Pergamon

European Management Journal Vol. 15, No. 5, pp. 523–536, 1997

0263-2373/97 $17.00 + 0.00

PII: S0263-2373(97)00032-7

Managing the Tug-of-War Between Supply and Demand in the Service Industries

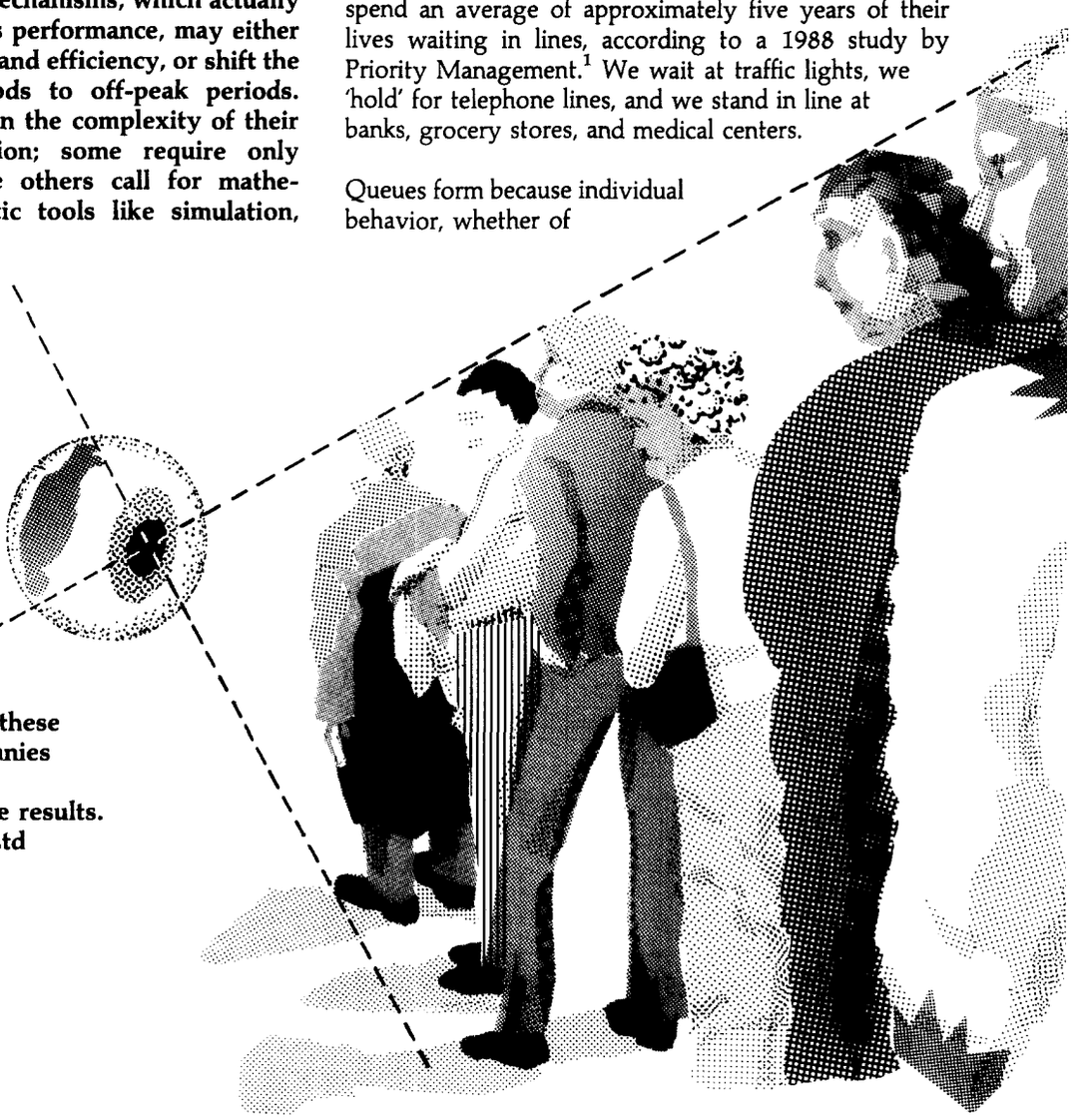
GABRIEL BITRAN, *MIT Sloan School of Management, Boston*
SUSANA MONDSCHEIN, *University of Chile*

In a service organization, a number of mechanisms may be used to match a limited supply of services with an unpredictable demand for those services. Tactical and operational mechanisms, which actually enhance the organization's performance, may either increase absolute capacity and efficiency, or shift the demand from peak periods to off-peak periods. These mechanisms differ in the complexity of their design and implementation; some require only qualitative analysis while others call for mathematical models or analytic tools like simulation, queuing theory, and mathematical programming. Finally, perceptual mechanisms, which alter only the customer's perceptions of the organization's performance, may also be used to maintain customer satisfaction when delays in service are unavoidable. Most service firms will want to use a mix of these mechanisms; airline companies have used many different mechanisms with favorable results.

I Introduction

Queues are almost unavoidable in our lives. People spend an average of approximately five years of their lives waiting in lines, according to a 1988 study by Priority Management.[1] We wait at traffic lights, we 'hold' for telephone lines, and we stand in line at banks, grocery stores, and medical centers.

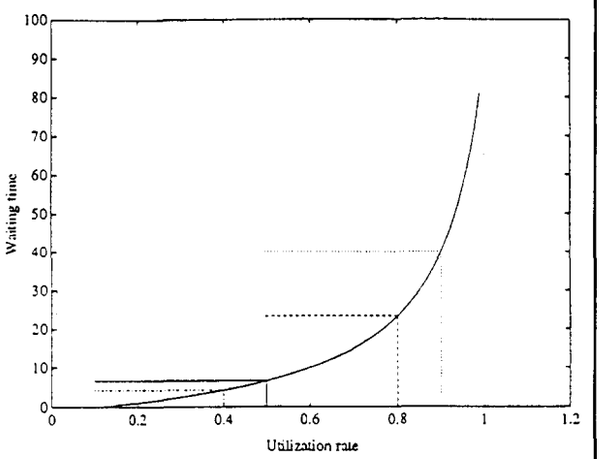
Queues form because individual behavior, whether of



humans or of machines, is unpredictable. For example, many customers may arrive simultaneously at a service facility, or a crucial piece of equipment may break down; either may result in a long wait for service. Some of the uncertainties that produce queues can be more effectively managed and minimized than others. Preventive maintenance can significantly reduce equipment malfunctions, but unevenness in the flow of customers or service requests can be predicted only crudely, if at all.

For service-oriented businesses, which include the service industries — such as transportation, travel and lodgings, food, communications, entertainment, and personal services, including health care — as well as almost any type of retailing, the ability to match the available supply with the current demand can be a major determinant of success. Fortunately, a variety of management techniques are available for solving such problems. Figures 1, 2, and 3 give a graphical illustration of the important quantitative relationships in queue formation.

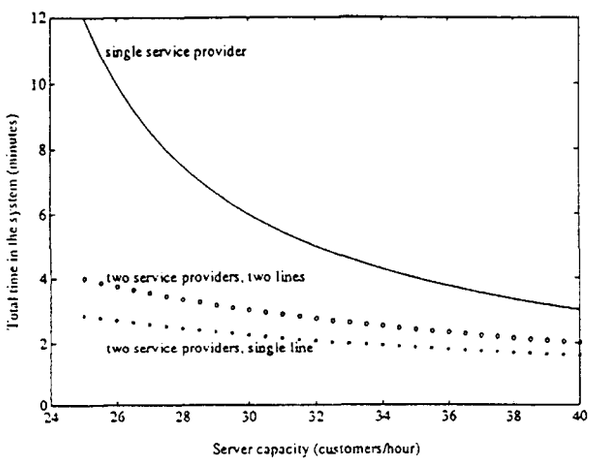
Strictly speaking, queues are formed when unpredictable demands give rise to conflicts over the use of available resources. Moreover, the random fluctuations of both the arrival times and the sizes of the demands will cause queuing to occur even if, on average, the demands do not exceed the capacity of the service facility. Consider, for example, a single facility that receives an average of 10 calls per hour, each of three minutes duration; if four calls are received during a five-minute period, then a queue of waiting calls will result.



The utilization rate is defined as the ratio between the average number of arrivals per unit time and the average capacity of the system per unit time. An interesting relationship can be observed between the utilization rate of the system and the average waiting time: if the utilization rate increases, the effect on the average waiting time depends largely on the current utilization. In other words, the higher the utilization rate, the larger the increase in average waiting time. Thus facilities that operate close to their capacity are more unstable; a computer breakdown or an employee's absence can cause great congestion. On the other hand, in facilities with low utilization rates, the resulting delays may be negligible. Clearly there is a trade-off between increasing the efficiency of the system and delivering consistently high-quality service, as measured by waiting times.

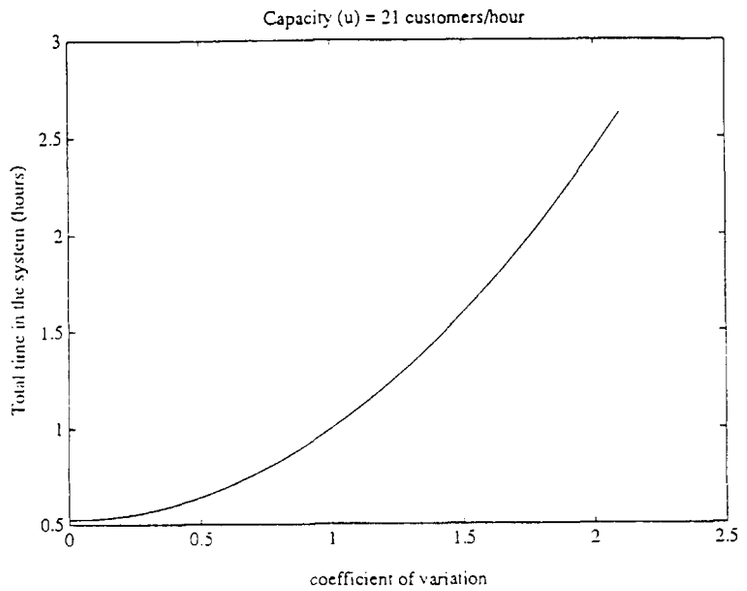
Figure 1 Average Waiting Time as a Function of the Utilization Rate

Two main mechanisms have been used to reduce waiting time (and idle capacity), and therefore to increase the quality of delivered services: (i) to increase the capacity of the service-providing system and (ii) to shift demand from peak to off-peak periods. The tools available to operations managers for this purpose range from the traditional ones of pricing and promotional techniques that smooth the peaks and valleys of



The basic relationship between the utilization rate of the system and the average waiting time has important implications for the design of service facilities. For example, this figure shows the total time per call (waiting time plus service time) for different service provider capacities, in a service facility that handles an arrival rate of 20 calls per hour (we consider all exponential interarrival time and an exponential service time). The service systems modelled consist of: a single service provider; two service providers and two separate waiting lines (one for each service provider); and two service providers and a single waiting line (calls are answered for the first available service provider). The efficiency of the facility increases dramatically when an additional service provider is available. For example, using a single service provider with a capacity of 25 customers per hour results in a total time per call of 12 minutes. However, the total time is reduced to less than 3 minutes when an extra service provider is added to the system, assuming a single waiting line in both cases. On the other hand, if our goal is to have an average total call time of less than 10 minutes, then we can either add a second service provider or increase the productivity of the first one. In the latter case, the service provider should reduce the average service time from 2.4 minutes to 2.26 minutes. We also observe from Figure 2 that it is more efficient to have a single queue, assuming that all customers belong to a single market segment and therefore request similar services.

Figure 2 Relationship Between Total Time Per Call and Modelled Different Service Provider Capacities



The total time, in a single-service provider system, varies according to the coefficient of variation of the service time. The coefficient of variation is a measure of the variability of the service time with respect to its mean; it is defined as the quotient of the standard deviation of the service time and the average service time. The average service time remains unchanged. Observe that the total average time increases considerably when there is more variability in the customers' service times. In such situations, it is possible to reduce the waiting time without increasing the number of service providers by, for example, segmenting the customers according to their service times, say by establishing different queues for higher and lower numbers of transactions per customer.

Figure 3 Relationship Between Total Time in a Single-service Provider System and the Coefficient of Variation of the Service Time

demand over time; through new technologies, especially automation and information systems, that improve efficiency and add value to the service; to quantitative facilities design and planning methods, such as yield management with its sophisticated mathematical underpinnings. However, even when a system's capacity has been increased to optimal levels and demand has been shaped efficiently, people may still have to wait. Under these circumstances, managers may fall back on 'perception' management techniques that allay the anxieties and dissatisfactions of customers waiting to be served.

The goal of this paper is to study these mechanisms, analyzing the tools that are available to managers at tactical and operational decision levels. For this purpose, we present an extensive description of these mechanisms based on cases taken from real-life and reported in academic literature. Our analysis presents a unified view of this topic, enhancing the existing literature with new mechanisms that have been developed or implemented only recently in the service industry, such as yield management and sophisticated information systems.

II The Distinctive Attributes of Service Operations

Service-oriented businesses, while similar to manufacturing businesses in many respects, produce 'projects' with attributes that are markedly different than those of the manufacturing sector.[2]

In the service industries, the provision and delivery of the product occur simultaneously; in other words, services are 'consumed' the instant they are 'produced'. Customer service representatives take telephone orders when clients want to buy. Medical doctors and nurses must be available when emergencies occur.

The intangibility of services means that they cannot be inventoried as tangible goods are. Thus service managers are deprived of an important buffer that their counterpart in manufacturing firms may use to withstand fluctuations in demand. This lack of inventory has two main consequences. First, service facilities may be idle for long periods: city hotels have low occupancy rates on weekends, telephone operators receive few calls after midnight, and air conditioner technicians are rarely called during winter months. Second, large queues may build up at peak times; in that circumstance, the consequences of insufficient capacity are dissatisfied customers and poor perceptions of service quality, which in turn can lead to loss of profits. This phenomenon can be observed in technical support centers at computer companies during normal business hours, waiting lists at restaurants on Saturday nights, and crowded retail stores during the Christmas season.

Managing supply and demand, then, must be a key task of service managers. As stated by Sasser,[3] 'when service managers plan, rather than react, they can successfully fit their capacity to the demand for their products.' Lovelock[4] complements this idea by pointing out that

even where the fluctuations (in demand) are sharp, and inventories cannot be used to act as a buffer between supply and demand, it may still be possible to manage capacity in a service business ... But, for a substantial group of service organizations, successfully managing demand fluctuations through marketing actions is the key to profitability.

Table 1 Mechanisms for Matching Supply and Demand

Tactical level	Location	Pre-selling
	Modular facility design	Direct marketing
	Sharing capacity	Price differentiation
	Technology	Information to customers
	Information systems	Promotion and sales
	Standardization	Complementary services
	Floating staff and part-time employees	Preventive maintenance of users' equipment
	Cross-training	
	Extending business hours	
	Preprocessing	
	Preventive maintenance	
	Cooperation with competitors	
	Complementary services	
Operational level	Downgrading of products	Pricing
	Overbooking	Loss leaders
	Service length	
	Scheduling	
	Engaging customers	
	Batching the service delivery	

III Mechanisms for Matching Supply and Demand

The manager's goal in a service organization is to keep customers satisfied at a reasonable cost. This goal can be achieved by first improving the organization so that better service is provided within acceptable cost constraints, and then, if possible, improving the customer's perception of the service. In other words, it is possible both to *do* better and to *seem* better.

(A) Operations Management: How to Do Better

Two main dynamics can be used to match a limited supply of services with an unpredictable demand, and thereby to enhance operational efficiency. The first dynamic calls for managing capacity, by means of increasing the efficiency of service operations or the actual capacity of the service facility. The second depends on managing demand, notably by shifting demand from peak periods to off-peak periods.

Although strategic decisions can directly affect service delivery, we are interested in exploring the range of tactical mechanisms and operational decisions available for use by managers within any given strategic framework. Therefore, we study the dynamic responses that service companies can take in the medium and short terms to respond to changes in demand patterns. Table 1 summarizes the responses that are described below.

The applicability of most of these responses relies on the company's ability to segment its market in homogeneous groups. Market segmentation allows companies to invest their limited resources more effectively. For example, the Parker House hotel, had, at some point, segmented its customers into more than 11 categories, which allowed the sales division to direct its marketing efforts selectively according to the known demand profile. Thus, mini-vacation packages and bus tours were promoted to tourists for weekends and for July and August, when the demand from corporate customers was low. 'Consider New England during the middle of October,' explained one manager. 'For us success at that time is to have 100 percent walk-in transient business at the rack rate — and to have raised the rate the day before. It wouldn't be in our best interest to have booked a group far in advance at a very low rate when we know we're going to get excellent, high-rate transient business at this time of the year.[5]

On the other hand, if the service company does not segment its market, it must satisfy the needs of heterogeneous customers with a single set of services; under those circumstances, the best outcome that could be expected would be an average performance, which is not sufficient to win in the long run. The trend that is now observed is the change from mass markets to micro-markets, even 'markets of one.' In fact, in the automobile industry, manufacturers sometimes produce as few as 20,000 units of certain models of car.

The responses differ in the complexity of their design and implementation. Some, for example, cross-training, engaging customers in the delivery of the service, and preprocessing, usually require only qualitative analysis for their design. Others require more sophisticated methodology. For example, the application of yield management techniques (see Box 1) requires mathematical models to determine the appropriate pricing structures. A similar analysis is sometimes required when extending hours of service or scheduling employees. Analytic tools like simulation, queuing theory, or mathematical programming can be very useful in these decision processes.

1. Managing Capacity

(a) **Tactical mechanisms**

There are a number of medium-term decisions that are used to manage the supply of services. We describe below examples of tactical mechanisms that have been used to introduce flexibility in the supply of services or to increase the efficiency of the service facilities.

(1) **Location: mobile or distributed services**: One of the most important decisions that service managers must make is where to locate the service facility. 'Flexibility' of location can increase the utilization of the service facility, by allocating dynamically the supply of services closer to the potential demand. This mechanism is useful when the demand varies geographically over a period of time. For example, hospitals distribute their ambulances, according to the historical demand patterns, in different sites of the city in order to be closer to their potential customers. Food trucks and moving libraries are other examples where a flexible location leads to higher utilization rates.

(2) **Modular facility design**: Analogous to 'flexibility' in the capacity of the service facility. During off-peak hours, part of the facility can be closed, thereby reducing operational costs. Modular design is effective in supermarkets, post offices, and retail stores, where only a fraction of clerks or cashiers is needed at off-peak times. However, the facility must be designed to avoid damaging customers' perceptions of service quality. The Vice President, Employee Relations of the United States Postal Service noted that patrons complained when they saw postal employees who were at their work stations but performing other tasks instead of helping the patrons who were waiting for service.[6] In a well-designed facility, customers' will not be drawn to the temporarily unused capacity. Phone-based services, where a varying number of reps can be scheduled according to the daily demand profiles, are therefore particularly well-suited to modular facilities.

(3) **Sharing equipment**: This is useful when there is expensive, underutilized equipment, which can be used at different times by different firms. Redbanc, the Chilean network of teller machines, is an example. Banks have to install teller machines over a vast geographical area, but they are very expensive to install and maintain. Thus, Chilean banks have created Redbanc, where each bank can install any number of tellers provided that these can be used by competitors' customers. The bank that owns the machine receives a fee proportional to the number of transactions conducted by its competitors.

(4) **Technology to save time**: Technology can significantly increase the efficiency of the service facility and therefore enlarge the actual capacity of the system. For example, banks are using technology to make their front-office workers much more efficient. Cleveland's Society National Bank has automated routine customer-service work so that 70% of phone calls are handled through a voice mail system. This has freed reps to help customers who really need assistance.[7] Technology has also helped Frank's Nursery and Crafts Inc. to reduce long checkout lines that resulted from obtaining credit card authorizations via telephone calls, averaging 45 seconds each, to a credit-reporting agency. Instead, the company installed a $4 million satellite system that connected its stores directly with Visa USA Inc. and reduced the authorization time to only seven seconds.[8] Federal Express has also incorporated automation in many steps of its service delivery process. As stated in Zemke:[9]

Waiting at airports around the country is Federal Express' fleet of more than sixteen thousand computer- and radio-dispatched delivery vans, each one primed to handle the day's incoming deliveries by 10:30 a.m. Using sophisticated scanners, computer terminals, and on-line information systems, Federal Express drivers can manage their time and routes efficiently and accurately. One by-product of the computers: drivers can be made aware of overdue accounts and special delivery conditions. The same system allows customers to find out where their packages are within thirty minutes. If FedEx can't find it in thirty minutes, it's delivered free.

(5) **Information systems to add value and increase efficiency**: Information technology has been one of Federal Express' most important competitive advantages. The main value of customer service systems generally lies in: (i) preprocessing, which increases the efficiency of a service facility by speeding up access to all information that is required to deliver the service, for example customers' names, addresses, phone numbers, and related service preferences and characteristics; and (ii) point-of-service analyses, which boost sales revenue by allowing quick analysis of, for example, the customer's past history.

FedEx's COSMOS (Customer, Operations, Service, Master On-line System) consists of: an order-entry system for customers to request package pickups; a continuously updated record of each package's progress; financial records for billing purposes; and a relational database of customer transactions. This powerful system, installed in 1979 and continuously updated to increase its capability, allows the company to speed up order processing, to segment it markets, and to deliver better quality service. As a FedEx executive explains, 'Examining the order blank that appeared on the video terminals used by customer service agents, COSMOS could be programmed to identify high-volume customers,' information that could be used to give special treatment to those customers.[10]

(6) **Standardization; do it *their* way**: Companies that sell only a few highly-standardized products have an advantage when it comes to increasing the efficiency of their service facilities. This is because, first, employees in general are more productive when their time is allocated among fewer tasks. Second, standardization allows more accurate demand forecasting, which is particularly important in service industries since traditional inven-

tories are not possible. Fast food chains are a good example of service standardization. McDonald's food preparation and delivery system is based on assembly-line techniques: cook a dozen burger patties at a time, garnish them identically, wrap them up, and keep them in a warming bin so customers can walk up to the counter or slide through the drive-through and leave fast. Customers understand that the food provider will also let you have it 'your way', but you have to wait. In the food service business, the 'priority' customers are the ones who request the standard product.

(7) **Floating staff and part-time employees; more people**: Adjusting staffing levels to accommodate peak demand is a useful alternative for some businesses. The cycle of demand peaks varies by business type, and may correspond to certain hours of the day (telephone companies, mass transit system), certain days of the week (restaurants), certain weeks of the month (banks) or certain months of the year (resorts, Christmas sales). L.L. Bean, for example, employs approximately 65% more people during the Christmas season, which for that company starts in the Fall.

(8) **Cross-training; more skilled people**: Cross-training, which permits one employee to perform more than one job, increases worker flexibility. Cross-trained employees can be switched to bottleneck tasks during peak times, and then switched back to their regular activities at off-peak times. At Domino's Pizza, where 80% of all orders are placed during 20% of the store's business hours, most employees can perform more than one of five crucial jobs: driver, order-taker, pizza-maker, oven-tender, and router (a router determines the most efficient routes for delivery drivers to follow).

(9) **Extending or redistributing business hours; more time**: To meet growing demand without expanding physical capacity, businesses can adjust their working hours. It is an attractive alternative when increases in demand are temporary. Extending working hours is also an alternative to capture customers who are not able to access the service during regular business hours. Frequently, companies adjust their schedules for their convenience, and not their customers'. Lovelock[11] mentions five main factors that are driving the move toward extended hours. These are (i) economic pressure from consumers, (ii) changes in legislation, (iii) economic incentives to improve asset utilization, (iv) availability of employees for 'unsocial' hours, and (v) growth of automated self-service facilities. A growing number of businesses are using this mechanism; for example, supermarkets, banks through automated teller machines, retail stores, and customer support services. An example of extended services is offered by a Chilean bank (Banco de Crédito e Inversiones), which provides information, consultation, and services that are available world-wide, around-the-clock, via the Internet.

(10) **Preprocessing**: The nature of some services allows the execution of certain tasks before the service is actually delivered. These preprocessed parts become a buffer to face peak times. The feasibility of preprocessing is highly correlated with the degree of standardization of products and processes. Many service companies use the preprocessing of information to shorten the service time. For example, every time customers call Pizza Hut to place an order, their telephone numbers are the only information required for complete identification, including of their previous order. Similar examples are found in hospitals, courier services, and repair companies.

(11) **Preventive maintenance of equipment in facilities**: Preventive maintenance during off-peak demand periods is an indirect method for increasing the capacity of the system during peak demand, simply by reducing the number of break-downs.

(12) **Cooperation with other competitors; cost-averaging**: It is common to find situations in the service industry where competitors make agreements to satisfy the excess demand faced by one of the companies. In this process both parties gain: one increases the profit and the other decreases the tangible and intangible costs of turning down customers. For example, managers with overbooked hotels usually send customers, whose requests cannot be served, to other hotels in the city. As a manager of the Parker House hotel pointed out, 'when guests were turned away because of overbooking, they were referred to Boston's best hotels, such as the Ritz, the Copley Plaza, and the Hyatt Regency, the latter across the river in Cambridge. More price-sensitive guests were directed to middle-rank hotels or motor lodges.'

(13) **Complementary services**: This mechanism, especially suitable for highly seasonal services, permits two or more products to be offered at different times of the year in order to establish a more homogeneous demand. For example, equipment maintenance companies may offer air conditioning services during summer and heating services during winter.

(b) Operational Decisions

The following are examples of mechanisms usually used to manage the capacity of the system in the short run to respond effectively to variations in demand.

(1) **Downgrading of products (or equivalently, upgrading of customers)**: This is the case when there exists a natural ordering of all products, so that every service is an acceptable substitute for those that are 'worse' than it. A good example of this situation can be found in the hotel industry, where suites can be substituted for deluxe and standard rooms, and deluxe rooms are an acceptable substitute for standard rooms. This feature adds flexibility to the definition of capacity where the availability of an individual product can be seen as the aggregated capacity of all products that have the same or a higher quality than it. Bitran and Mondschein[12] developed a methodology to efficiently rent hotels rooms to different classes of customers

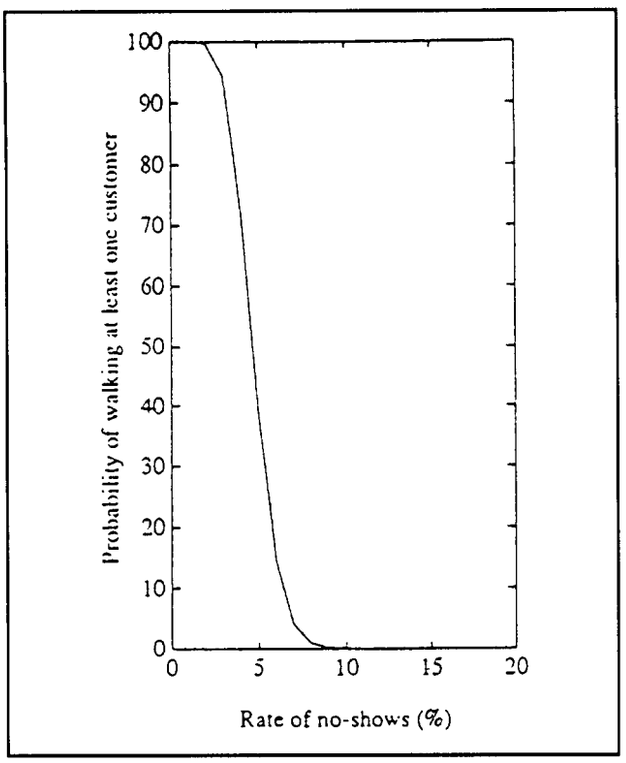


Figure 4 A Model of Motel Over-booking

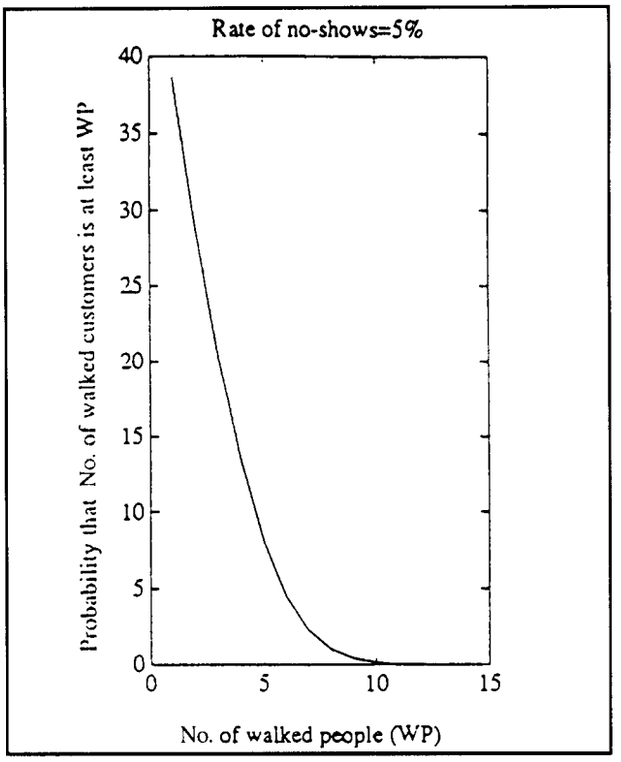


Figure 5 Probability of Turning Away at least a Given Number of Customers for a No-show Rate of 5 per cent

considering the possibility of downgrading. Using real data from a medium-size hotel, it is possible to obtain increases of approximately 15% in profits when using this mechanism.

(2) **Overbooking**: When reservations are allowed and the show-up rate is less than 100% overbooking is a viable mechanism for increasing the actual utilization. Service managers who deal with overbooking face trade-offs between addition costs: the cost of turning away the overbooked customers who cannot be served, and the opportunity costs of idle capacity. Usually overbooking may be used in conjunction with other mechanisms, such as customer compensation or downgrading of products, in order to decrease its associated costs.

Consider a motel with 300 rooms that allow overbooking of 15 rooms (5%). Figure 4 shows the probability, as a percentage, of having to turn away at least one customer who holds a reservation, for different rates of no-shows. For example, if the no-show rate is 5% (i.e., on average, 5% of customers do not show up) then with probability 0.39 the hotel will have to reject at least one customer. However, if the no-show rate increases to 10%, which is fairly realistic in the hotel industry, then the probability of having to turn away a customer is almost zero. Therefore, in this case there is no doubt that to overbook the hotel is the best decision to maximize profits. Figure 5 shows the probability of turning away at least a given number of customers for a no-show rate of 5%. We observe, for example, that although the probability of turning away at least one customer is approximately 0.39, the probability of turning away more than five decreases to less than 0.05.

(3) **Service length**: In many businesses, service providers can control, to some extent, the duration of the services rendered. Thus, services can be curtailed when there are too many customers waiting, or extended to maintain higher utilization of the facility. Benihana, a chain of Japanese restaurants, has mastered the art of managing the duration of service according to demand. In these restaurants dinner is delivered as a 'show', which the chefs serving the dinner control, according to the number of customers waiting in the cocktail lounge. Hairdressers, restaurants, and sight-seeing tours are other examples of service companies which can adjust the length of their services according to the work load.

(4) **Dynamic scheduling; 'intelligent' scheduling**: Flexible scheduling provides an important tool to manage the capacity of the system. For instance, it is crucial for a delivery company to have a dynamic scheduling system to program efficiently the collection and delivery of packages. Federal Express dispatchers schedule their trucks in order to handle the day's incoming deliveries by 10:30 a.m. Emergency systems also use dynamic scheduling as a key element for matching the fixed capacity (ambulances, trucks, etc.) with the uncertain demand. For example, Chilectra, the Chilean electrical utility, classifies emergencies in five categories according to priority and then dynamically schedules the emergency trucks to minimize the total expected time that they spend moving from one emergency site to another. The software developed for this purpose uses operations research techniques such as network design, routing algorithms, and heuristics.

(5) **Engaging the customers in the delivery of the service**: Rafaeli[13] observes that this is the most proactive strategy a service employee can adopt. An experienced cashier at a supermarket says, 'I simply tell the customer what to do. They do not mind it because things seem to move quicker. And it helps me to do my work.' Another alternative for engaging customers in service delivery is to ask them to fill out necessary paperwork while they are waiting to be served. For example, travelers usually must fill out forms when they are standing in line to be served by international police at airports, and patients must fill out medical forms before seeing their doctors.

(6) **Batching the service delivery**: 'Batching' services is an excellent way to increase efficiency and flexibility when the service can be delivered simultaneously to a group of customers. Moreover, some service companies have the flexibility to increase the batch size to respond to surges in demand. For example, taxis at airports can serve several customers at the same time during some 'emergencies'. Other examples of batch processing that take advantage of the clear economies of scale include tours of museums or historic sites, which do not start until a sufficiently large group has gathered, and transportation to airports, where buses pick up people at different places before departing.

2. *Managing Demand*

(a) **Tactical mechanisms**

Several medium-term mechanisms attempt to change demand patterns by modifying the behavior of customers. The success of these techniques relies strongly on demand elasticity; thus a price differential will have an impact on the demand pattern only if customers have the flexibility of requesting the service at different moments in time. We describe below some of the most frequently used mechanisms.

(1) **Pre-selling**: Pre-selling the productive capacity of a service facility can be seen as building up 'inventories of customers'. Excess demand at one time period can be shifted by moving it to another period. Reservation systems, which reduce some of the uncertainty in demand, are frequently used in airlines, restaurants, hotels, rental car companies, hairdressers, and others who deal with time-sensitive 'perishable' services. Of course, the major problem with reservation systems is 'no-shows', customers who make reservations that they do not honor, often without incurring any cost. Service companies who depend on reservations will also usually overbook their facilities, and run the attendant risk of having to turn down paying customers.

(2) **Extending the marketplace into the home**: Increasingly, companies are able to pursue prospective customers directly into their homes or offices by means of catalogs, phone calls or television advertisements. In fact, the catalog sales industry is one of the fastest growing segments within American business. A daily average of 30 million catalogs were sent during 1991, selling $50 billion (US) during that year.[14] Using direct marketing, catalog companies can shape the demand by controlling the timing and the quantity of their mailings, i.e., when and how many catalogs to mail every week. In other words, they smooth the demand by shifting it in time.[15] Direct marketing can also boost demand in off-peak periods by offering coupons, gifts, or special discounts. Its application transcends the catalog sales industry and can be found in a variety of service companies including banks, fast food chains, and retailers.

(3) **Price differentiation: smoothing demand over time via price incentives**: This strategy is frequently used to shift demand from one period to another. Managers use a differential pricing scheme to encourage people to use the service facility at off-peak times. This mechanism has been implemented in a Chilean telemarket, Europa, that sells supermarket products by phone. Because most Chileans are paid at the end of the month, peak demand occurs during the first half of every month. In order to smooth the demand, Europa began offering a 5% discount on orders placed during the last two weeks of the month. With this simple strategy, Europa improved the quality of its service, reducing the percentage of busy telephone lines and the number of complaints for not delivering orders on time. Furthermore, Europa can now meet demand with fewer trucks and employees.

The effectiveness of price differentiation mechanisms depends highly on the price elasticity of demand. Discount fares for rooms during weekends in city hotels, night rates for long-distance telephone calls, and special matinee prices for movies are all examples of effective uses of this mechanism. But the lower the price elasticity, the lower its impact on demand. For example, a price differentiating scheme was recently implemented in the Chilean subway system for the purpose of reducing the high level of congestion at peak times. The measure was very controversial because some experts believed that the effect on redistributing the demand would be limited due mainly to the low flexibility in commuters' job and school schedules. The subway authorities claim that this measure was successful.

Finally, this mechanism can also boost demand. Potential customers may begin using the service at lower price points, increasing the utilization rate of the service facility.

(4) **Informing customers about work load**: This simple mechanism attempts to 'educate' customers about the profile of utilization of the service facility so they can modify their purchasing behavior and, whenever possible, to request the service during low utilization periods. For example, a few years ago, while Fidelity Investments brokerage was upgrading its facilities, they provided customers with information about the distribution of calls received during the day, so they had the option of calling at a different time when the queues might be shorter. This was a successful temporary measure.

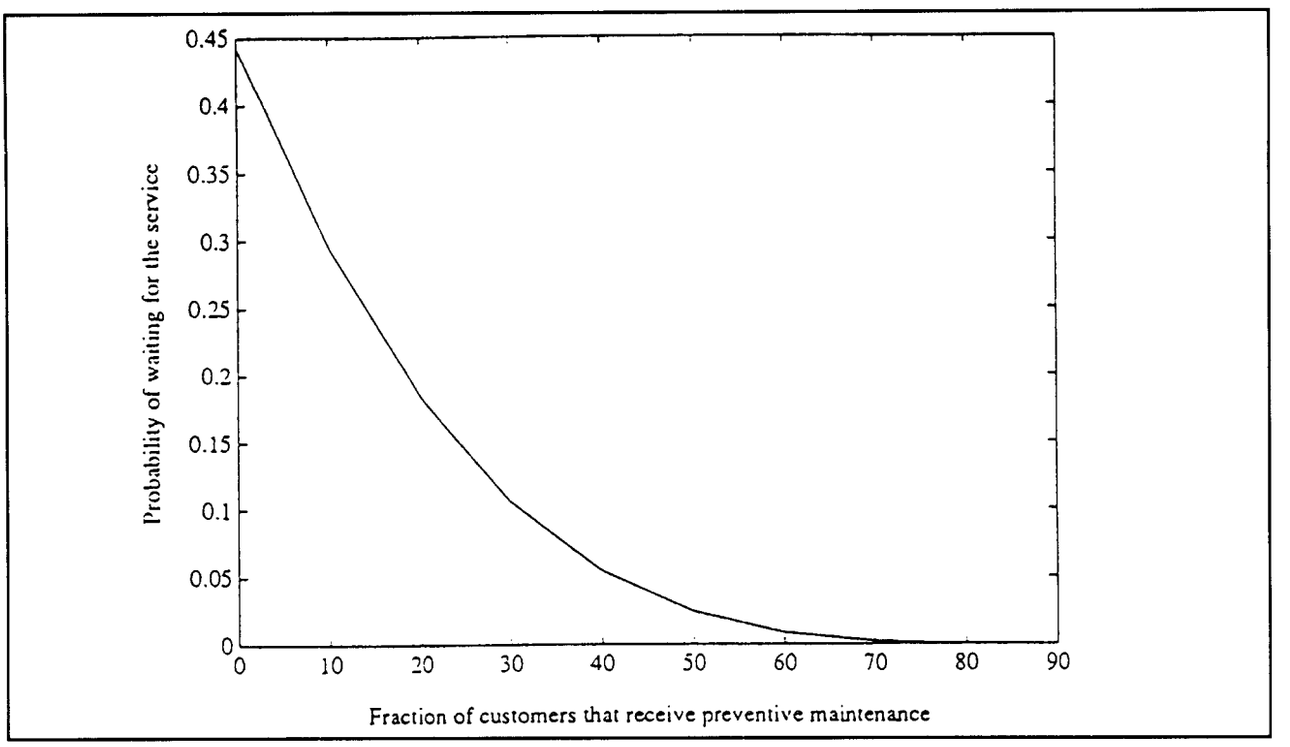


Figure 6 Model of Probability that a Customer Requesting a Service will have to Wait, Assuming Different Levels of Preventive Maintenance

(5) **Promotions and sales**: Retail stores usually plan their big sales and promotions at the beginning of the season, even though they still do not know which products will be on sale and at what prices. However, using past experience they know that boosting demand is required for some fraction of products that will not sell well. Promotions and sales are more necessary for seasonal or perishable products, such as high-fashion clothing, special foods for holidays, and airline tickets.

(6) **Complimentary services; alternative to price differentiation**: Another mechanism for smoothing demand by attracting customers during off-peak periods is to offer complimentary services. For example, malls can offer child care during weekdays, or a party of two pays only for the more expensive main course in a restaurant from Mondays to Thursdays.

(7) **Preventive maintenance of users' equipment**: This mechanism is used to reduce the demand at peak times. For instance, service companies that repair heating and air conditioning systems for a fixed annual charge usually perform preventive maintenance to the customers' equipment breakdowns for the firm to handle efficiently will occur. Consider for example an organization with five employees and an average arrival of four job repairs per day. An employee spends on average one day to fix the problem. We assume that if the company performs preventive maintenance to a certain fraction of customers' equipment, then the arrival rate of jobs will decrease in the same fraction. Figure 6 shows the probability that a customer requesting the service will have to wait, assuming different levels of preventive maintenance.[16] For example, if there is no preventive maintenance, then on average 44% of customers have to wait. However, with preventive maintenance provided to 20% of customers, then, on average, only 18% of customers have to wait. This percentage is reduced to only 2% if preventive maintenance is performed for 50% of all customers.

(b) Operational decisions

Some examples of short term mechanisms include pricing and daily specials.

(1) **Pricing**: Pricing is one of the most common mechanisms for influencing demand. There are two pricing policies that are usually implemented for yield management (see Box 1). First, dynamic pricing strategy that depends on the remaining capacity (or inventory), the time left until the end of the planning horizon, and the future demand (which is, in general, a function of the price path). Often, this policy is referred to as 'hidden price' because the manager does not know the customers' reservation prices, i.e., the maximum price a customer is willing to pay for the product. Therefore, she faces the trade-off between losing the customer surplus due to a low price and losing a potential sale due to a high price.[17] The second strategy, known as 'revealed price', assumes that the demand can be segmented according to the maximum price that customers are willing to pay for the product. For example, there may be different fees depending on whether the customer is an employee of a firm that has a special agreement, the customer uses a particular credit card, or the customer is a government employee. Given the market

segmentation, prices are kept constant during the planning horizon. Thus, the goal of the service managers is to set the target levels for each of the market segments, which can be fixed or dynamically updated in time.

(2) **Loss leaders**: This includes all special promotions that are designed in the short term either to get rid of unsuccessful products or to attract people to the stores. For example, supermarkets usually offer a few products at very low prices (many times they lose money on those products) to increase the flow of customers into the store.

(B) Perceptions Management: How to Seem Better

Given that, even in a 'well-designed' facility, people will have to wait at least some of the time, perception management mechanisms can be valuable. These mechanisms 'manage' the time perception of customers waiting to be served. A customer's perception of a ten-minute wait can be very different depending on the environment in which the wait takes place. For example, Katz, Larson, and Larson[18] reported the impact that an electronic news board, installed in a bank, had on customers' perceptions of their waiting experience; the news board made the time spent in line more palatable, interesting, and relaxing. A survey showed that overall satisfaction with the service received from the bank increased when the news board was present. Sasser, Olsen and Wyckoff[19] give the example of 'a well-known hotel group that received complaints from guests about excessive waiting times for elevators. After an analysis of how the elevator service might be improved, it was suggested that mirrors be installed where guests waited. The natural tendency of people to check their personal appearance substantially reduced complaints, although the actual wait for the elevators was unchanged.' Maister[20] establishes eight propositions of the psychology of queues, which are:

1. Unoccupied time feels longer than occupied time.
2. Pre-process waits feel longer than in-process waits.
3. Anxiety makes waits seem longer.
4. Uncertain waits are longer than known, finite waits.
5. Unexplained waits are longer than explained waits.
6. Unfair waits are longer than equitable waits.
7. The more valuable the service, the longer I will wait.
8. Solo waiting feels longer than group waiting.

Customers' behavior and perception are context-specific; they are influenced, for example, by cultural aspects, expectations, and time when the service is needed. For example, studies show that a 'reasonable' wait for customers depends on when they request the service; at peak times they are willing to wait longer. Cultural aspects are also important because they influence the way people perceive time. A study made by R. Levine, a professor of psychology at California State University, and reported in *Across the Board*[21] in 1992, finds that the pace of life and perception of time differ widely in American cities. This study measures the amount of time taken by certain activities in 36 American cities (nine cities in each of the four regions determined by the census: the Northeast, Midwest, South, and West) and concludes that cities can be classified as type A where there is a sense of time urgency, hostility, and competitiveness, and as type B where there is a slower, and more relaxed attitude. In this study Boston is classified with the largest index of pace of life (time urgency) and Los Angeles with the lowest index.

Therefore, every solution should be tested before implementing it in the service facility, because there could be some hidden effects that are not considered in advance. For example, Katz, Larson, and Larson[22] reported on the effects of an electronic clock installed in a bank, which, by means of an 'electric eye' at the entrance and exit of the queue, estimated the length of waits. In this study, the authors hypothesized that 'a wait where the length is known in advance would be less stressful than an open-ended wait,' so they hoped that the clock-phase respondents would have a lower stress level. However, control-phase and clock-phase respondents did not rate differently either their stress levels or their overall satisfaction with the service received. The authors explain that 'this may be because the clock made respondents more aware of the time wasted standing in line.'

IV Finding the Most Effective Mix

What is the best mix of mechanisms for a firm? To answer this question, service managers have to understand the nature of their service organizations. In what follows we present four dimensions that must be considered when making this decision.

A What Drives the Demand?

It is important to understand the underlying causes that determine the demand fluctuations over time. These fluctuations are due to a combination of uncertainty in the customers' purchasing pattern and inherent seasonalities of the service.

1 Customers' habits

When the demand pattern is largely influenced by customers' usual practice and not by clear economic reasons, there may be many opportunities to manage demand proactively. Because customers do not have exogenous constraints that prevent them from modifying their behavior, price differentation, complimentary services, promotions and sales, and informing customers about work load may prove to be suitable mechanisms to smooth the demand over the planning horizon. Customers can benefit (save time, money or receive a better quality service) just by requesting the service at another moment in time.

2 Actions by third parties

Many times the customers' purchasing pattern is determined by other people or organizations that establish clear constraints on when the service will be required. For example, peak times in public transportation are usually determined by work and school schedules, some peak times at banks are determined by tax payments, and retail stores' demand is higher after working hours and during weekends. In these cases, it is easier to implement mechanisms to manage the supply of services — that is, to manage capacity through both tactical and operational means — rather than to shift the demand. Thus, extending or redistributing business hours, sharing equipment, managing the service length, and dynamic scheduling can be good alternatives to satisfy the demand. If the degree of control from third parties is low, then some demand management mechanisms may also be appropriate; for example, informing customers about work load and offering complimentary services.

3 Unpredictable events

Almost all types of demand carry a degree of uncertainty. The importance of this uncertainty in the design of a service is a matter of degree. For example hospitals, and fire and police stations are service organizations where this feature is one of the most important considerations in the design of the service facility. The problem of matching supply and demand in these cases is addressed, in part, by designing the system to satisfy the demand instantaneously with high probability; queues or reservation systems are not possible. These organizations have only limited prospects for managing demand in the short term. To match the supply with the non-predictable demand, they use mechanisms like location, automation, modular facilities, cross-training, and share capacity with competitors.

B What Constrains the Delivery of Service?

The delivery of a service may require a customer to be physically present in the service facility for all, some, or none of the servicing.

1 Customer present throughout the delivery of the service

Examples include restaurants, hairdressers, and transportation systems. Factors like location, layout, and degree of automation play a central role in the design of the service facility. The mechanisms of managing the perceptions of waiting time, and the use of reservation systems 'to inventory customers' are effective for such services.

2 Customer present for part of delivery

For services such as car repairs, where the customer leaves the car to be serviced, and medical tests, where specimens are collected for later laboratory analysis, the customer needs to present only for part of the time. Under these circumstances, the reception of customers becomes most important, and service managers may, for instance, make use of information systems containing customers' personal information, and may choose convenient locations. Extended business hours for customer service, as well as the use of part-time employees to increase the capacity of the front office at moments of high demand, are also effective.

3 Customer not present

Catalog sales, and telecommunications are examples of this type of service organization. Automation of telephone steps, scheduling, and batching the delivery of the service provide alternatives to manage the supply and demand.

C What Employee Skills Are Needed?

Service operations can also be classified according to the degree of qualification required by their employees. Thus, in companies requiring less education or specialization, the supply of services may be managed by cross-training, part-time employees, and floating staff provide. On the other hand, companies that depend on highly-skilled labor, usually professional services, are constrained by more rigid supplies and, therefore, must turn to demand management techniques.

D How Specialized is the Equipment?

The alternatives of sharing capacity, and managing demand are the most suitable mechanisms to use in service companies that require expensive and specialized equipment.

V A Study: Airline Companies Use Multiple Mechanisms

To achieve the goal of delivering quality service while maximizing profits, airline companies must constantly deal, in the short run, with matching the uncertain demand for flights with their fixed capacity (number of planes and number of seats in a plane). In the process of purchasing the service, a customer faces different stages where he must interact somehow with the airline or travel agency. Some of these are the reservation and purchasing of tickets, the check-in, the preboarding, the flight, and the check-out. In all these interactions customers evaluate the quality of the service. Hence, each of these stages must be carefully designed.

The first contact that the customer has with the airline or travel agency is when he decides to make a reservation for a flight. The access to the system plays a key role in the purchasing process; busy telephone lines during peak hours can prevent customers from buying the service. Therefore, scheduling the appropriate number of people to answer the phones can be as crucial as having departures and arrivals on time.

Later on, in the check-in process, the queuing phenomenon is observed again. At this stage, passengers do not have an alternative to waiting in line to be served. During the preboarding process, 'captive' customers must wait to board the flight. Occasionally, they experience delays due to bad weather conditions, or mechanical and personnel problems. For example, in the United States in 1991, flights averaged a 34-minute delay prior to boarding and over 300 flights a day were delayed.[23] Taylor[24] shows that delays can have a negative effect on customer evaluations of service and suggests that service providers should attempt to either shorten delays for service or change the consumer's wait experience so that it results in less uncertainty and anger.

Filling a flight with the 'right' customers is one of the hardest tasks that managers in the airline industry face. Notice the difference between a plane that takes off with half its seats empty and a furniture company that in a given month faces a demand equal to half its production. The excess furniture production can be inventoried to satisfy the demand in future periods, but the excess passenger capacity has zero value after the take-off. Acknowledging this challenge, American Airlines has developed a decision support system (see Box 1) for the reservation process for each flight; the system maximizes revenue by controlling reservation availability via three major functions: overbooking, discount allocations, and traffic management. As stated by the managers of American Airlines,[25] when comparing the role of yield management to the inventory control function for a manufacturing company, 'the nature of today's marketplace in the airline industry makes yield management absolutely essential to profitable operations.'

Finally, at the check-out stage, customers have to wait for their baggage. There is an interesting case[26] where customers complained vehemently about lengthy luggage handling delays, although customers never experienced a waiting time longer than eight minutes, which is an acceptable standard in industry. A critical analysis showed that the waiting time consisted of one-minute walking time from the plane to the luggage carousel and seven-minute waiting time at the carousel. The most likely explanation about the customers' dissatisfaction is the perception of injustice because passengers who arrived even after them but without baggage could start the working day before them. The solution to this problem consisted of deliberately inserting delays in the system so almost all passengers would leave the airport together. With this purpose, the authorities moved the aircraft disembarking area far away from the luggage main area so the walking time was increased from one to six minutes (the total waiting time remained unchanged). After this change the number of complaints dropped dramatically. Martin,[27] who reported this case, calls the solution an example of 'perception management.'

VI Conclusions and Comments

In any service organization, a number of mechanisms may be used to match a limited supply of services with an unpredictable demand for those services. Tactical and operational mechanisms, which actually enhance the organization's performance, may either increase absolute capacity and efficiency, or shift the demand from peak

Box 1 Yield Management

Yield management is a technique that has been created and developed in the last 20 years with several successful applications. This technique is suitable when three features are present:

- The product or service is perishable: after a moment in time it is not available or it ages (for example, hotel rooms, flight seats, and seasonal and fashion clothing),
- there is a fixed capacity or inventory that cannot be modified in the short run, and
- different market segments are willing to pay different prices for the same product.

Several definitions of yield management have been given in the literature. American Airlines officials describe it as 'selling the right seats to the right customers at the right prices.'[28] Belobaba[29] defines yield management as the efforts to manage, using both pricing and seat inventory control, the revenue mix of passengers carried. The seat inventory control problem is 'to determine the number of seats to make available to each fare class from a common or shared inventory (i.e., the coach cabin of the aircraft) so as to maximize total expected revenues for a scheduled future flight leg departure.'

Weatherford and Bodily[30] present an extensive review of yield management. They propose a new name — Perishable-Asset Revenue Management (PARM) — and define it more broadly 'to be the optimal revenue management of perishable assets through price segmentation.' The key questions that PARM attempts to answer on a unit-by-unit basis are:

- How many units should be made available initially at various price levels?
- How should this availability of units change over time as the time of actual availability approaches; that is, when should certain price levels be closed out or opened up?

Several successful applications of yield management have been reported in the literature. For example, American Airlines, one of the leading users of yield management, estimates that t his accounts for about 5% (or $500 million) of annual revenues.[31] Duke University Diet and Fitness Center (DFC) in Durham, North Carolina, has also reported important gains using this method to increase capacity utilization by knowing whether and when to apply discounts and how to manage capacity and improve revenue yields beyond 100%. At the time the system was implemented, DFC projected an increase in revenue of 10% because of the yield management techniques.[32] Other applications have been reported in the hotel industry and retailing.

periods to off-peak periods. These mechanisms differ in the complexity of their design and implementation; some require only qualitative analysis while others call for mathematical models or analytic tools like simulation, queuing theory, and mathematical programming. Finally, perceptual mechanisms, which alter only the customer's perceptions of the organization's performance, may also be used to maintain customer satisfaction when delays in service are unavoidable. Most service firms will want to use a mix of these mechanisms; airline companies have used many different mechanisms with favorable results.

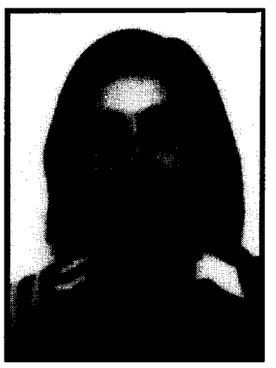
The quality of service that a firm can provide will depend on management's skill in choosing the right mix of mechanisms for its own particular business environment, and the carefulness with which the chosen measures are implemented. One very important overarching principle should unify whatever measures a firm puts into place: customers need to be aware, always, that their satisfaction drives your management decisions.

Notes

1. Priority Management, Pittsburgh, PA, 1988.
2. See, for example:
   Sasser, W. (1976) Match Supply and Demand in Services Industries. *Harvard Business School*, 44–51;
   Lovelock, C. (1983) Classifying Services to Gain Strategic Marketing Insights. *Journal of Marketing*, **47**, 9–20;
   Bowen, D. and Cummings, T. (1990) Suppose We Took Service Seriously. In *Service Management Effectiveness*, ed. D. Bowen, R. Chase, T. Cummings (Jossey-Bass), 1–14;
   Flipo, J. (1991) On the Strategic Implications of Tangible Elements in the Marketing of Industrial Services. In *Service Quality, Multidisciplinary and Multinational Perspectives*, ed. S. Brown, E. Gummesson, B. Edvarsson, and B. Gustavsson (New York: Lexington Books), 123–134;
   Bitran, G. and Lojo, M. (1993) A Framework for Analyzing Service Operations. *European Management Journal*, **11**(3), September, 271–282.
3. Sasser, W. (1976) Match Supply and Demand in Service Industries. *Harvard Business Review*, November–December, 44–51.
4. Lovelock, C. (1983) Classifying Services to Gain Strategic Marketing Insights. *Journal of Marketing*, **47**, 9–20.
5. The Parker House (B), Harvard Business School Case, 0-580-152, 1983.
6. William Henderson, Vice President, Employee Relations of the United States Postal Service. Talk's title: 'Service Improving Strategies of the United States Postal Service', at MIT Symposium *Services Industries: Decision Technologies to Achieve Productivity Improvement*, Boston, 1993.
7. Gleckman, H. (1993) The Technology Payoff. *Business Week*, June 14, 57–68.
8. *Ibid.*
9. Zemke, R. with Schaaf, D. (1989) *The Service Edge: 101 Companies That Profit from Customer Care*, (New York: Penguin Books USA), 479.
10. For an extensive review of the FedEx case, see Lovelock, C. (1994) *Product Plus: How Product + Service = Competitive Advantage*, (New York: McGraw-Hill, Inc), 121–141.
11. *Ibid*, 285–286.
12. Bitran, G. and Mondschein, S. (1995) An Application of Yield Management to the Hotel Industry, Considering Multiple Night Stays. *Operations Research*, **43**(3), May–June, 427–443.
13. Rafaeli, A. (1989) When Cashiers Meet Customers: An Analysis of the Role of Supermarket Cashiers. *Academy of Management Journal*, **32**(2), 245–273.
14. Holiday Hopes of the Catalog Industry: Merrier Christmas, Happier New Year. *The Wall Street Journal*, December 2, 1992.
15. For an extensive description, see Bitran, G. and Mondschein, S. (1995) A Comparative Analysis of Decision Making Procedures in the Catalog Sales Industry. (Cambridge, Massachusetts: MIT Sloan School of Management, Working Paper).
16. We assume that the interarrival time of jobs and the service time are exponentially distributed.
17. See, for example, Bitran, G. and Mondschein, S. (in press) Periodic Pricing of Seasonal Products in Retailing. *Management Science.*
18. Katz, K., Larson, B. and Larson, R. (1991) Prescription for the Waiting-in-Line Blues: Entertain, Enlighten, and Engage. *Sloan Management Review*, Winter, 44–53.
19. Sasser, W.E., Olsen, J. and Wyckoff, D.D. (1979) *Management of Service Operations: Text, Cases and Readings* (New York: Allyn and Bacon).
20. Maister, D.H. (1984) The Psychology of Waiting in Lines. *Harvard Business School Note* 9-684-064, Rev. May.
21. The Fastest City in America. *Across the Board*, April 1992, 52–53.
22. Katz, K., Larson, B. and Larson, R. (1991) Prescription for the Waiting-in-Line Blues: Entertain, Enlighten, and Engage. *Sloan Management Review*, Winter, 44–53.
23. Air Transportation Association, 1992.
24. Taylor, S. (1994) Waiting for Service: the Relationship Between Delays and Evaluations of Service. *Journal of Marketing*, **58**, April, 56–69.
25. Horner, P. (1991) The Best in MS. *OR/MS Today*, August, 42.
26. Larson, R. (1987) Perspectives on Queues: Social Justice and the Psychology of Queuing. *Operations Research*, **35**(6), November–December, 895–905.
27. Martin, A. (1983) Perception and Value Management. *Think Proactive*, 85–101.
28. Horner, P. (1991) The Best in MS. *OR/MS Today*, August, 42.
29. Belobaba, P. (1989) Application of a Probabilistic Decision Model to Airline Seat Inventory Control. *Operations Research*, **37**(2), March–April, 183–197.
30. Weatherford, L. and Bodily, S. (1992) A Taxonomy and Research Overview of Perishable-Asset Revenue Management: Yield Management, Overbooking, and Pricing. *Operations Research*, **40**(5), September–October, 831–844.
31. Horner, P. (1991) Eyes on the Prize. *OR/MS Today*, August, 34–38.
32. Chapman, S. and Carmel, J. (1992) Demand/Capacity Management in Health Care: An Application of Yield Management. *Health Care Manage Rev*, **17**(4), 45–54.

GABRIEL BITRAN, *Sloan School of Management E53-390, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.*

Gabriel Bitran is Professor of Management Science at MIT Sloan School of Management. He is the Editor-in-Chief of Management Science *and a member of the editorial boards of several management journals including the* European Management Journal. *His research interests lie in the field of operations management in manufacturing and the service industry, and he has published and consulted widely in operations management.*



SUSANA MONDSCHEIN, *Department of Industrial Engineering, University of Chile, P.O. Box 2777, Santiago, Chile.*

Susana Mondschein is Assistant Professor of Operations Management in the Industrial Engineering Department of the University of Chile. Her research interests include the study of optimal decision-making under uncertainty, and service management problems. She has worked on applications to retailing, catalog sales, hotel reservations, pollution control, and urban transportation among others.