is the opportunity to auction limited capacity. As capacity fills, we would like to sell the remaining capacity at the highest price. Remember, demand that exceeds our capacity, generates no revenue. Auctioning the last available capacity does extract the maximum revenue.

Despite the theoretical attractiveness of auctions, they are rare. Airlines do auction overbooked seats by offering greater and greater payments to flyers willing to surrender their current seats for seats on the next flight. Ushers may extract higher and higher tips for the best of the remaining seats. Tickets for sporting events may increase as the stadium fills. However, social pressures and transaction costs usually make auctions infeasible.

## ***Shadow Prices and Multiple Constraints***

### **Defining A Shadow Price**

To understand how to construct pricing strategies for meeting peak demand with multiple prices, we need to understand the concept of a shadow price. Shadow prices represent the additional profit that we would make if we could expand the underlying resource. The shadow price is higher when the capacity or resource constraint is more important. Capacity constraints with larger shadow prices are usually hard, i.e., difficult to relax, and are more likely to be binding when demand increases. Constraints with small shadow prices are less likely to be binding and, when binding, are relatively soft. When a constraint is non-binding, it has a shadow price of zero.

For example, suppose we own a restaurant. We face a capacity constraint on our ability to seat dinner parties. The constraint is based on the underlying resource consisting of the number of tables. During the off-peak Monday-night period, there are sufficient tables to meet demand and the shadow price for a table is zero. During the peak Friday-night period, there are insufficient tables to meet demand. It may be impossible to seat additional patrons when all tables become filled and so the resource, consisting of tables, is binding. In that case, the shadow price on the resource consisting of the number of tables is strictly positive, i.e., it is greater than zero.

Suppose that during Friday-night, when all tables are filled, we can earn \$40 from a one-hour dinner at a table for two people. Here, the shadow price of the table resource would be approximately \$40. A \$40 shadow price implies that we would be willing to pay up to \$40 to add an additional table during Friday-night. It also implies that anything that consumes a two-person table for one hour, on Friday-night, should be priced at \$40 more than during off-peak Monday night.

Finally, note that we have failed to discuss how to set optimal capacity. Although that decision is important, it may be possible to use price to adjust for mistakes in setting capacity. Hence, pricing decisions can compensate for suboptimal capacity (Skiera and Spann 1998) and reduce the loss in customer welfare (Berg and Tschirhart 1988). This later finding is important when uncertainty and long lag times can cause many firms to operate at suboptimal capacity (Bar-Ilan Sulem, and Zanella 2001).

### **Using Shadow Prices to Set Peak-Prices**

We see that a shadow price represents the opportunity cost of a resource. As a service consumes more of a limited resource, i.e., one with a binding capacity constraint, the service requires a higher price. The price should reflect the off-peak profit plus the shadow price of the limited resource. The optimal peak-price is the off-peak price plus the shadow price of the resource being consumed.

Consider our restaurant example. Suppose a dessert causes a table to be occupied for another six minutes, we should charge  $\$40 \times 6 \text{ minutes}/60 \text{ minutes}$  or \$4 more for the two

desserts during Friday night than during the off-peak period. Alternatively, we could offer a \$ 4 discount from the peak-price during off-peak periods.

Hence, the true cost of a service during the peak-period includes the shadow price of the capacity it consumes. While an after-dinner coffee may consume no resources during the off-peak period, that coffee consumes a valuable table during the peak-period. A restaurant should charge more for the coffee during the peak period, or offer a large discount during the off-peak period. When charging different prices is not possible, the restaurant should discourage after-dinner coffee during the peak-period.

Note that the concept of a shadow price implies a different strategy for managing a restaurant's offers. It suggests that we could offer a new appetizer rather than a new dessert. The appetizer would need to be ready-made, so that we could serve it quickly. Unlike desserts, we can serve appetizers while patrons already would be seated at a table waiting for their main courses. In that way, the appetizer consumes no capacity and provides additional revenue during peak demand. Moreover, a filling appetizer may become a substitute for dessert. That would free additional capacity because the average patron, who eats an appetizer rather than a dessert, may spend less time at the table.

Finally, remember that, during times of off-peak demand and non-binding capacity constraints, the shadow price for desserts becomes zero, because they do not consume a table. Consequently, we should have different prices for desserts during the peak and off-peak period. We should, in contrast, offer the approximately the same price for appetizers during the peak and off-peak periods, or possibly, charge less for appetizers during times of peak demand. Any service that does not consume a constrained resource should have the same price during the peak and off-peak period.

## Shadow Prices With Multiple Services

When we offer several services, the shadow price may become more complex. The sum of the resources consumed by the separate services may be less than the required resources for all of the services. This situation often occurs when customers simultaneously use several services.

Consider again our restaurant example. Suppose we have two desserts, an ice cream and a fruit dessert, that are equally expensive to prepare. The ice cream dessert requires 3 minutes of preparation by the server while the fruit dessert requires no preparation. Assume that, because of this preparation, dinner at a table that orders the ice cream dessert is 3 minutes longer than a table that orders the fruit dessert. According to our prior analysis, the ice-cream dessert should cost more. The shadow price of the ice cream dessert is  $\$40 \times 3 \text{ minutes}/60 \text{ minutes}$  or \$2. Hence, the ice cream dessert should cost \$2 more than the fruit dessert during Friday-night.

These computations assume that the table would be available 3 minutes sooner were it not for the ice cream dessert. This assumption holds if one person orders the ice cream dessert while the other orders the fruit dessert. Were both people at the table to order the ice cream dessert, the second dessert consumes no capacity because the table would not be free 3 minutes sooner. When the table orders two ice cream desserts, we should charge each only \$3/2 or \$1.50 more for the ice cream dessert than the fruit dessert. Of course, in practice, we may be unable to implement this optimal pricing strategy.

Now suppose that one person at the table orders still another dessert that takes five minutes to prepare. In that case, the ice cream dessert consumes none of the table resource constraint. The table would be occupied anyway and the ice cream dessert has no effect on the time the table is occupied. Hence, the peak-price of the ice cream dessert may depend on the dessert ordered by the other person at the table. The shadow price for the ice cream dessert can be \$3.00, \$1.50, or \$0 more than the off-peak price. It depends on what other dessert is ordered.

In practice, we may not know which desserts a table will order, nor can we charge a different price for different pairs of desserts. We can, however, determine the peak-price for the ice cream dessert based on our expectations about the likelihood of different orders. Suppose, for example, ice cream desserts had a 20% share of all of our desserts, while fruit desserts had a 40%

share, we might expect a shadow price of  $20\% \times \$3.00$  plus  $40\% \times \$1.50$  or \$1.20. Hence, we give diners a \$1.20 discount for the ice cream dessert during off-peak periods.

## Multiple Constraints

So far, we have considered a single capacity constraint on a single resource. With one resource, each service consumes some of that resource. A shadow price is a price that represents the value of the resource. When the capacity constraint is non-binding, the shadow price is zero. When the constraint becomes binding, it has a positive shadow price. The shadow price represents the importance of the resource in generating profits. Hence, we need to increase the prices of services by the amount of the resource consumed times the shadow price.

In this section, we consider the situation when the service provider faces multiple constraints and more than one may be binding at times of peak-demand. For every binding constraint, there is a shadow price associated with underlying resource. During times of peak-demand, we have a several shadow prices for each underlying resource. The price of a service needs to increase, during times of peak demand, by the quantity of each resource consumed times the respective price of each resource.

Conceptually, we start by identifying all capacity constraints on limited resources. We should then determine which constraints are most likely to be binding during times of peak-demand. We should determine which of those binding constraints, when relaxed, would generate the most incremental profits. These are our primary constraints. These are the constraints that have the highest shadow prices.

Constraints, whose relaxation, leads to less incremental profits, are secondary constraints. These constraints have smaller shadow prices. Finally, constraints whose relaxation leads to no incremental profits are not binding and have no shadow prices.

Let us return to our restaurant example. In that example, the constraint on the number of tables is our primary constraint. It is the primary constraint because one additional table could generate a profit of \$40 per hour, a relatively high shadow price.

In addition to tables, we face constraints on other resources such as the number of servers and the amount of food in the kitchen. The number of servers is a secondary constraint. We might stretch current services across more tables and allow servers to devote less time to each table. That action would generate less than \$40. Slower service would result. It might also suggest produce somewhat neglect patrons who might decrease tips and, consequently, require us to pay servers larger salaries to compensate for the lower tips. Mover, fewer servers may suggest less revenue from appetizers, alcoholic drinks and other revenue encouraged by servers. At this point, let us only consider the service time implications.

Suppose, for example, having one less server reduces available service time per table. Although we may have some surplus of server time, having one less server may still slow service time. Suppose, for example, the average service time per table increases, on average, 1 minute per table per hour for our 30 tables. Hence, losing one server costs us 30 minutes of service time. In that case, the shadow price for a server is  $\$40 \times 30 \text{ minutes}/60 \text{ minutes}$  or \$20 per hour. In sum, we would be willing to spend up to \$20 to add an additional server during this hour. Adding an additional server would provide 30 minutes of empty table space capable of generating \$20 in revenue.

Of course, this calculation is only approximate. We might want to do a more detailed calculation considering other effects of slow service time, including failure to obtain customer orders for drink refills and the consequence of providing customers less attention. For illustrative and conceptual purposes, however, our \$20 computation illustrates the general approach. Moreover, our attempt here is not to advocate complex mathematical computations, but to suggest a conceptual framework on how to think about capacity constraints.

Returning to our example, we also have a constraint on the amount of food in the kitchen. This tertiary constraint may be non-binding because the amount of food may be sufficient. At worst, we may have insufficient quantity of a popular item and may need to substitute a less

popular item. We could assign a small shadow price to reflect the cost of goodwill and, perhaps, the smaller profit margin associated with the surrogate item.

We can now use these shadow prices to determine the prices of different services. When we price off-peak, no capacity constraints are binding so we set prices without consideration of capacity. We call these prices the non-binding prices. When we approach capacity, the best or optimal prices for each service should equal the non-binding price plus the sum of the shadow prices of the resources consumed at capacity.

A dish that takes longer to prepare, causes a longer occupation of a table. For example, a dish that takes 3 minutes more than average to prepare, causes a table to be occupied by 3 minutes longer. The cost of that dish, during peak-times, should increase by  $\$40 \times 3 \text{ minutes}/60 \text{ minutes}$  or \$2. That same dish might also require 2 minutes more of server time to either prepare the dish or present the dish to the customer. During times of off-peak demand, this server time is non-binding and should not influence the price. During times of peak-demand, the server time is binding and the peak-price should increase to include the cost of the table space as well as the additional server time. At a rate of \$20 per hour, for example, the 2 minutes of server time costs us  $\$20 \times 2 \text{ minutes}/60 \text{ minutes}$  or 67¢. Hence, a dish, that takes 3 minutes longer to prepare and occupies a server for 2 minutes of presentation time, should have a price of \$ 2.67 higher during the peak period than during the off-peak period.

Conceptually, therefore, we should think of peak-prices as off-peak prices plus the sum of the shadow prices of resources with binding constraints. This conceptualization of an optimal peak-price with multiple resource constraints provides numerous pricing and non-pricing implications.

For example, this reasoning suggests a focus on services that consume the most important resource. Consider, for example, a large party. In large party, it is more likely that some member of the party will order something that requires maximum preparation or server time than in a small party. Hence, large parties are more likely to occupy more table space, and more of other resources. The implication is that we should charge more for meals at larger tables. Alternatively, we should offer larger off-peak discounts to tables with larger parties. Although practicality may limit our ability to do so, we should recognize the strategic implications. We should recognize the objectives of pricing strategy given binding capacity constraints.

There are also many non-pricing strategic implications. For example, we might want to develop special services to offer during times of peak-demand. These services would either be complementary or substitutable. In either case, we should develop services that do not consume valuable capacity. For example, serving drinks to waiting customers does not consume table capacity. Selling magazines to passengers on a train does not consume seating capacity on the train. We discuss these non-pricing strategies in the next section.

### ***Non-pricing Strategies for Peak Demand***

#### **Problems Raising Prices**

Despite the obvious advantages of raising prices when facing peak-demand, there are also many reasons that service providers are reluctant to raise prices during periods of peak demands. Most of these reasons are long-term in nature and relate to the long-term consequences of price increases. For example, non-profit service providers may fear public outcry and political repercussions. Private-sector service providers may fear that price increases may encourage competitors to enter the market. Price increases may encourage our customers to try competitive service providers and, if they like competitive services, we may lose those customers off-peak. We may also permanently lose those customers. Price increases may also generate ill will among customers leading to alienated customers and diminished off-peak demand. This reason is frequently cited for the failure to raise prices for popular sporting events and music concerts. Another reason is that price increases may encourage industrial customers to make long-term adjustments, such as changing procedures or finding alternative suppliers that might diminish long-term demand. Finally, price increases might conflict with a long-term growth strategy.

They may be counterproductive for building customer loyalty, lowering costs (Bailey and White 1974) or getting customer referrals (Biyalogorsky, Gerstner and Libai 2001).

There may also be short-term factors discouraging a temporary price increase during times of peak demand. These factors include transaction costs involved in changing the price. For example, retailers might have problems continuously changing prices. Direct marketers may have problems continuously changing catalog prices or other announced prices.

## Non-pricing Objectives

When discussing non-pricing strategies, we should remember that the objective of public sector services is different than private sector services. Public sector services tend to look for creative ways of decreasing demand such as brown-outs for electricity conservation, express lanes for car-pooling during peak rush-hours and ordinances that outlaw the watering of lawns during specific hours.

Private sector services, in contrast, usually have the objective of extracting additional profit from the peak demand period. It is usually insufficient to merely decrease peak demand. Therefore, non-pricing mechanisms for decreasing demand must have the property that they also increase profits and there are only two ways to do that. To increase profits without changing price, we can do one of the following. The first is to cut the cost of service delivery. The second is to shift revenue to future periods, increasing the profitability of future periods. At this point, we discuss non-pricing strategies for reducing costs.

As we said early, an off-peak discount is often equivalent to a peak price increase. This is also the case for non-pricing strategies. A decrease in service during the peak cuts costs, and it can be framed as an increase in service during the off-peak. No matter how we present the strategy to customers, the idea is increase peak profitability by decreasing peak costs. Examples of profitable non-pricing strategies for decreasing during peak periods include providing less service, forgoing investments to off-peak periods, increasing customer waiting times, giving less manufactured product and increasing customer participation in the service delivery process. Let us discuss each of these strategies.

## More Service Off-peak

The most obvious way to cut costs during peak periods is lowering the level of service. A retail floor employee may spend less time helping customers during peak periods and spend a greater percentage of total time handling transactions. A bus service may make fewer stops during times of peak demand and only stop at designated locations. A bakery may only offer stock items during times of peak demand but offer more customization (Anderson and Shugan 1991) during times of off-peak demand. An accountant may also offer only standard services during times of peak demand and much more customized services during off-peak periods. Off-peak seating at entertainment events may be more spacious.

To accomplish this strategy, it is necessary to be somewhat ambiguous about the nature of the service. The added services during the off-peak period are somewhat special and unexpected by the customer. Off-peak services may be more creative and go the extra mile. The best or optimal level of service during peak demand is usually lower than the optimal level of off-peak service.

The higher optimal level for off-peak service does have a disadvantage. Optimal off-peak activities often involve more service for every provider. For that reason, service providers often consider off-peak periods as more competitive. They attribute some of that competition to having reduced primary demand (i.e., industry demand for all providers), but having the same number of competitors. The additional reason for that increased competition is the surplus of capacity and the fact that optimal service levels are higher for all service providers.

It is, therefore, very intuitive that off-peak periods bring different problems. In the peak demand period, we need to work at capacity and stretch our resources to the limit. In off-peak demand periods, we face decreased demand, increased competition and the need to deliver more

service. Off-peak periods present the opportunity to provide better service and get customer loyalty to carry us through the peak-demand periods.

### Forgoing Investments to Off-peak Periods

There are many activities done by service providers that represent important investments in the future. These activities include the training of employees, improvement of service facilities, developing relationships with key customers, giving employees vacation time and experimenting with new procedures and new delivery systems. All of these activities are important for the future of the service. Off-peak demand periods provide an excellent opportunity for these activities.

### Longer Customer Waiting Times

One of the most frequent consequences of peak demand is an increase in the waiting time for customers. At supermarkets, amusement parks, medical facilities, banks, copying services and many other services, customers can face very long waits during the peak period. In each case, there is a greater chance that a larger number of customers will arrive at the same time. Anytime the number of recent arrivals outnumber the number of available servers, a waiting line or queue develops.

Service providers can shorten the queue by adding more servers. Banks can add tellers. Supermarkets can add checkout people. Accounting firms can add accounts. Unfortunately, the cost of adding people is very high because the peak demand for one service provider usually coincides with the peak demand for competitive service providers. Consequently, every service provider is trying to hire temporary, but already trained, help at the same time. The result is an inability to systematically add servers.

Another problem with adding servers occurs when the service provider experiences daily or hourly peaks. It is very difficult to hire an additional employee for just a few scattered hours during the day. A restaurant, for example, might find it difficult to get waiters and waitresses who would only work during the lunch and dinner hours. An accounting firm would have great difficulty finding accountants to work just a few months out of the year.

As we noted earlier, it is often best to build insufficient capacity to meet peak demand. The result is, unfortunately, longer waiting times. We should plan for these wait times and try to manage them. Some firms provide entertainment during the wait, they providing seating for waiting customers, they provide comfortable waiting areas and many firms try to carefully control their queues. Some restaurants, for example, encourage waiting customers to visit the lounge and enjoy a cocktail until their table is ready. Some large buildings install mirrors in waiting areas so that customers can preen themselves while waiting for the elevators, restaurants and other services in the building.

### Less Manufactured Products

The easiest way to cut costs is to provide less manufactured products during peak and more off-peak. Examples include two-for-one drinks, early-dinner specials that include a free dessert, an accounting firm that provides a free service during the off-season and a season pass to several sporting events for just a little more than the price of one peak-season event.

Bundling is a useful way to implement this strategy (Oren, Smith and Wilson 1985). Here, the service provider offers a bundle of items during the off-peak period that includes items consuming capacity-constrained resources. Of course, during the off-peak period, these resource constraints are non-binding and, therefore, have a shadow price of zero. The early-dinner special at a restaurant, for example, includes a free dessert. The free dessert causes the party to stay at the table longer, consuming the table resource. During peak-hours the shadow price would be high and the corresponding cost to the service provider would be high. During off-peak hours, however, the shadow price is low and there is no capacity cost to the service provider. The cost

of the dessert is only the direct cost of the dessert, there are no indirect costs associated with using the table resource.

## Increasing Customer Participation

During times of peak-demand, the server resource is binding because all servers are busy. This creates a high shadow price for the server resource. As we discussed earlier, the service provider could hire additional servers but when all competitors share the same times for peak demand, there may be an insufficient number of temporary servers. This situation creates the opportunity to allow the customer to take a greater role in production of the service.

For example, during times of peak-demand, a retailer may either charge more for gift-wrapping services or require consumers to do their own gift-wrapping. During times of peak demand, an airline may require passengers to check some carry-on baggage that, during off-peak times, the airline would allow on the plane. Banks encourage the use of automated equipment such as ATM during times of peak-demand. Restaurants have buffet self-service on busy nights.

For most services, users already participate in the production of the service. The user's time is an important resource. During peak-periods when all servers are busy, the user's time carries a much smaller shadow price than the servers' time.

## Substitute Peak Services

One of the best non-pricing strategies for peak-demand management is the development of new services that have low shadow prices. After developing these services, we attempt to switch customers to these new services from existing services with very high shadow prices during times of peak-demand.

Consider the example of an amusement park where afternoons bring peak daytime demand. During that peak time, the park can add substitute amusements such as parades, fireworks, outdoor shows, outdoor concerts and pageantry. These substitute amusements siphon people from amusements with high shadow prices such as popular rides with limited system capacity (e.g., a limited number of people per vehicle or a small number of vehicles), indoor shows with limited seating and "stop-and-go rides" that shut down for loading and unloading.

We see that the key to developing new substitute services is to go beyond user benefits. In addition to user benefits, we must focus on the resources the new services consume. These services may provide slightly less benefits to users, during off-peak periods, than existing services, but during peak-periods, these services provide a good value. Once we consider shadow prices, these substitute services have the same customer benefit at a much low cost and, probably, price to the user. Hence, they provide the user with greater value.

## ***Off-peak Marketing Strategies***

### ***The Objective***

For non-profit services, shifting demand from peak to off-peak periods may be socially desirable (Burness and Patrick 1991, Radas and Shugan 1998). It may lessen congestion, pollution, conserve societal resources and provide other socially desirable outcomes. For profit-making services, the situation is different.

As we recall, decreasing peak demand was not itself an objective. Similarly, shifting peak-demand to off-peak periods is itself not an objective. It is only important when the shifted demand produces additional profits. That usually occurs when the demand lost from the peak-period does not significantly decrease revenues during the peak-period. For example, consider the postal service's campaign to get people to mail early for Christmas (Shostack 1984.). In this example, the same packaged would have been mailed later at the same postage. It is profitable to shift demand because the shifted demand has no impact on total demand but decreases peak period costs. It is also profitable to shift demand to off-peak when the peak demand would have been otherwise lost.

We must remember, however, that merely shifting demand often generates no additional profits (Radas and Shugan 1998, 1999). If we follow a good off-peak strategy, we need not necessarily concern ourselves about whether the demand comes from peak periods or from increasing the overall off-peak demand. We could very well decrease profits by shifting customers to the off-peak. We must worry about which customers we shift. We want to keep price-insensitive customers buying at the peak because these customers allow us to charge higher prices during the peak period. So we want to shift only price sensitive customers to the off-peak period. Accomplishing this objective requires a specific separating strategy.

We either need to identify price sensitive customers or, make an offer that causes price-sensitive customers to reveal themselves. For this reason, many firms keep extensive databases on customers. For example, the nation's largest hotel chain, Marriott Hotels & Resorts, uses a four million-person database of Honored Guest Awards members to plan off-peak promotions.

### ***Competitive Factors Influencing Our Strategy***

During periods of peak demand, we lose some customers because of our inability to meet all demand. With insufficient capacity, we must ration customers and deny service to some users. These users may be lost either temporarily or permanently. If losing customers during the peak were inevitable, it would be better to lose them temporarily than permanently by shifting lost customers to the off-peak period. However, we must remember that demand shifting is not necessarily a good strategy. It depends on the cost of demand shifting relative to other ways of increasing off-peak demand.

Consider a situation when a service provider believes that demand may exceed supply. The service provider may attempt to switch customers to an off-peak period by signaling customers, that is, telling customers that the service provider is near capacity. For example, consider reservations as a technique for signaling customers. Users may call an airline, a restaurant or another service provider. The service provider may then tell some users that no reservations are available. That information may cause some users to shift their purchases to off-peak periods. Some airline passengers may make reservations on another flight. Some diners may make a dinner reservations at a different time. Patients at a health maintenance organization may make reservations on a less busy day. Unlike an HMO, however, in other industries some customers may switch to competitive services. Some airline passengers may call competitive airlines. Some diners may visit competitive restaurants. In these cases, the service provider permanently loses these users by trying to shift demand.

It might be a better strategy to over-book or raise prices as demand nears capacity. Unable to forecast demand exactly, a restaurant must choose between possible long customer waits and possible lost customers from charging too much (with excess capacity). If the restaurant fails to take reservations, for example, customers may risk a very long wait for service. If the restaurant takes reservations and some customers fail to arrive, the restaurant may have empty tables while their reservation systems shifts customers to competitors. Strategies such as reservations are complex and we discuss those strategies later.

Remember here, our goal is to extract the maximum profits during the peak period. We never want to shift customers merely to increase off-peak demand. It is only profitable to shift customers when that is the most profitable way to increase off-peak demand. In other words, we must compare the profitability of demand shifting to the profitability of merely creating new demand during the off-peak period. Remember, that demand shifting may be only one of many methods of generating off-peak demand.

The most important factor influencing the profitability of a demand shifting strategy is the nature of the competition and competitive demand. In some cases, our peak demand coincides with our competitor's peak demand. In other cases, our peak demand is independent of competitor's demand. The precise timing of competitive peaks is very important. Let us consider why.

With more competitors, it becomes more difficult to shift demand because our competitors may capture our lost customers during the peak-period. Competitors, of course, may be unable to accept our lost customers when these competitors share the same peak-demand

periods as us. Consider, for example, Florida resort hotels during Christmas. All of these hotels face the same peak, that is, during Christmas. When our hotel fills, our competitors also fill. Consequently, some demand must shift to off-peak and some people will postpone their Florida trip to an off-peak month.

We see that when all competitors share the same peak, demand shifting becomes easier. It is also easier to extract maximum profits from the peak period. It is easier to either raise prices or execute an alternative non-pricing strategy when competitors are also at capacity. When competitors have a different peak period, we would have less ability to increase prices or extract additional profits because users could easily get service from a competitor. Moreover, when competitors are not at capacity, they are most attractive to our users because our competitors will likely have lower off-peak prices and higher off-peak levels of service.

The timing of competitive peak periods also influences our capacity and service strategy. When we face long periods of peak demand, and we face the same periods of peak demand as competitors, we have a relatively easy task managing demand. We can invest in larger capacity because we can more easily extract profits during periods of peak demand. In contrast, suppose that we face short periods of peak demand and that we face different periods of peak demand than competitors. Here, we have a far more difficult task. We cannot invest as much in capacity because we are unable to extract sufficient profits during the peak period for two reasons. First, the brevity of the peak period shortens the duration of peak profits and, thereby, decreases our total peak profits. Second, competitors offer credible alternatives for customers during times of our peak demand. So we are less able to extract profits during the peak because we would quickly lose customers to our competitors.

### ***Pricing Strategies***

To some extent, all price increases during times of peak-demand indirectly shift demand to off-peak periods. A higher rate for telephone calls during peak hours often causes some callers to shift their calls to lower-rate off-peak periods. Hence, we may not need an explicit strategy for shifting demand.

When demand is perfectly predictable, we need not consider shifting when setting the peak price. We should increase the peak price until the capacity constraint is just binding. In other words, we increase the peak price until peak demand exactly equals our capacity. We should have no excess capacity and no excess demand.

We do, however, need to consider shifting when we set the peak price when demand is somewhat unpredictable. Shifting may further increase the optimal price during the peak-period. Suppose, for example, that we overestimate peak demand and our price is higher than optimal. With that price, our peak demand falls below our maximum capacity and we unnecessarily lose some sales.

Shifting demand lessens this loss because some of those unnecessarily lost sales are merely shifted to the off-peak period. We still enjoy many of those sales, albeit at the lower off-peak price. So shifting decreases the loss from incorrectly forecasting demand and accidentally pricing too high during the peak period.

Faced with uncertain demand, the presence of shifting makes us willing to risk a higher price during the peak period than otherwise. Suppose, for example, we manage an airline that flies two flights. One flight is during the peak late afternoon. The other flight is in the off-peak evening period. Assume that without capacity constraints, we would maximize profits with a price of \$400. At a price of \$400, there would be 200 travelers wanting to board the late-afternoon flight.

Now assume the plane's capacity is 150 passengers. We should continue to increase the fare until only 150 of the original travelers still want to board the plane. Suppose that happens at \$450. At \$450, exactly 150 passengers want to board the plane.

Finally, assume we are wrong and only 149 passengers want to board the late-afternoon plane. The plane has one empty seat on the flight. We lose the opportunity to enjoy revenue from that empty seat. The loss, however, might be small when demand shifting is present because the lost passenger flies the evening flight. Although that passenger pays a lower fare, we enjoy the

revenue from that fare. Without demand shifting, the total fare is lost and we would adopt a lower peak price.

## ***Non-pricing Strategies***

### **Multiple Services with Different Peaks**

We said that excess capacity should not effect our off-peak decisions. We should not develop off-peak services to fill excess capacity. However, the optimal capacity decision depends on the extent of demand and duration of the peak and off-peak period.

We should purchase less capacity with a short peak period and a long off-peak period. We should purchase more capacity with a longer peak period. In other words, it is hard to justify a very large capacity to satisfy demand for only a short time period.

With several services, each generating peak demand during a different time period, we can extend the period of peak demand. With a longer period of peak demand, the optimal capacity increases. In other words, we can justify a larger capacity given more demand. Consequently, with counter-cyclic services, that generate peak demand during different time periods, we invest more in capacity. With a larger capacity, we can serve more customers during the longer periods of peak demand and we receive a higher corresponding profit for those periods. The result may be better service during every peak period.

Consider, for example, a manufacturer's representative for several winter-sports manufacturer. The representative may offer numerous services from order taking to constructing in-store displays. To provide our services, we employ a number of sales representatives and own a small warehouse. All of our services are highly seasonal, with demand growing strongly at the end of summer, peaking at the beginning of the winter and minimal during the spring.

Were we to begin distributing summer sports equipment, we would maintain a larger number of full-time employees and a larger warehouse, because we could now generate more profit from having additional capacity. In other words, by carrying summer sports equipment, we would have more employees and a larger warehouse throughout the year. With those additional resources, we could provide more service to our original winter sports equipment manufacturers.

We see that smoothing demand with new off-peak services, we can deliver better service during peak periods. That creates a strong incentive to smooth demand because it increases our ability to compete in all times periods. The problem is being sufficiently creative to find services that have different peaks and, that we can profitably deliver with the same resources.

Children's Software writers, who face peak demand near Christmas, would find it profitable to develop financial software, peaking in the winter, if their programmers could write for both markets. Mortgage companies, who face peak demand during the construction season, would find it profitable to develop other services such as automobile loans, that peak during the autumn.

### **Smoothing Demand with Bundling Across Time**

As we discussed earlier, we would like to substantially increase price during times of peak demand. Sometimes, unfortunately, there are upper limits on the extent to which we can increase prices. Huge price increases may convey the perception of price gauging. Moreover, when facing uncertain demand, constantly changing prices may create transaction costs. This is true in many industries where service providers announce prices in advance. A ski resort, for example, that books rooms months in advance, would find it difficult to suddenly increase prices after a heavy snow.

Here, bundling services across time provides a possible solution. Consider, for example, a season pass for a series of ten football games. Suppose we would like to charge \$20 for the less popular games and \$50 for the more popular games. Public relations, unfortunately, prevents us from charging more than \$35 for a game. Here, we charge \$35 for all the games, but we limit the number of single tickets to a very small number. We offer season passes to all ten games for

\$35×10 or \$350. Note that the season pass price is equivalent to charging \$20 for the less five popular games plus \$50 for the five popular games, because  $20 \times 5$  plus  $50 \times 5$  equals \$350.

We see that bundling across time provides a way of hiding the individual prices. In that way, we effectively charge more during the peak periods than our single period price suggests. We transfer profits from the peak period to the off-peak period by bundling purchase in each period.

Note that bundling may have an additional advantage when time-sensitive people are also price insensitive. In those cases, the price insensitive customers may not attend events during the off-peak. They may buy a bundled ticket and only use tickets for popular events. As a result, we incur lower costs during the off-peak because we incur no costs for those users, who have a ticket, but fail to attend off-peak events.

## Advertising and Offering Coupons at the Peak

The peak demand period provides an excellent opportunity to contact customers who may be interested in off-peak service. Unlike customers reached through less directed tools, such as television advertising, the peak identifies high potential customers. These users have already revealed a propensity to like our service because they have already used the service. It is now only necessary to interest them in additional off-peak services. This strategy is viable for all services that could be used again during the off-peak such as lodging services, transportation services, restaurant services, insurance services and many others.

When customers use the service during the peak, we can expose them to advertising concerning our off-peak service. These advertisements depend on the type of services. Restaurants place clear plastic displays, often called tents, on the table to advertise off-peak specials and attempt to enhance repeat usage.

There are many other creative ways of using peak periods to reach potential off-peak customers. We can target customers during the peak with coupons for off-peak services. During the peak, we can offer discounted tickets or passes for off-peak periods. We can also inform customers about price reductions and special promotions associated with off-peak usage of our service.

We must, however, be careful not to overdue our selling effort during the peak-period and subject customers to unrelenting advertising. Customers, unhappily waiting in line, may find advertising unattractive. In one case, a supermarket installed 13-inch color television monitors mounted at each checkout counter. The television monitors exposed customers, waiting in its queues, to paid advertising. They went beyond advertising items at the supermarket and included general commercials. Customer response was overwhelmingly negative. Customers complained that the commercials were annoying, distracting and noisy.

## Giving Priority to Loyal Customers

Another non-pricing strategy for using peak demand to increase off-peak demand is the establishment of priority queues. With priority queues, loyal customers get special priority during peak periods. Here, we define a loyal customer as a user who uses the service off-peak. Therefore, priority queues during the peak period encourage loyalty and enhance demand during the off-peak period. Customer loyalty leads to increased retention (Bolton 1998) and higher future profits.

There are examples of priority queues in many industries. Airline gives priority seating to frequent flyers that show demand during both peak and off-peak periods. Restaurants give priority seating to regular customers. Loyal industrial customers may get faster delivery during times of peak demand. Banks sometimes have special lines for regular commercial customers. Any services that schedule appointments can give priority to loyal customers and encourage off-peak demand.

We see that giving priority during the peak to loyal customer can help create off-peak demand. This strategy works best when loyal customers are also price insensitive customers such

a business travelers or high-income individuals. This strategy may not work as well when our loyal customers are also price sensitive customers. Let us consider the case when our most loyal customers are also more sensitive to price than our average customers at the peak.

Consider, for example, Disney world resort customers in Florida. Here, Disney enjoys regular visits by Florida residences throughout the year. Florida residences are quick to take advantage of off-peak specials and unexpected lulls in peak-demand. These loyal residences, however, are far more flexible with their travel plans than customers who fly to Disney from the north or corporations who must schedule conferences on particular dates. Florida residences, consequently, are both more loyal, while being far more price sensitive.

When loyal customers are more price-sensitive, it becomes more costly to give loyal customers higher priority. By giving priority to price sensitive customers, we may need to lower the optimal peak price to account for the high price sensitivity during the peak. As a consequence, we increase off-peak profitability at the cost of peak profitability. In general, demand shifting is not a good strategy. It is better to focus merely on generating off-peak demand, than shifting peak demand.

## Requiring Reservations

There are many ways reservations can help service providers stimulate off-peak demand. The most obvious way is to shift customers, who would otherwise be lost at the peak, to the off-peak. With reservations, we are able to fill to capacity and then transfer all excess demand to the off-peak. When a user calls, we try to match that user with a time period when capacity is available. We may also be able to extract the highest price by shifting the user to the most desirable time slot whose price does not exceed the user's reservation price.

Reservation systems work best when either little competition exists or our competitors face the same peak demand that we do. When users lack alternatives, we have maximum flexibility in assigning reservations. During the peak, we can completely book our capacity at very high prices. When either a user is unable to reserve our service at the peak, because of insufficient capacity, or when the user desires a lower price, we can provide the user with an off-peak reservation. The lack of user alternatives encourages the user to accept our off-peak reservation in lieu of switching to a competitor.

With many competitors that face different peak demand, the situation changes. We would have less flexibility. When we deny a user a peak-reservation and offer an off-peak reservation, the user may prefer a competitive peak-reservation. We, consequently, observe reservations in industries where few alternatives exist, such as physician or other professional services, or when all competitors face the same peak demand, such as hotels and restaurants.

Beyond competition, user preferences may also affect the profitability of reservation systems. Remember that a reservation is an insurance against either being denied service or a long wait for service. Some reservations provide more complete coverage than others. Greater the coverage is a greater benefit to users. Greater coverage also produces a greater cost for the service provider. Reservation systems are costly to the service provider because of both administration costs and the opportunity costs associated with customers who fail to use their reservation. Customers who fail to honor their reservations often cause less than ideal idle capacity during peak-periods and considerable opportunity costs. An airline, for example, may fly with an empty seat that would have otherwise earned a premium fare.

## Promoting to New Customers

The cost of giving free service, on a trial basis, is cheaper during the off-peak periods. Here, we only incur the incremental cost associated with the additional service. We offer trial services at a lower price. The intent, here, is to overcome the initial hesitation that a customer might have with our service. Our goal is to get that customer to continue using our service after the trial period. Therefore, it is important that we only offer trial prices to clients who have the

potential to become permanent customers. We should also avoid offering trial prices to existing customers who may only seek to obtain our service at lower prices.

Trial prices are common in many service industries. In promoting cellular phones, for example, one company offers free time during off-peak hours to potential new clients.

## **Capacity-Constrained Strategies with Unpredictable Demand**

### ***The Impact of Unpredictable Demand***

#### ***How Unpredictable Demands Decreases Service Quality***

The previous section discussed service strategy when demand for our service had predictable peaks. This section discusses service strategy when demand is relatively constant, but characterized by sudden and unpredictable peaks. This type of demand is found in many service industries when several customers, by chance, arrive at the same moment.

Consider, for example, a bank in the middle of the afternoon. Suppose there is one teller on duty. About 30 customers arrive each hour and it takes approximately 1 minute to service a customer. Customers, in our example, arrive independently. In other words, the arrival of one customer neither increases nor decreases the probability that another customer will arrive sooner or later. Moreover, in this example, the bank has sufficient capacity to serve all customers and, therefore, should not deny service to any customers. In fact, we would expect the teller would be idle 30 minutes of every hour because the teller would be with customers only 30 minutes of every hour. In other words, the bank has twice the capacity needed to meet the demand of 30 minutes of teller time per hour.

From a quick look at the situation, we might conclude that our system is inefficient because the teller is idle fifty percent of the time. We might also conclude that this excess capacity will provide a very high level of service because customers need not wait in line. Despite the presence of excess capacity, however, long lines may still develop at the bank. Suppose, for example, that rather than arriving in two-minute intervals, by chance, the 30 customers all arrive at the same time. In that event, the one teller starts to serve the first customer, and the other 29 customers begin to wait in a line or queue. The queue starts with a length of 29. With a 1 minute per customer service time and assuming that no new customers arrive, it takes 29 minutes before the teller serves the last person in line. We see that the average customer, in the queue, waits  $435/29$  or 15 minutes for service. Moreover, the last customer in the queue waits 29 minutes for service.

It is unlikely that many customers will wait 29 minutes or more, but it is very likely that some customers will wait. In fact, we can compute the average time we would expect customers to wait. Computing the average wait time requires a mathematical formula from queuing theory. We present that formula later in this section. That formula reveals that with an average 2-minute arrival rate for customers and an average service time of 1 minute, we predict that the average customer will wait 1 minute. Stated another way, the average customer spends 2 minutes in the bank and half of that time is spent waiting in line. Hence, customers spend fifty percent of their time waiting in line while tellers are idle spend fifty percent of their time.

In our example, the arrival of customers was independent and evenly distributed throughout the hour. This situation implies that we do not expect customers to arrive together, except by chance. Chance arrival by customers, of course, suggests a high probability that customers will sometimes arrive together. In addition, when customers arrive in groups, because they are traveling together as friends or coworkers, the situation is worse. Expected queues will be longer and the average customer will spend more time in the bank.

In general, we can think of capacity as a measure of our ability to provide service quality and queue length as a measure of actual service quality. Our previous example suggests that as demand becomes unpredictable, we require more capacity to provide the same level of service quality. We can have many idle servers, yet, with some positive probability, many of our

customers may still suffer long waits in queues. In the next section, we discuss how different types of demand uncertainty effect service quality.

## ***Factors Influencing Service Quality and Waiting Times***

### **Time for Service Delivery**

Two key components of uncertainty are service times and customer arrival rates. Service time is the time it takes for a customer to receive complete service. Customers often prefer faster service to slower service, keeping constant the benefits from the service. Despite customer preferences for faster service, service time is important for another reason. Service time has a dramatic impact on the total time that the average customer spends in the system.

The time the average customer spends in the system is the sum of the time being served and the time waiting for service. For example, at a health care facility, patients must first wait for care and, subsequently, obtain care. Their total time in the system is the sum of the time waiting for care plus the time actually receiving care from, say, a physician. Were one patient to spend more time with the physician, it would increase that patient's time in the system increase. In addition, the longer service time for the one patient would delay all patients currently waiting in line as well as any new patients who arrive during the delay. Hence, delays in service time can quickly increase the total time spent by all customers in the service system.

We often measure service time as the rate at which service occurs. The rate is the number of customers served per unit time. For example, if it takes 2 minutes to serve one customer, we say the service has a rate of  $\frac{1}{2}$  or .5 customers per minute. The service rate is a function of several factors including organization of the service, server training and customer preparation. A well-organized service sequences tasks to maximize efficiency. A retail check-out clerk, for example, may follow an efficient script of first asking for all merchandise, asking the type of payment, removing price tags for the merchandise, entering prices from tags into the computer, bagging the merchandise while printing the receipt and so on. Server training assists servers in maintaining efficiencies by developing contingent procedures for handling special problems. Finally, work progress analysis and other special methods can reduce server time.

The service rate varies from industry to industry. In some industries, the average service rate is relatively short. In other industries, the average service rate is relative long. In the next section, we explore the effect of service rate on service strategy and customer waiting.

### **How often Customers Arrive**

For some services, customers arrive frequently but service time is relatively fast. By frequent arrivals, we mean that we expect many customers to arrive in each time interval. By relatively fast service, we mean that a server can service many customers within a short time. Examples of frequent arrival/fast service situations include express lines in grocery stores, tollbooths on interstate highways, collection of tickets at entrances to theaters and directory assistance by telephone companies. For these services, one server usually serves many customers each hour. The server spends little time with a customer and queues move quickly. As a consequence, these services can provide higher levels of service with less capacity.

To see why, consider the unlikely, but possible, event of multiple arrivals. Suppose that 30 customers usually arrive per hour and all 30 customers arrive at the same time. If the service is fast, the line moves quickly and, despite the unfortunate timing, the last customer in the queue still has a relatively short wait. This reasoning suggests a less disastrous impact associated with simultaneous arrivals. Fast service allows us to quickly overcome the problems associated with simultaneous arrival.

Now let us consider the situation of infrequent customer arrival with relatively slow service times. For example, consider an interstate automobile towing and on-site repair service. The facility may expect very few calls in any day, but a single call may take several hours of time. With situations of infrequent arrival but slow service rates, we expect one server to spend more time with customers. We also expect shorter lines because we expect fewer customers.

Despite these expectations, infrequent arrival but slow service situations require more capacity to provide the same level of service. In other words, when lines do develop, our customers wait much longer than in the frequent arrival/fast service case.

To see way, let us consider our previous example where, on average, 30 customers arrive per hour and it takes, on average, 1 minute to serve a customer. That situation is a frequent arrival/fast service case. For the infrequent arrival/slow service case, consider a situation where only 6 customers arrive per hour but it takes 5 minutes to serve each of them. In each case, we expect one server to spend 30 minutes of every hour serving customers. In the first case, on average, each hour the server would spend 1 minute with each of 30 customers and 30 minutes idle. In the second case, on average, each hour the server would spend 5 minutes with each of 6 customers and 30 minutes idle. The two cases appear somewhat similar.

When it comes to average waits, however, the two cases are very different. Queuing theory formulas tell us that the average wait for the frequent arrival/fast service case is 1 minute while the average wait for the infrequent arrival/slow service case is 5 times longer, or 5 minutes. The average time spent in the service system for the frequent arrival/fast service case is 1 minute waiting and 1 minute being serviced, for a total of 2 minutes. The average time spent in the service system for the infrequent arrival/slow service case is 5 minutes waiting and 5 minutes being serviced, for a total of 10 minutes.

We see that in both cases we have similar capacities because in each case our server is idle, on average, fifty percent of the time. In every hour, our server spends about 30 minutes with customers. However, when customer arrivals are infrequent and service is slow, our customers wait 5 times longer. To keep average waiting times constant, for the two cases, we would need more servers in the second case. We conclude that infrequent arrival/slow service providers require more capacity to provide the same waiting times as frequent arrival/fast service. Both cases may have the same amount of idle capacity, but the expected wait times are very different. It, therefore, costs us more to provide the same waiting times and customers must pay more for the benefit of the same waiting time when service is slow. The consequence is an inherent efficiency problem in infrequent arrival/slow service situations.

### ***Strategic Implications***

#### **Technological Constraints on Service Quality**

##### ***Uncontrollable Service Rates***

In many cases, the nature of the business dictates the quality of the service in terms of service rates. For example, a travel agency that books vacations will have slower service than a tollbooth on an expressway. The technology of a travel agency service requires more time spent with customers to evaluate options and assess customer preferences. The technology of a tollbooth, in contrast, allows very fast interactions with customers and minimum contact.

Services such as unemployment agencies, traffic court, eye-glass examinations and hospital emergency rooms, will require more lengthy service times than services such as dry cleaning drop-offs, car washes and toll booths. These differences in service time are a function of the service technology and are somewhat beyond the control of the service provider.

Services that have technologically slow service rates will face a technological disadvantage. Not only must they provide more servers to compensate for slower service rates, but they must add still more servers to keep waits short. Here, a service that has twice the service time must have more than twice as many servers to remain even with the faster service.

Consequently, the payoff from improving service rates is much higher for services with slower service rates than services with faster service rates. Services with low service rates gain on both time with the server as well as shorter waits when they improve service. Unfortunately, services with slower rates often require more customization of their service, which limits their ability to improve service rates.

With slower service rates, we require more capacity because each server will be busy longer. For example, when a server takes 6 minutes to serve a customer, we will need more

servers than when each server takes only 1 minute. When server takes 6 minutes, that server is busy for six minutes and will be unable to service another customer for six minutes. Consequently, we need more servers to accommodate multiple arrivals. Otherwise, when multiple customers arrive in close proximity, we will either lose some customers or decrease the quality of service by increasing average waiting times.

### *Controllable Customer Arrival Rates*

Similar to service rates, the arrival rates of customers are somewhat of a function of the service technology. Travel agencies will have few arrivals per hour while tollbooths will have many arrivals per hour. Museums will have a continuous flow of customers while new car dealers may encounter more sporadic arrivals. As a service provider, we should seek to influence arrival rates. Service technology, however, may place limits on our influence.

Consider a situation when the rate of service is relatively slow. As we discussed earlier, the average wait time will be higher. It may take, for example, an hour to fully serve one customer. In that case, the unfortunate event of multiple arrivals has a more adverse effect on waiting time. With one server and the arrival of three customers, we will be able to serve only one of the customers. The second customer must wait one hour while the third customer must wait two hours. Our average wait times will be one hour.

As a service provider, we can insure against multiple arrivals by having more servers. Of course, having more servers suggests more idle capacity. More idle capacity, in turn, suggests a higher fixed cost and a higher rate to the consumer. In other words, customers must pay for insurance against long waits by subsidizing excess capacity.

The compromise between excess capacity and insufficient capacity depends on both the arrival rate and the incremental fixed cost associated with additional capacity. When the arrival rate increases, we are better able to justify more capacity. More frequent arrivals suggest a high probability that many customers will be in the service delivery system. As the number of customers increase, we expect more even demand and better utilization of capacity. We also expect to be better able to manage average customer waiting time because demand is more evenly distributed over time.

### *Combining Arrival Rates and Service Rates*

To get the same average wait time, we said that arrivals that are more frequent and slower service require more capacity than less frequent arrivals and faster service. In general, industries with slow service and frequent arrivals require the highest capacity. Industries with slow service and infrequent arrivals require the second highest capacity. Industries with fast service and frequent arrivals require the next highest level of capacity. Industries with fast service and infrequent arrivals require the least capacity.

Table 3 provides examples from different industries of the four cases. Table 4 illustrates the expected wait times for these different cases, holding constant capacity (i.e., the number of servers).

**Table 3: Arrival and Service Rate Examples**

	Many Customers	Few Customers
Slow Service	Unemployment Agency Traffic Court Hospital Emergency Room Dental Service Tax Preparation	Travel Agency Eye Glass Service Clothes Alteration New Car Dealer Many Repair Services
Fast Service	Dry Cleaning Drop-off Convenience Store Car Wash Parking Garage Toll Booth	Off-peak ATM Self-serve Gas Station Vending Machine Expensive Boutique Check-out

		Museum Admission
--	--	------------------

**Table 4: Expected Waiting Times**

	Many Customers	Few Customers
Slow Service	Very Long Waits	Long Waits
Fast Service	Medium Waits	Short Waits

To determine whether the appropriate strategy involves higher levels of service and shorter waiting times, we must determine whether our customers are willing to pay for decreased waiting times. In industries such as medical services, where insurance companies and government agencies sometimes limit pricing strategies, there may be little incentive to decrease waiting times. In other industries, such as business travel, users may place great value on their time and be very willing to pay for shorter waits.

In many cases, some customers will prefer a higher price for insurance against a long wait while other customers would prefer a lower price and a risk of a long wait. It is, therefore, necessary to segment the market into customers who are willing to risk a wait and customers who are willing to pay more to avoid the risk of a wait. Remember that it is always best to provide sufficient service so that some servers are idle some of the time.

This segmentation can be either explicit or by self-selection. With explicit segmentation, we attempt to identify and classify each customer by his or her willingness-to-wait. Having identified the customer, we then offer that customer the appropriate service. For example, we might announce two services -- a premium service and an economy service. The economy service provides a basic service at a minimum price. The premium service consists of an additional bundle of benefits including insurance against long waits.

### Strategic Implications of Larger Scale

Suppose we have two service providers. The first provider has one server, the service rate is .5 minutes and customers arrive, on average, every 2 minutes. The second service provider has twice the capacity and twice the arrival rate. The second provider has two servers, a service rate of .5 and customers arrive, on average, every minute. In this situation, the first service provider will have 60 percent more customers in line and customers will spend 80% more time in line.

Hence, when the service rate is slow, the penalty imposed on customers is greater than when service rates are higher. This penalty is greater even when the slow service rate provider has additional services to compensate for the slower service rate. An unlucky customer, who arrives shortly after another customer, will need to wait a long time. The customer would need to wait the entire time required for service. The longer the service time, the longer the wait unless there were two servers. With two servers the second customer would not need to wait. As we will see, however, a service with twice the service time and twice the servers still has a greater risk of a long wait than a server with both half the number of servers and half the service time.

Now suppose, there are two servers but customer arrivals occur at twice the rate. Of course, in this case, lines will be longer than with a more infrequent arrival rate. Never the less, as we mentioned, average waits will be less than the situation with only one server but half the arrival rate.

Although it is possible to demonstrate this conclusion with some complex mathematical calculations, we will, instead, examine the intuition underlying the conclusions. Suppose that we have two service providers, A and B. Provider A has one server and gets one customer every two hours, i.e., an arrival rate of .5 customers per hour. Provider B has 2 servers and gets ten customers every two hours, i.e., an arrival rate of 1 customer per hour.

Provider B is similar to two provider A's. For simplicity, suppose that Provider B has two lines, one for each server. Each line would behave similar to provider A. People arrive in each line and wait. The wait in each line at Provider B is, on average, identical to the wait at

provider A. In short, provider B looks completely identical to provider A except there are two lines rather than one.

However, provider B offers an important advantage over provider A. When one line at provider B is busy, a customer can switch to the other line. When one line at provider B gets longer than the other line, customers can switch to the shorter line. In fact, provider B can make the whole process extremely efficient by having one line and, as customers move to the front of that line, assigning them to the first available server.

In other words, as we scale-up the operation, the average wait decreases. Double the arrival rate and add twice as many servers and the wait goes down. Triple the arrival rate and add three times as many servers and the wait goes down further. Increasing the scale decreases the probability of a long wait.

There are, therefore, important economies-of-scale associated with services. These economies will be most important when increasing the size of the operation proportionally increases arrivals. In short, when doubling the size of the operation doubles the market, we should double the size of the operation. The operation should continue to increase in size until further increases in the size of the operation fail to generate sufficiently large increases in arrivals.

The advantage of larger size or scale is particularly important in industries that have both large markets of customers and exhibit very slow service rates. When an industry's service technology requires long service times, service providers in that industry gain an important competitive advantage by growing larger. Larger service providers gain the capacity to provide better service by merely doubling their size. This advantage continues as long as the market is sufficiently large.

New entry, in contrast, is easier in smaller markets with more rapid service. In these markets, a small service provider can serve the entire market or a segment of that market at approximately the same cost as a larger service provider. Holding other factors constant, there is little advantage to scale from the perspective of waiting times.

## ***Profiting from Imbalances in Service Capacity and Demand***

### ***Profiting from Excess Capacity***

#### **Is Excess Capacity Needed?**

As we increase the number of servers, we provide our customers with benefits. We decrease the expected wait times and insure our customers against a long wait. We also incur a higher fixed cost for the capacity that enables us to deliver these benefits. These benefits are desirable for us provided that we are able to increase our price to cover that capacity.

With a constant marketing strategy, it is seldom profitable to have sufficient capacity to cover the maximum possible demand. It depends on the distribution of customer arrivals. Suppose, for example, we must decide on the service capacity of a retailer. Here, we must decide the size of the floor service staff.

We would certainly have additional staff on weekends, holidays and other periods when we anticipate above-average demand. During weekdays, we would have a smaller staff. However, it is possible to have unexpected busy periods each weekday.

Suppose, on average, there are two customers in the store that require service. There is a 5% chance that 5 customers will arrive together and all require service. There is a 1% chance that 10 customers will require service and a .1% chance that 50 customers will require service.

Were we to maintain 50 service employees, 48 employees would usually be idle. The prices for our products must include the salaries of the 50 service employees. Few of our customers would pay the higher prices to insure against not receiving service with a probability of .1%. Hence, we would not employ 50 service employees.

Were we to maintain 5 staff people, 3 would usually be idle. Again, our retail prices must include the salaries of the 3 idle employees as well as the 2 busy employees. Our decision

about the number of staff people depends on our service strategy. We must decide whether to be a high service/high price or lower service/lower price provider

Whatever our strategic decision, we will have excess capacity because we would never hire sufficient service employees to meet the maximum possible, but improbable, demand. It is always possible, with some small probability, that demand will exceed the optimal capacity. Stated another way, optimal capacity always is less than the maximum possible demand.

The consequence, therefore, of our capacity decision is to have periods when demand apparently exceeds capacity. This conclusion, however, may be false when we allow the marketing strategy to vary. We might implement a marketing strategy that reduces demand during periods of insufficient capacity. In general, these marketing strategies involve some type of rationing.

Finally, we should end this section by remembering that we can always eliminate excess capacity by lowering the price of our service. At a sufficiently low price, we would generate sufficient demand to fill capacity. A hotel, for example, at a sufficiently low room rate would fill the rooms with permanent guests. This point illustrates that filling capacity is not a goal and is merely a deceptive distraction for the service provider.

## Selling Excess Capacity

As authors, such as Lovelock (1992) note, off-peak periods with excess capacity provide an opportunity to rest, plan for the peak, train new employees, accomplish required maintenance and develop new services consistent with the overall service strategy. All of these activities provide indirect help that contributes to the efficiency and profitability of the service provider at the peak. They are real benefits that are often difficult to measure.

Beyond these indirect benefits, service providers often seek direct revenue gains from excess capacity. We should again remember that filling excess capacity is not a goal in itself. Just as a manufacturer should not produce products, which the manufacturer is unable to sell, a service provider should not lose money merely to fill capacity.

It would be foolish to switch high-paying peak customers to off-peak by offering deep discounts. We only would be forgoing possible profits. Off-peak periods should not cannibalize peak period profits. We should only try to maximize revenues given the possible resource of excess capacity.

Having said that, let us consider how we can use excess capacity. With excess capacity, we can often produce additional service at very little additional cost. That is, of course, the key to using excess capacity. We must try to accomplish this task by not shifting specific buyers from higher priced purchases at the peak. A restaurant, for example, should not have early-bird dinner specials that merely take customers from the peak. These specials, rather, should target customers who would be unwilling to pay the higher prices at the peak. They may also target customers who we are unable to serve at the peak because of insufficient capacity.

Another approach to profiting from excess capacity is to sell that capacity. A public utility, for example, could sell excess electricity or telephone line capacity. This strategy is feasible when demand is unpredictable and uncorrelated across the industry. Here, when one electric utility has reached capacity, another may be well below capacity and can sell the first utility some of its excess capacity.

To sell excess capacity, it is necessary to establish a spot market. On this market, buyers and sellers can negotiate short-term agreements for excess capacity. In many cases, these spot markets take the form of sub-contracting. An advertising agency, for example, faced with demand that exceeds capacity might subcontract work to an agency operating at under capacity. A hotel, whose demand exceeds capacity, may send guests to another hotel that is operating below capacity.

Remember that when demand is unpredictable, the timing of excess capacity is also unpredictable. Hence, we must develop procedures long before the excess capacity develops. The hotel, for example, that absorbs the overflow from another hotel, must develop a long-term

relationship with the second hotel. The same is true for rental car companies, airlines and other services that either strategically overbook or just find themselves with unexpected demand.

## ***Profiting from Insufficient Capacity***

### **Peak Strategies with Price Flexibility**

Previously, we discussed the pricing decision when demand varied by season but varied in a predictable way. In this case, when demand for the service was certain, the pricing decision was straightforward. As noted earlier, the optimal price is independent of capacity provided we have sufficient capacity at the optimal price. When we have insufficient capacity at the optimal price, we should set a sufficiently high price so that demand exactly equals available capacity. Hence, we must first determine the optimal price without a capacity constraint. Only then can we determine whether capacity is sufficient. Moreover, we should make no attempt to fill excess capacity, but instead, ignore it by merely setting price to maximize profits.

In this section, we turn our attention to the case of uncertain demand. When demand is uncertain, the situation can become more complex. When demand is uncertain, the capacity constraint is probabilistic or uncertain. In short, we do not know whether we have sufficient capacity. Only after setting our price do we observe actual demand. At that point, demand may or may not exceed available capacity. The outcome is uncertain.

When demand is uncertain, the old rules for certain demand may or may not hold. The best pricing rule depends on whether we have the ability to change prices when unexpected peaks in demand develop. That ability is usually specific to the service industry. In industries such as consulting, account, law, architecture, air travel and so on, a service provider can change prices when unexpected demand develops. A consultant who gets unexpectedly busy, for example, can raise prices for new projects. As the consultant reaches capacity, the price can approach infinity and the consultant declines new business regardless of the client's willingness to pay.

When the pricing decision is flexible and we can change our price as unexpected demand conditions develop, the old pricing rules for certain demand apply. We should again ignore capacity and price at that price that maximizes profits. We should only consider capacity constraints when we have insufficient capacity to meet demand at the optimal price. In that case, we should raise the price until demand exactly equals capacity.

### **Peak Strategies with Price Inflexibility**

When price is inflexible, the situation is far more complex. This situation occurs in many industries, such as restaurants, retailing and most personal services. For these industries, we post prices before we know demand conditions. It becomes very difficult, after demand becomes known, to change prices. Retailers can markdown merchandise with disappointing demand, however, they cannot suddenly raise prices on an unexpectedly busy day.

Here, the pricing decision is much more complex because the optimal price always depends on the probability that demand will exceed available capacity. Hence, the optimal price depends on capacity regardless of whether the capacity constraint is binding. The mere possibility of a binding capacity constraint changes the optimal price.

Suppose, for example, that a car wash finds that the best price for a car wash, without considering capacity constraints, is ten dollars. Lowering the price below ten dollars would generate additional customers, but the number of additional customers fails to compensate for the loss in profit per wash. Raising the price above ten dollars would increase the profit per wash, but that increase in profits does not compensate for the decrease in the total number of car washes caused by the higher price.

Now suppose that, at ten dollars, the car wash attracts sufficient demand that long waits can develop. To be precise, twenty-percent of the time, the car wash permanently loses customers who are unwilling to wait in line. So, twenty-percent of the time, the car wash can raise prices and do the same number of washes. Hence, the car wash should raise prices, continue to operate at capacity and earn greater profits.

The car wash, however, is unable to predict when a busy period will develop. Moreover, it is unable to temporarily raise prices when a busy period develops. Hence, the car wash must raise prices during all periods with the anticipation that during busy times, the higher price will generate greater profits. During slack periods, unfortunately, the higher price generates less profit. After balancing price and demand, the best price with a capacity constraint will be higher than the best price without the capacity constraint regardless of whether the capacity constraint is binding.

Remember that this conclusion is only valid when the service provider is unable to change prices when a peak develops. As we have said, the usual cause for this situation is that the peak was unexpected. Other causes for this might include problems with image. Many service providers may feel that their image with the public would suffer if they were to raise prices at times of peak demand (e.g., consider increasing the price of fresh water after a hurricane). The loss in image would suggest a loss in goodwill and an eventual loss in sales. Given these conditions, the service provider would not have a different price during the peak and non-peak. Note, however, that the best price is now greater than the off-peak price and less than the peak price. In short, customers pay more during off-peak periods so that the service provider does not need to raise prices at the peak.

## ***Rationing: Balancing Service Capacity and Demand***

### ***Price Rationing***

For a variety of reasons, many economists argue that rationing capacity on price is the most efficient means of allocating capacity. In short, most economists argue that when demand exceeds available capacity, the capacity should go to the highest bidder. If the demand for a theater exceeds seating capacity, for example, the available seats should go to the highest bidders. Alternatively, the theater could raise the price of each seat until the number of buyers exactly equals the number of seats.

One of the primary justifications for price rationing is supply-side efficiency. As the price of a scarce resource increases, the service provider enjoys more revenue from the available capacity. As that revenue increases, the service provider finds it profitable to expand available capacity. Without the additional revenue, there would be less incentive to expand capacity.

This fact is true for both the service provider and competitive service providers. Were a theater, for example, to consistently find that demand exceeded capacity, the theater might raise its price. With a higher price, the theater would find that additional capacity is more profitable. Competitive theaters would also find that additional capacity is more profitable. As a consequence, the entire industry would build additional capacity. The industry would enjoy increased profits from higher prices. Those profits would encourage existing service providers to increase capacity and increased profits would attract new service providers into the market.

The combined service capacity of all competitors continues to increase until the price drops. This drop in price suggests a balance between the demand for the service and its supply. Were we to allocate capacity on a basis other than price, there would be no incentive to increase capacity to meet demand. The result would be less than optimum levels of capacity.

A second argument for price rationing is that customers reveal their true preferences for the service. For example, consider a theater that is hosting two events, a musical production and a drama. For both events, demand for tickets exceeds the number of available seats. Suppose, rather than raising price, the theater allocates scarce theater tickets for two events using a random lottery. Here, chance determines which customers receive which ticket. In this case, customers who are more interested in the drama may receive a ticket for the musical. Moreover, some customers, who have only a moderate interest in either event, may receive, by chance, tickets to both events.

We see that price rationing is often a desirable means of balancing capacity and demand. It is an efficient way of reducing demand while increasing revenue. It is also an efficient way of allocating capacity in the event that demand still exceeds capacity.

Here, a waiting line or queue develops. We use price rationing to prioritize customers in the queue and determine who receives service first. Those customers willing to pay a higher price move to the front of the queue. More price-sensitive customers risk waiting in line longer for the reward of a lower price. This type of queue is known as a priority queue.

A perfect priority queue would prioritize customers exactly according to their willingness to wait as expressed by their willingness to pay. The customer willing to pay the highest price would take the first position in the queue and be the first to receive service. The customer willing to pay the second highest price would take the second position in the queue. The customers would queue so that the last customer in line bids the lowest price for immediate service.

In reality, we seldom see pure priority queues except in highly technical markets such as queues for computing resources allocations, natural gas delivery services or financial markets. The complexity of the priority queue makes it difficult to implement in its pure form. Far more often, we see random queues with priority rules such as first-come first-serve. Never the less, the priority queue remains an important concept.

A simple means of simulating a priority queue is a useful goal. Commonly practiced methods such as yield management (Desiraju and Shugan 1999; Smith, Leimkuhler and Darrow 1992) create weaker forms of priority queues. These methods fail to perfectly align customers with their willingness to pay. Although they are not perfect, these methods do achieve some degree of success toward that goal.

Yield management, for example, creates a partial priority queue by segmenting customers by their willingness to pay and, then, allocating capacity by segment (Dana 1998, 1999a, 1999b, 2001). The better the yield management system, the more the system resembles a priority queue. An airline, for example, tries to allocate limited seating first by full-priced first class, full-priced business class, full-priced economy class followed by discounted fares. In this way, the airline creates an imperfect, but effective, priority queue.

Yield management also allows overselling of capacity (Biyalogorsky, Carmon, Fruchter and Gerstner 1999). Overselling is useful when demand is uncertain and advance purchases are not completely binding. Overselling also gives the seller the potential to make very profitable buyback transactions during the spot period. When capacity is insufficient, sellers can buyback capacity at lower prices than the expected spot price.

However, recent research suggests that non-price rationing, although inefficient for society, may be more profitable than price rationing (Gilbert and Klemperer 2000). This situation occurs when customers must make large sunk investments to enter the market.

### ***Why Non-price Rationing***

The discussion of rationing is important both in this section, which discusses unpredictable demand peaks, and the last section, which discussed predictable demand peaks. Rationing capacity is important whenever demand for the service exceeds the capacity needed to supply the service. It is less important whether the service provider has the ability to forecast when capacity will be insufficient to satisfy demand.

There is, however, at least one important difference between rationing when demand is unpredictable. The difference comes in the ability to react to changes in demand. In the last section, the service provider knew well in advance when demand would exceed capacity. The service provider had sufficient time to plan for the event and take necessary actions. A tour operator, for example, could announce higher prices during periods of peak demand and lower prices during off-peak periods. An accounting firm could similarly announce higher prices for tax services during the peak season.

When demand is unpredictable, the service provider knows that demand will sometimes exceed capacity. The service provider, however, does not know when that will occur. In this case, the service provider is unable to announce higher prices during periods of peak demand. That inability often leads to forms of non-price rationing.

Consider a taxi service, for example, may suddenly find itself faced with higher demand because of inclement weather. The service would like to increase fares during these unexpected

peak-periods. Unfortunately, the service seldom has the ability to take that action. The taxi service, therefore, resorts to other forms of non-price rationing.

The taxi service may, for example, focus on more profitable routes. Drivers may focus on airports, hotels and locations where long hauls are more likely. The service may also attempt temporarily to expand capacity by postponing breaks for drivers. Calling in additional drivers and deploying all available cars. We later discuss other methods for incorporating temporary capacity through flexible capacity planning.

We see that uncertainty in forecasting peak demand may prevent price rationing. In addition, there are other arguments against price rationing such as perceived equity, simplicity in implementation and legal constraints. Perceived equity, which occurs in services such as health care and public utilities, suggests that customers who are unable to pay a very high price should still be allowed access to the service. Simplicity in implementation, which occurs in industries such as amusement parks and toll roads, suggests that it would be operationally costly to implement complex rationing systems because of difficulties in explaining and maintaining those systems. Finally, legal constraints often prevent many forms of price rationing because of either regulatory rules or laws against price discrimination.

Before discussing forms of non-price rationing, we should again note that higher prices do solve the capacity constraint problem. We can reduce demand to the level of capacity with higher prices. Moreover, higher prices provide the long-term incentive to increase capacity to serve more customers. However, there are equity arguments against price rationing.

Consider the National Development Bank (i.e., Banco Nacional de Fomento, BNF), in Ecuador. This bank had the responsibility for granting credit services to the agricultural sector. The bank, on an equity basis, decided to reject price rationing in favor of non-pricing methods for allocating available credit. Given non-pricing rationing, the demand for credit exceeded available funds. Therefore, the bank used methods such as higher requirements for collateral, a more involved application process and increased delays in granting loans (Morrison 1994). The consequence was a bias in favor of larger loans to customers who could endure both the more onerous application process and the long wait for the loan. This outcome left small farms at a distinct disadvantage, leaving many small farms strapped for funds.

Now consider the health care system. Here, policy-makers rejected price rationing for donor hearts as inequitable. A system of first-come first-served became the policy. Unfortunately, this method of rationing scarce donor hearts has another drawback. The first come-first served method rejects hearts for sicker patients who often find waiting more costly and sometimes die waiting for a heart (Frank 1992).

Obviously, equity concerns are important when choosing a rationing system. However, service providers must face a difficult compromise between equity and gaining additional profits that could help expand future capacity. Higher levels of profit justify both additional investments in capacity as well as maintaining higher capacity off-peak. Service providers who are unable to generate sufficient funds during times of peak demand should choose lower levels of overall capacity.

Finally, non-price rationing might be directed at gaining customer loyalty or rewarding better customers. This concept is consistent with the Zeithaml, Rust and Lemon (2001) concept of a customer pyramid.

### ***Public Attitudes Toward Rationing***

When determining a ration plan, we should consider both the profitability of the rationing plans as well as customer opinion. This is a difficult compromise because customers often prefer less profitable rationing methods. Moreover, many customers are somewhat unhappy about priority queues when they view the service as essential. Fortunately, they do recognize the need for rationing when faced with limited capacity.

One survey, for example, found that many people would support some form of rationing for public health-care services. Forty-four percent of those surveyed, would prefer the same rationing rules regardless of the patient's ability to pay. Forty percent of the respondents said

they preferred rationing of treatments based on the likelihood of success rather than rationing based on first-come first-served. Twenty-nine percent of the respondents said that we should deny smokers health care services diseases caused by smoking. Twenty-six percent of the respondents said that children should have priority over the elderly. Finally, only sixteen percent of the respondents thought that rationing should not be used for privately funded health care services<sup>1</sup>.

### ***Advance Selling with Capacity Constraints***

Advance selling of capacity can be extremely profitable. Shugan and Xie 2000 show that giving buyers the ability to advance purchase capacity provides many potential advantages. If the advance price is discounted, it allows a much larger increase in demand than the same price discount in the spot period (Shugan and Xie 2000, Xie and Shugan 2001). Lee and Ng (2001) also find this result. Shugan and Xie (2002) show that this result survives the introduction of competition.

Advance selling can also allow price-discrimination where more price-sensitive buyers advance purchase allowing high spot prices (Desiraju and Shugan 1999, Dana 1999a). It does, however, often require the sequential servicing of buyers (Rosen and Rosenfield 1997). Advance selling can also help to efficiently adjust uncertain demand to fill more capacity across services (Dana 1999b).

When buyers want to insure future capacity (Png 1989), the advance price can be at a premium to the spot price (Shugan and Xie 2002). Sellers profit from selling at a higher advance price than selling at the lower spot price.

Advance selling can also decrease the pressure to attract new buyers with lower prices (Serel, Dada and Moskowitz 2001). Advance selling impacts the compromise between selling only to current customers in the spot period at high prices and trying to increase demand by attracting new customers. The reason is that by advance selling, spot capacity is reduced and, with less spot capacity, there is less incentive to attract more buyers (Shugan and Xie 2002).

Finally, recent research examines the connection between bundling, rationing and advance selling. DeGraba and Mohammed (1999) suggest that by advance selling in bundles and subsequently selling individually, a multi-service seller can earn more profit than individually selling each service. Given limited capacity and possibly insufficient capacity, less price-sensitive (i.e., high-valuation) customers will advance buy. The reason is that they expect rationing when units are individually sold. It is these bundled purchases that cause the shortages that result in rationing. In this case, the bundle's price exceeds the sum of the individual prices. Future research might further integrate capacity constraints with product line strategy (Noble and Gruca 1999).

### ***Conclusions and Future Research***

This chapter discussed many ways in which capacity constraints impact the marketing strategies for service providers. Although not all service providers face capacity constraints, many do. This section reviews some of the more significant conclusions in this area. We now review eleven important conclusions of this chapter.

First, capacity often dictates service strategy. Second, service providers can only compete on capacity when different buyers want different levels of service. Third, operational adjustments to changing capacity become a marketing strategy because they have a direct impact on customer satisfaction and demand. Fourth, marketing strategies must change during times of peak demand to emphasize services or aspects of services requiring less capacity. Fifth, it is important how we communicate with buyers about changes in price and quality during peak demand. Sixth, an important marketing decision is whether to announce peak prices or retain

---

<sup>1</sup> Northwestern National Life Insurance Company, *Americans Speak out on Health Care Rationing*, Minneapolis: The Company, November 1990.

flexibility about them. Seventh, non-pricing rationing of capacity during times of peak demand creates the opportunity to reward loyal customers and sell new services. Eighth, there is an essential conflict between providing more service per customer and creating waiting times that diminish total service quality. Ninth, price rationing is often the most profitable form of rationing but it does create problems and is often difficult to implement. Tenth, advance selling of services provides enormous potential to increase profits. Eleventh, we can create sophisticated marketing strategies by combining dynamic pricing strategies, non-price rationing, bundling, advance selling and changing service quality over time.

We expect that future research on marketing strategies in the presence of capacity constraints will continue. We expect future research will focus on new strategies made possible by advancing technology. These strategies include more strategies involving more complex pricing, strategies involving advance selling and strategies that continuously adjust over time to adapt to changing conditions (possibly, instantaneously). We need a deeper understanding of how to market bundles of services each facing different capacity constraints. We need a deeper understanding of how to implement multi-service strategies where different services face different competitive and demand conditions. We need a deeper understanding of what are the best ways to implement the rationing of capacity. Finally, we need a deeper understanding of how to build capacity consistent with our service strategy.

## References

- Allen, W. Bruce and Dong Liu (1995), "Service Quality And Motor Carrier Costs: An Empirical Analysis," *The Review of Economics And Statistics*, 77(3), 499-510.
- Anderson, Eugene W. (1995), "An Economic Approach to Understanding How Customer Satisfaction Affects Buyer Perceptions of Value," *Marketing Theory and Applications*, Vol. 6, edited by David W. Stewart and Naufel Vilcassim.
- Anderson, Eugene W. and Claes Fornell (1999), "The Customer Satisfaction Index As A Leading Indicator," *Handbook of Services Marketing and Management*, Dawn Iacobucci and Terri Swartz (eds.), New York: Sage.
- Anderson, Eugene W. and Claes Fornell (2000), "Foundations of the American Customer Satisfaction Index," *Journal of Total Quality Measurement*, 11 (7) S869-S882.
- Anderson, Eugene W. and Steven M. Shugan (1991), "Repositioning for Changing Preferences: The Case of Beef versus Poultry", *Journal of Consumer Research*, 18 (2, September) 219-232.
- Anderson, Eugene W., Claes Fornell, and Donald R. Lehmann (1994), "Customer Satisfaction, Market Share, and Profitability," *Journal of Marketing*, 58 (3, July) 53-66.
- Anderson, Eugene W., Claes Fornell, and Roland Rust (1997), "Customer Satisfaction, Productivity, and Profitability: Differences Between Goods and Services," *Marketing Science*, 16:2, 129-45.
- Bailey, Elizabeth E. and Lawrence J. White (1974), "Reversals in peak and off-peak prices", *Bell Journal of Economics and Management Science*, 5, 75-92.
- Bar-Ilan, A, Sulem AS, and Zanello A (2001), "Time-To-Build and Capacity Choice," *Journal of Economic Dynamics & Control*, 26 (1) 69-98.
- Berg, Sanford V. and John Tschirhart (1988), *Natural Monopoly Regulation*, Cambridge University Press, Cambridge.
- Bergen, Mark, Shantanu Dutta and Steven M. Shugan (1996), "Branded Variants: A Retail Perspective" *Journal of Marketing Research*, Winter, XXXIII, 33 (1) 9-19.
- Berry, Leonard L. (1980), "Services Marketing is Different," *Business Horizons*, (May/June), 24-29.
- Biyalogorsky E., Z. Carmon, G. Fruchter, and E. Gerstner (1999), "Overselling with Opportunistic Cancellations," *Marketing Science*, 18 (4) 605-10.
- Biyalogorsky E, E. Gerstner, B. Libai (2001), "Customer Referral Management: Optimal Reward Programs," *Marketing Science*, 20 (Winter, 1) 82-95
- Bolton, RN (1998), "A Dynamic Model Of The Duration Of The Customer's Relationship With A Continuous Service Provider: The Role Of Satisfaction," *Marketing Science*, 17 (1), 45-65.
- Brock, William A. and Jose A. Scheinkman (1985), "Price Setting Supergames With Capacity Constraints," *The Review of Economic Studies*, 52 (3), 371-382.
- Burness, H. Stuart and Robert H. Patrick (1991), "Peak-Load Pricing with Continuous and Interdependent Demand", *Journal of Regulatory Economics*, 3 (1) 69-88.
- Chu W., E. Gerstner, and J. D. Hess (1998), "Managing Dissatisfaction, How to Decrease Customer Opportunism by Partial Refunds," *Journal of Service Research*, 1 (2) November, 140-154.

- Chu, W., E. Gerstner, and J. D. Hess (1995), "Cost and Benefits of Hard Sell," *Journal of Marketing Research*, 32, (February) 97-102
- Compte O, Jenny F, and Rey P (2002), "Capacity Constraints, Mergers and Collusion," *European Economic Review*, 46 (1), 1-29.
- Crew MA, CS Fernando, PR Kleindorfer (1995), "The Theory Of Peak-Load Pricing - A Survey," *Journal Of Regulatory Economics*, 8 (November, 3) 215-248.
- Dana, James D. Jr. (1998), "Advance-Purchase Discounts And Price Discrimination In Competitive Markets," *The Journal of Political Economy*, 106 (2), 395-422.
- Dana, James D. Jr. (1999a), "Equilibrium Price Dispersion Under Demand Uncertainty: The Roles Of Costly Capacity And Market Structure." *Rand Journal of Economics*, 30, 632-660.
- Dana, James D. Jr. (1999b), "Using Yield Management To Shift Demand When The Peak Time Is Unknown", *The Rand Journal of Economics*, 30 (3), 456-474.
- Dana, James D. Jr. (2001), "Competition In Price And Availability When Availability Is Unobservable", *Rand Journal of Economics*, 32 (3) 497-513.
- DeGraba, P and R Mohammed (1999), "Intertemporal mixed bundling and buying frenzies," *Rand Journal Of Economics*," 30 (Winter 4) 694-718
- Desiraju, Ramarao and Steven M. Shugan (1999), "Strategic Service Pricing and Yield Management," *Journal of Marketing*, 63 (January), 44-56.
- Folland, Sherman, James T. Ziegenfuss, Jr. and Paul Chao (1988), "Implications of Prospective Payment Under DRGs for Hospital Marketing," *Journal of Health Care Marketing*, 8 (December), 29-36.
- Frank, Jacqueline (1992), "Many On Heart Waiting List May Not Need Transplants Study," *Reuters News Service*, (November 19).
- Gilbert RJ and P Klemperer P (2000),"An equilibrium theory of rationing," *Rand Journal Of Economics*, 31 (Spring, 1) 1-21
- Greenidge, C. D. (1983), "Let Customers Go With The Flow; Analyzing Store Layout," *Skiing Trade Monthly News*, 7 (January) 10
- Hart, C., J. Heskett, and W. E. Sasser Jr. (1990), *The Service Management Course*. N.Y.: Free Press.
- Haskel, Jonathan and Christopher Martin (1994), "Capacity And Competition: Empirical Evidence On UK Panel Data," *Journal of Industrial Economics*, 42 (1), 23-44.
- Hauser JR, DI Simester, B Wernerfelt (1994), "Customer Satisfaction Incentives," *Marketing Science*, 13 (4) 327-350
- Hauser, JR (2001), "Metrics Thermostat," *Journal Of Product Innovation Management*, 18 (May 3) 134-153.
- Herk, Leonard F. (1993), "Consumer Choice and Cournot Behavior in Capacity-Constrained Duopoly Competition," *The RAND Journal of Economics*, 24(3), 399-417.
- Hess J. D. and E. Gerstner (1998), "Yes, Bait and Switch Really Benefits Consumers," *Marketing Science*, 17 (3) 273-282.
- Hirschberg, JG (2000), "Modelling Time Of Day Substitution Using The Second Moments Of Demand," *Applied Economics*, 32 (June, 8) 979-986
- Hoffman, K.D. and Bateson, J.E.G. (1997). *Essentials of Services Marketing*, Orlando, Florida: The Dryden Press

- Kurtz, David L. and Kenneth E. Clow (1997), *Services Marketing*, Wiley, USA.
- Lee KS and Ng ICL (2001), "Advanced Sale Of Service Capacities: A Theoretical Analysis Of The Impact Of Price Sensitivity On Pricing And Capacity Allocations," *Journal Of Business Research*, 54 (3) 219-225.
- Lovelock, C (1992), "Seeking Synergy in Service Operations: Seven Things Marketers Need to Know About Service Operations," *European Management Journal*, 10 (March, 1) 22-29.
- Lovelock, C., P.G. Patterson and R. Walker (1998), *Services Marketing*, Prentice-Hall, Sydney.
- Morrison, Andrew R. (1994), "Capital market imperfections, labor market disequilibrium and migration: a theoretical and empirical analysis," *Economic Inquiry*, 32 (2, April) 290.
- Noble P.M. and T.S. Gruca (1999), "Industrial Pricing: Theory And Managerial Practice," *Marketing Science*, 18 (3) 435-454.
- Onkvisit S. and Shaw J.J. (1991), "Is Services Marketing "Really" Different?", *Journal of Professional Services Marketing*, 7 (2) 3-17.
- Parasuraman, A., Zeithaml, V. A., and Berry, L. (1985), "A Conceptual Model of Service Quality and its Implications for Future Research," *Journal of Marketing*, 49 (Fall) 41-50.
- Perrakism, Stylianos and George Warskett (1983), "Capacity And Entry Under Demand Uncertainty," *The Review of Economic Studies*, 50(3), 495-511.
- Png IPL (1989), "Reservations - Customer Insurance In The Marketing Of Capacity," *Marketing Science* 8 (Summer, 3) 248-264.
- Radas, Sonja and Steven M. Shugan (1998), "Managing Service Demand: Shifting and Bundling," *Journal of Service Research*, 1 (1, August) 47-64.
- Radas, Sonja and Steven M. Shugan (1998), "Seasonal Marketing and Timing New Product Introductions," *Journal of Marketing Research*, 35 (3, August) 296-315.
- Rose, Judah and Mann, Charles (1995) "Unbundling The Electric Capacity Price In A Deregulated Commodity," *Public Utilities Fortnightly*, 133 (22, December).
- Rosen S, AM Rosenfield (1997), "Ticket Pricing," *Journal Of Law & Economics*, 40 (October, 2): 351-376
- Rust, R., A. Zahorik, and T. Keiningham, *Service Marketing*, Harper Collins, New York, 1996.
- Sawhney MS and J. Eliashberg (1996), "A Parsimonious Model For Forecasting Gross Box-Office Revenues Of Motion Pictures," *Marketing Science*, 15 (2) 113-131.
- Serel, Dogan A, Maqbool Dada and Herbert Moskowitz (2001), "Sourcing Decisions With Capacity Reservation Contracts," *European Journal of Operational Research*, 131 (3) 635-648
- Shmuel Oren, Stephen Smith and Robert Wilson (1985), "Capacity Pricing," *Econometrica*, 53(3), 545-566.
- Shostack G. Lynn (1977), "Breaking Free from Product Marketing", *Journal of Marketing*, 41 (April) 73-80.
- Shostack, G. Lynn (1984) "Old Ways Of Selling Just Won't Do In New World Of Services," *The American Banker*, Marketing Management, (August 8).
- Shugan, Steven M. (1994), "Explanations for Service Growth", in *Service Quality*, Richard Oliver and Roland Rust, Ed., Sage Publications, 223-240.

- Shugan, Steven M. and Jinhong Xie (2000), "Advance Pricing of Services and Other Implications of Separating Purchase and Consumption," *Journal of Service Research*, February, 2 (February) 227-239.
- Shugan, Steven M. and Jinhong Xie (2002), "Advance-Selling Strategies with Competition," University of Florida working paper.
- Shugan, Steven M. and Ramarao Desiraju (2001), "Retail Product-line Pricing Strategy when Costs and Products Change", *Journal of Retailing*, Spring, 77 (1) 17-38.
- Shugan, Steven M. and Sonja Radas (1999), "Services and Seasonal Demand," in *Handbook of Services Marketing and Management*, Teresa A. Swartz, Dawn Iacobucci (Eds.), Sage Publications, 147-170.
- Sirdeshmukh, Deepak, Jagdip Singh, and Barry Sabol (2002), "Consumer Trust, Value, and Loyalty in Relational Exchanges," *Journal Of Marketing*, 66 (Winter, 1), 15
- Skiera, Bernd and Martin Spann (1999) "The Ability to Compensate for Suboptimal Capacity Decisions by Optimal Pricing Decisions", *European Journal of Operational Research*, 118, 450-463.
- Smith BC, JF Leimkuhler, RM Darrow (1992), "Yield Management At American-Airlines," *Interfaces*, 22 (1) 8-31
- Spence, A. Michael Entry (1977), "Capacity, Investment And Oligopolistic Pricing," *The Bell Journal of Economics*, 8 (2), 534-544.
- Spicer, John and Dan Bernhardt (1997), "Durable Services Monopolists Do Better Than Durable Goods Monopolists," *The Canadian Journal of Economics*, 30 (4a), 975-990.
- Szymanski, David M. and David H. Henard (2001), "Customer Satisfaction: A Meta-analysis of the Empirical Evidence," *Journal of the Academy of Marketing Science*, 29 (Winter), 16 - 35.
- Valarie A. Zeithaml and Mary Jo Bitner (2000), *Services Marketing: Integrating Customer Focus Across the Firm*, Second Edition, New York, NY: McGraw-Hill Companies.
- Van Cayseele P and Furth D (2001), "Two Is Not Too Many For Monopoly," *Journal of Economics*, 74 (3) 231-258.
- Wernerfelt, B (1995), "The Resource-Based View Of The Firm - 10 Years After," *Strategic Management Journal*, 16 (March, 3) 171-174.
- Xie, Jinhong and Steven M. Shugan (2001), "Electronic Tickets, Smart Cards, and Online Prepayments: When and How to Advance Sell," *Marketing Science*, 20 (3) 219-243.
- Zeithaml VA, RT Rust and KN Lemon (2001) , "The Customer Pyramid: Creating And Serving Profitable Customers," *California Management Review*, 43 (Summer, 4) 118
- Zeithaml, Valarie A. (2000), "Service Quality, Profitability and the Economic Worth of Customers: What We Know and What We Need to Learn," *Journal of the Academy of Marketing Science*, 28 (1, Winter) 67-85.