8

Demand and capacity management

Introduction

In service industries the matching of capacity and demand is particularly difficult. There is either too much demand for the capacity, putting a strain on resources, or too little demand, giving rise to unused capacity and a loss in revenue. This is known as the *perishability factor*, whereby revenue from a unit of capacity, e.g. a hotel bedroom that's not sold on a particular day, is lost forever, unlike a product that if unsold can be stored in a warehouse for sale at a later date. Therefore services need to develop some understanding of demand patterns. While the level of demand can sometimes be outwith the organization's control, strategies are available for measuring demand, along with capacity, with a view to bringing them into balance.

One technique that is becoming highly significant for demand and capacity management in the service sector is that of yield or revenue management. By setting different prices for different time periods organizations seek to increase the yield or revenue from the service.

One problematic symptom of excess demand is queuing. Whilst reservations and advance booking can mitigate its occurrence, in some situations it still occurs and will remain so with the inevitable frustrations for customers.

8.1 The basic problem: perishability

If a manufacturer of cars, washing machines or furniture fails to sell some of its products on a particular day or over a certain period they are stored for sale at a later date when demand increases. For organizations providing a service, e.g. hotels, trains, their ‘product’ in the form of rooms and seats is not capable of being stored. Where a hotel room on a particular day and a train seat on a particular journey lies empty the revenue that could have been gained is lost and can never be regained. The ‘products’ have in effect perished. The situation is the same for services such as counselling where the unit of capacity is time. Perishability is a significant issue for services then and can best be summed up as follows:

- If demand far exceeds capacity it cannot be met as in manufacturing, by taking goods from a warehouse.
- If capacity far exceeds demand the potential revenue from that service is lost.

Where service capacity is largely fixed and demand is subject to variation, organizations can experience any of the following situations:[1]

- Excess demand – the level of demand exceeds maximum available capacity
- Demand exceeds optimum capacity – the service is less than adequate
- Demand and capacity are well balanced – this is the level of optimum capacity
- Excess capacity – demand is below optimum capacity.

From Figure 8.1 it is evident that the level of capacity utilization will impact on the quality of service. As the name implies, the optimum is the best level in most cases for all parties concerned – customers, employees and the organization. Operating at maximum capacity, however, or even beyond as with some transportation services, is seen as a desired feature of some service situations, namely nightclubs and bars. (The trend is now for ‘vertical bars’, standing room only, where most of the tables and chairs are removed allowing a much higher number of customers to enter.) Otherwise services must aim to occupy that optimum capacity zone. There has been little comment as to where it lies. For an airline it is said to range from 65 to 75%[2] and for most services the optimal capacity appears to be between 70 and 90%[3] of maximum

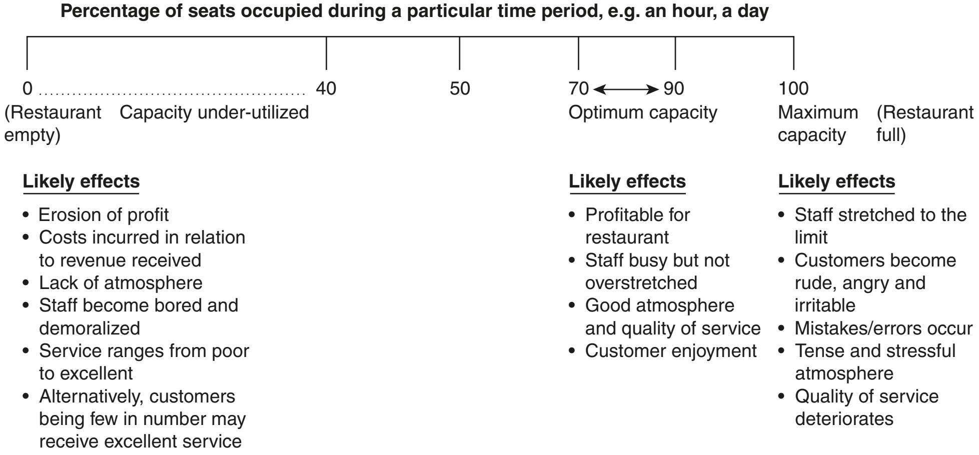

Percentage of seats occupied during a particular time period, e.g. an hour, a day
0
40
50
70 ⟷ 90
100
(Restaurant empty)
Capacity under-utilized
Optimum capacity
Maximum capacity
(Restaurant full)
Likely effects
• Erosion of profit
• Costs incurred in relation to revenue received
• Lack of atmosphere
• Staff become bored and demoralized
• Service ranges from poor to excellent
• Alternatively, customers being few in number may receive excellent service
Likely effects
• Profitable for restaurant
• Staff busy but not overstretched
• Good atmosphere and quality of service
• Customer enjoyment
Likely effects
• Staff stretched to the limit
• Customers become rude, angry and irritable
• Mistakes/errors occur
• Tense and stressful atmosphere
• Quality of service deteriorates

Figure 8.1 The impact of capacity utilization in a restaurant

capacity. Where specifically it may lie will be conditioned by the service type or situation. Customers, in general, could be said to feel happier if less than all the seats on a train or in a restaurant were occupied. Equally from the organization's viewpoint operating at the maximum means there is no slack to allow for addressing a range of problems that may arise, e.g. demanding customers, employee errors, diagnostic difficulties (car repair). Just how much of a challenge arriving at the optimum can be is illustrated by the matter of how many patients a doctor should see in one hour.

8.2 Service capacity: resources and assets

Service organizations draw on the following resources and assets in varying degrees depending on the type of service:

- Physical facilities designed to house customers, the obvious ones being hotels, hospitals, aeroplanes, schools. Each facility will be defined in terms of number of rooms, beds, seats and classrooms.
- Physical facilities designed for processing customers and their possessions, examples would include washing machines (launderette), computer technology (banks), X-ray equipment (hospitals), turnstiles (football stadium).
- Labour is a key element in the provision of service, e.g. waiters in a restaurant, cabin crew on an aeroplane, tellers in a bank. A number of services rely heavily on labour where the service comes to the customer, as in breakdown services, cleaning, gardening, roof repair, postal services.
- Time is a resource that serves as the basis upon which several services may be sold, e.g. a consultant, lawyer, plumber, car repair, counsellor.

There will inevitably be periods when capacity is under-utilized. On the other hand, a service's capacity will come under strain when demand is high. In such circumstances the flexibility of that capacity to meet demand will be tested.

8.3 Service demand

To anticipate and alleviate pressures on capacity, services need to have, as clearly as possible, an understanding of demand patterns. Two questions come into play. First, by how much does demand vary or fluctuate? There may be extreme variation to very little. Secondly, is the variability able to be predicted? If a predictable cycle or pattern is detectable its duration may be as follows:

- One day (varies by hour)
- One week (varies by day)
- One month (varies by day or week)
- One year (varies by month or season).

The causes of these cyclical variations may be many and will vary by type of service. Causal factors for a bus service are likely to be employment/school hours, shopping behaviour and entertainment. As there is a degree of stability to these factors

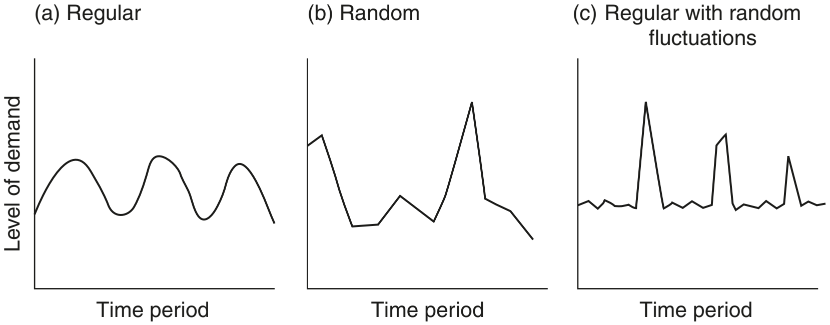


Figure 8.2 Demand patterns

demand will be largely predictable and vary by hour, less so by day. Consequently the bus company is able to schedule its fleet in tune with demand. Some services – health, breakdown, roofing, insurance – will be subject to random and extreme fluctuations in demand caused by factors outwith their control, e.g. weather, outbreak of illness. Overall, the demand patterns for specific organizations and across the service sector are not available for public consumption. However, it can be stated that demand for most service organizations will exhibit one of the following patterns (Figure 8.2).

8.4 Managing demand and capacity

Originally two strategies were suggested for managing demand and capacity: the first would involve adjusting capacity to match demand (defined as 'chasing demand') and the second, altering demand to match available capacity (known as 'level capacity').[4]

1 **Adjusting capacity to match demand**
A number of options are available for consideration:

- Extend the opening hours – this is not an option open to all service organizations. Where it is possible it is likely to occur only when demand levels are regarded as particularly excessive.
- Encourage employees to work harder – the requirement here is usually that of processing more customers per hour or per day. Although a mark of efficiency (more output from existing staff), service quality for customers may deteriorate.
- Cross-train employees – enables organizations to operate with fewer staff. Instead of being confined to handling few responsibilities staff are equipped to manage a variety of tasks and activities. It amounts to a move in the direction of job enlargement and some might say job enrichment, increasing employee motivation, satisfaction and morale. Not all employees will welcome it, particularly where there is seen to be little increase in commensurate rewards. The type of service and the organizational culture will be two prominent factors that need to be taken into account prior to such a move.
- Recruiting part-time employees – this is an option low in cost and potentially one that can be achieved quickly. Organizations should, of course, ensure that

part-time employees be given the same support and encouragement as given to full-time staff.

- Add facilities – usually in the form of table, chairs or other equipment. Just how much scope there is for this will depend on the initial configuration and layout designed to communicate a specific atmosphere and/or level of service. Adding facilities may change both.
- Hire or share facilities or equipment – may be in the form of additional physical space or vehicles required either on a temporary or recurring basis.
- Using customers as productive resources – up to this point all attempts at adjusting capacity have involved manipulating internal resources and assets. However some have suggested that organizations should regard customers as 'partial employees' and make a contribution to productive capacity.[5]
- Outsourcing (already referred to in Chapter 2) – for small to medium-sized organizations, in particular, calling on outside assistance is a valuable option in trying to meet market demand. Typical areas for outsourcing are technological and marketing support, employee recruitment and training, and Web development. Large organizations also outsource. Consider the recent case of British Airways and the outsourcing of its inflight catering to Gate Gourmet. Competitive pressures in the airline industry had forced this move. Unfortunately, the demand for lower costs led to industrial action by Gate Gourmet employees.

The above options, then, are aimed at increasing capacity to absorb demand. However, for service organizations there will inevitably be periods of time where capacity is under-utilized. Such a situation will remain so if attempts to encourage demand during these periods prove unsuccessful. It has been suggested that slack time be used 'productively as a time to train new employees, do maintenance on the equipment, clean the premises, prepare for the next peak and give the workers some relief from the frantic pace of the peak periods'.[6]

2 **Altering demand to match available capacity**

Whereas capacity management is a response to demand, demand management is an attempt to shift demand. Given the relative inflexibility of capacity organizations may seek to smooth demand by reducing the variability and fluctuation of existing patterns. Organizations can turn to the marketing mix for stimulating demand during periods of spare capacity or shifting demand during periods where capacity is operating at or near maximum. Of the '4 Ps' price and, to a lesser extent, place, offer the most potential in this area.

- Manipulate price – this will be discussed in more detail in the following section on 'Revenue Management'. The central role of price is to discourage too many customers from using the service during 'peak demand' periods and encourage more customers to select 'off-peak' periods. On price alone this strategy will only work if enough customers can be attracted by the lower prices available during low demand periods. Leisure, hospitality and transportation services would appear suited to this approach. However this strategy of price differentiation is, it is argued, not without risk. Customers may become acclimatized to the lower prices and expect them whenever the service is used. Equally there is risk to the organization's image in that lower prices may attract undesirable customers.[7] This would be particularly relevant for a service that regards itself as more upmarket or exclusive.

- Offer a mobile service – for a number of reasons consumers have welcomed the emergence of mobile services where the provider takes the service to the customer rather than or in addition to the customer having to visit the provider in some fixed location. Libraries have used this approach for many years, the service being particularly valued by the disabled and those living in remote locations. Other services that have found mobility an effective method of managing demand include breakdown and maintenance, blood donation and catering.
- Communicating with customers – the provision of information as to when demand is, or is likely to be, high appears to be a strategy not well adopted by service organizations. In particular for customers in our 'call centre society' it can be especially frustrating. Waiting is a feature of modern day society and will be addressed later in the chapter.
- Changing the service offer – for most organizations this is not an option. What they offer remains fixed. Where services with a sizeable facility like hotels experience significant seasonal fluctuations however, action may be taken to encourage varied usage of the facility when capacity is under-utilized.

8.5 Aligning demand and capacity: the options

Due to the relative inflexibility of service capacity coupled with the variability of service demand, aligning the two remains a challenge. Relatively recent writing,[8] incorporating the original thinking,[9] suggests four options are available to service organizations in determining a relationship between demand and capacity. They are:

- **Provide –** where sufficient capacity is available at all times to meet peak demand. This may mean periods of excess capacity but that is to be preferred to a situation where business is lost due to insufficient capacity.
- **Match –** where attempts are made to anticipate demand pattern so that capacity levels can be changed to accommodate. It would involve careful scheduling of work as well as considering sub-contracting or outsourcing.
- **Influence –** where demand patterns are changed, if possible, to obtain effective utilization of capacity. Responsibility for changing demand will lie with marketing deploying elements of the marketing mix as appropriate.
- **Control –** where capacity remains fixed in service situations that are unique and high cost resources are needed to provide the level of service expected. Consequently, variation in demand needs to be kept to a minimum.

Each of these options in relation to an actual demand pattern is shown in Figure 8.3.

Actual demand in Figure 8.3 is presented as more of a random rather than regular fluctuation across the six time periods. The basic question that arises is:

- What level of capacity should be maintained to satisfy the demand? Should a company provide excess capacity to improve customer service or minimize excess capacity to maximize resource utilization.[10]

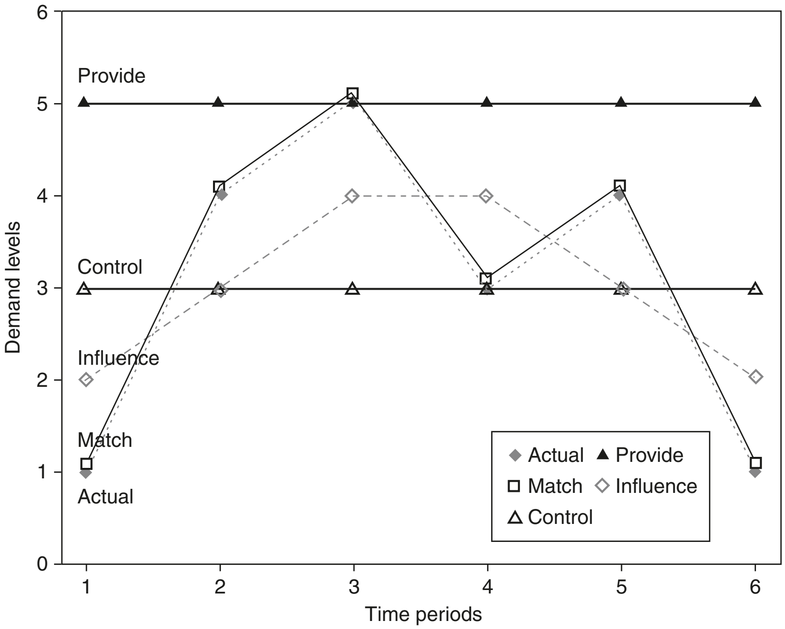


Figure 8.3 Demand management strategies – hypothetical
Source: Crandall and Markland (1996)[10], reproduced with permission from the Production and Operations Management Society

The following two options characterize the original thinking:[11]

- The 'match' option (managing the capacity) is shown as parallel to, but slightly higher than, the actual demand lines. This option is more easily managed with employees than equipment and facilities.
- The influence option attempts to smooth the actual and match capacity by shifting the peak demands to the low demand periods. Variable capacity then becomes less of a problem. It is slightly higher in low demand periods and slightly lower in high demand periods.

The more recent thinking embraces two further options:

- The provide option shown as the straight horizontal line at the top of Figure 8.3 presupposes that enough capacity is available to meet even the peak demands. It could be thought of as a wasteful option as resources will inevitably be underutilized, particularly through periods 1, 4 and 6.
- The control option is shown as the horizontal line in the middle. It is designed to meet average demand, using waiting lines or some other means of accommodating excess demand until capacity is available.

One piece of research[12] of selected service industries suggests an overall preference for the provide and match options over those of influence and control. Two broad reasons are advanced: companies' desire to improve customer service and the notion that companies understand how to change capacities better than they understand how to change demand patterns.

8.6 Yield management (also known as revenue management)

8.6.1 Definition

From the discussion so far it should be clear that service organizations face a challenge in terms of managing demand and capacity. Yield management is a technique designed to address that challenge. It is defined as 'provision of the right service to the right customer at the right time for the right price'.[13]

8.6.2 Where can yield management be applied?

Yield management is not (yet) suitable for all service organizations. Application has been most successful in services that have the following characteristics:[14]

- Relative fixed capacity – e.g. once a hotel has rented out all its rooms further demand cannot be met without substantial capital investment
- Perishable inventory – a major constraint for services is time or more specifically time during which a unit of capacity is available. If a hotel room (unit of capacity) is not sold for a particular date the revenue that would have been gained is lost
- Segmented markets – where the market for a service can be segmented according to certain criteria, e.g. price sensitivity
- Fluctuating demand – where the adoption of various pricing approaches enables the reduction of peaks and valleys in variable demand. Success in this regard results in more effective utilization of capacity
- Services that can be sold in advance through reservation systems – allows for better use of capacity
- Low variable to fixed cost ratio – in service pricing some contribution must be made towards fixed cost. The low level of variable cost, e.g. cleaning a hotel room, coupled with discretion in pricing means that the revenue expected from selling it is invariably greater than if it was not sold. That is why yield management is usually regarded as a profit-enhancing strategy.

Users of yield management fall broadly into three categories:

1. Sophisticated – airlines and large hotel chains are regarded as the classic users employing complex information systems and computer models for the purpose of analysing and predicting consumer demand patterns.
2. Moderate – theatres, trains, hairdressers, small/medium hotels use less advanced systems and technologies in the deployment of classic yield management techniques.
3. Potential – restaurants and golf courses are current and notable examples of services ready for yield management were it not for, in both cases, variability and unpredictability over the duration of service. For yield management to work there must be a fixed length of time within which the service is consumed.

8.6.3 How does yield management work?

Yield management works by recognizing and applying the following key elements.

Time

Time – when a service is consumed, e.g. hour of the day, day of the week, month of the year, is the key element in terms of determining how much a customer is likely to pay and consequently the yield that accrues to the organization. The sensitivity of customers as to when they consume the service is of prime importance in yield management. In addition to the times of consumption, the timing of any reservation may decide the price to be paid.

Demand – to enable yield calculations and assessments to be made, services need to classify demand periods and the variations between them. One possible classification can be seen in Figure 8.4.

In total there are 32 demand periods in Figure 8.4 (4 seasons [×] 2 days of week [×] 4 times of day). For yield management purposes it is a question of whether each of the 32 cells merit unique treatment in terms of price levels and customer profiles. Criticism has been made of the airline industry for its increasing fine-tuning of segmentation and combinations of fares, rules and conditions, only to find that '60–80 per cent of these products never generated a single sale of a seat to a customer'.[15]

Price discrimination – in conjunction with time, mentioned earlier, customers are also sensitive to price. Where a customer needs to use a particular service, e.g. the railways during, say, a peak period (time-sensitive), that customer will be willing to pay more (price-insensitive). The converse then is where a customer is unconcerned over, say, the time of travel (time-insensitive) and consequently will enjoy a discounted price if an off-peak period is selected (price-sensitive). Although time and demand levels are the main causal factors or rate fences in price discrimination, other factors (as we shall see) permit variation in pricing.

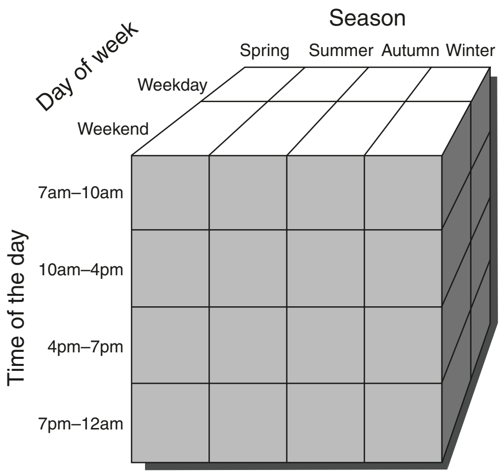


Figure 8.4 Variations in demand by time period

Rate fences

The price discrimination strategy or variable pricing as mentioned above should have some logic to it. To achieve this, yield management sets what is known as rate fences at time of usage. These are rules or conditions designed to make clear the reason for either a particular price or price differences in general. To be successful it is felt that customers need to perceive price differences as justified and fair. Where two customers are using the same service (a particular rail journey) there can be a difference in the price that each customer has paid. The customer paying the higher price may regard this situation as unfair. However, particular rate fences in this instance are designed to offer an explanation. For example, the customer paying the cheaper price has booked a seat in advance or falls into a category of customer (pensioner, student) which qualifies for a lower price. Although time of service usage is a major rate fence in yield management, other conditions or qualifications are in operation depending on the type of service. A number of rate fences of two types have been cited for restaurants:[16]

1. **Tangible rate fences**
   - Table location
   - Party size
   - Menu type
   - Presence or absence of certain amenities (e.g. bread on the table)
2. **Intangible rate fences**
   - Group membership or application
   - Time of day or week
   - Duration of use
   - Timing of booking
   - Walk-in or reservation
   - Type of reservation (guaranteed or not).

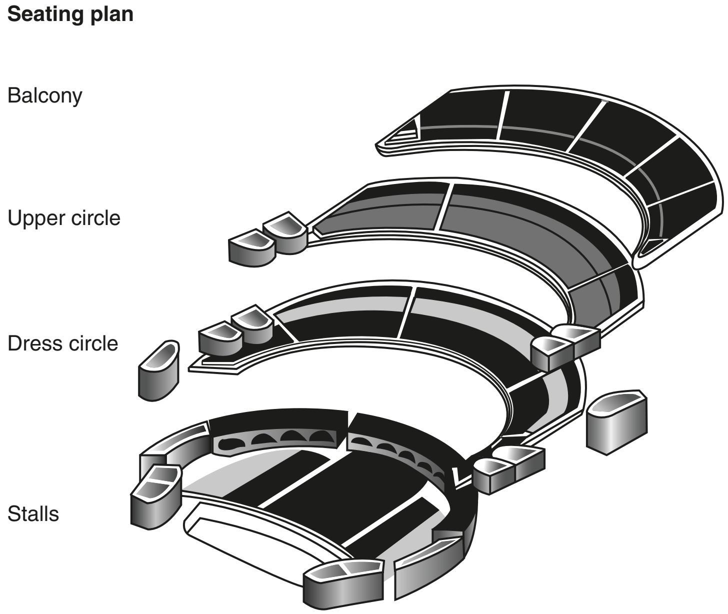
Another service industry open to yield management type activity is the opera. The ‘value’ of each seat is determined and priced according to what the customer is willing and able to pay. Seats vary in terms of ‘place utility’. In other words, all the seats do not deliver the same experience. Seat location then becomes a rate fence. Box 8.1 illustrates the seating plan of English National Opera. In line with seat location and differentiated experience, prices for the stalls and dress circle are higher than those for the upper circle and balcony. Other rate fences are in place qualifying for a price concession, e.g. students, children aged between 5 and 16 with an adult, senior citizens, tickets unsold three hours before a performance (standby tickets) and standing tickets (all seats have been sold but standing available at the back of the Dress and Upper Circles).

Yield measurement

The basic yield statistic is a straightforward measure. It is expressed as follows:

$$\text{Yield} = \frac{\text{Actual revenue}}{\text{Potential or Maximum revenue}}$$

Box 8.1 Seating plan – English National Opera



Source: http://www.eno.org/booking/seatingplan.php

Revenue potential is the revenue that could be secured if 100 per cent of capacity is sold at the maximum price possible. In more detail, yield is a function of price efficiency and capacity used, namely:

$$\text{Capacity utilization} \times \text{Price efficiency}$$

where

$$\text{Capacity utilization} = \frac{\text{Units of capacity sold}}{\text{Total units of capacity}}$$

and

$$\text{Price efficiency} = \frac{\text{Actual price}}{\text{Maximum price}}$$

(Note: terminology can vary according to industry, e.g. hotel capacity is occupancy rate, transport capacity is load factor, hotel price is rate, and so on.)

If we consider the hotel industry, yield would be found from:

$$\frac{\text{Rooms sold}}{\text{Rooms available for sale}} \times \frac{\text{Rate (price) of rooms sold}}{\text{Rate potential}}$$

For example, a hotel with 280 rooms sells in one night 140 rooms at £80 and a further 70 rooms at the maximum rate of £200. The yield would be

$$\begin{array}{rcccccccc} \text{Yield} & = & \dfrac{140}{280} & \times & \dfrac{80}{200} & + & \dfrac{70}{280} & \times & \dfrac{200}{200} \\ & = & 0.5 & \times & 0.4 & + & 0.25 & \times & 1 \\ & & \text{occupancy} & & \text{price} & & \text{occupancy} & & \text{price} \\ & & \text{percentage} & & \text{efficiency} & & \text{percentage} & & \text{efficiency} \\ & = & \multicolumn{7}{l}{0.2 + 0.25} \\ & = & \multicolumn{7}{l}{0.45 \text{ or } 45\%} \end{array}$$

$$\begin{aligned} \text{Maximum yield} &= 280 \times £200 \text{ (total number of rooms} \times \text{maximum rate)} \\ &= £56\,000 \\ \text{Actual yield} &= £56\,000 \times 45\% \\ &= 25\,200\ (140 \times £80 + 70 \times £200) \end{aligned}$$

In reality a hotel's capacity will be composed of a mix of rooms at varying prices. Suppose, in the example above, that 95 rooms are priced at £200 (maximum rate) and the remaining 185 at £80 (maximum rate), the maximum yield would be £33 800 (95 [×] £200 + 185 [×] £80). The percentage actual yield is 74.5% (£25 200 divided by £33 800). Price efficiency is maximized at 1 or 100%. On the other hand, average capacity utilization across the two categories of room is:

$$0.75 \text{ or } 75\% \qquad \left(\frac{140}{185} + \frac{70}{95}\text{divided by } 2\right)$$

With the above example in mind, measures have been established for measuring success in yield management (see Box 8.2).

Box 8.2 Revenue per available time-based inventory unit (REVPATI)

- **Airline:** Revenue per available seat mile
- **Hotel:** Revenue per available room night
- **Car rental:** Revenue per available car pay
- **Restaurant:** Revenue per available seat hour
- **Function space:** Revenue per available square metre per day part
- **Golf course:** Revenue per available tee time

Source: Kimes (2002)[17]

An example from the airline industry[18] illustrates the calculation of revenue per available seat mile with the additional consideration of cost. To begin with, operational expenses are calculated for each flight and then divided by the number of available seats on that flight. This means each seat has a fixed cost associated with it.

To make this figure even more meaningful the cost is then broken down to cost per kilometre. Therefore, each flight has an available seat kilometre (ASK), which is computed by multiplying the number of seats on the aeroplane by the distance of the flight. For example:

Boeing 747 with 400 seats
Flight: London to Johannesburg (10 000 kilometres)
$\therefore$ ASK = 4 million (400 × 10 000)

Full cost of flight = £210 000
$\therefore$ Cost per ASK = 5.25 pence per seat per kilometre
(£210 000 divided by 4 million)

To find out if the yield on revenue of the above flight is good, bad, or indifferent a further calculation is required. The marketing team needs to know the revenue per passenger kilometre (RPK) of each flight. Using the same flight and costs as above, the RPK would be worked out as follows:

380 passengers on board
Total revenue is £247 500

$$\begin{aligned}\text{Total passenger kilometres} &= 380 \times 10\,000 \\ &= 3\,800\,000\end{aligned}$$

Thus RPK = 6.5 pence, which gives a yield of 23.8% over the ASK.

The conclusion drawn was just how marginal airline operations can be. The challenge for the airline is one of managing the mix of passengers on the different classes.

8.7 Waiting and queuing

8.7.1 Why waiting matters

'Waiting is frustrating, demoralizing, agonizing, aggravating, annoying, time consuming and incredibly expensive.' This view of waiting is offered in an advertisement for Fedex, an American parcels carrier. Since one of the company's selling propositions is that they can deliver goods to customers sooner than their competitors, it is not surprising that they emphasize the disadvantages of waiting. However, it is clear that many service customers do find waiting a tiresome problem – to avoid which they are often prepared to pay extra. In James Gleick's book *Faster* we are invited to read about 'the acceleration of just about everything'.[19] In the book we read about fast food, Internet booking, and the chaos that occurs when speed is not matched by reliability – a problem we can see everyday at airports when people miss connections or are delayed because of a shortage of staff at check-in desks.

In a competitive environment the ability to provide a service without waiting confers two kinds of advantage. The first appeals to the customer – speed is one of the key service qualities which gives one firm an advantage over another – an advantage which sells cars in stock, fast food (even at the expense of taste) and photo-processing 'while you wait'. The second kind of advantage is operational – avoidance of waiting is often avoidance of waste. For example, an airliner which spends less time at an airport turn-round can make more trips, and more profit, in a day. In logistics increased speed in the supply chain means that goods for sale spend less time between factory and consumer, which improves cash flow, as well as increasing sales.

Waiting may be a lead time delay, between placing an order (or perhaps describing a symptom to a doctor) and delivery of the goods, or receiving appropriate medical treatment. 'Lead time' is the gap between signalling a demand, want or need – in commercial terms, generating an order – and satisfactory completion, or fulfilling the order. An efficient organization tells the customer what the lead time will be, meets the deadline consistently, and offers a faster response than competitors. Reducing lead time is one of the key efficiency aims of modern business. In services it is often achieved by self-service – for example by inviting hotel guests to help themselves from a breakfast buffet rather than wait for a waiter to take an order and later deliver it. In hospitals the ability to reduce waiting time between identifying the need for an operation and actually performing it is a key performance indicator – as well as being a matter of life and death.

In many service systems waiting often shows in the physical presence of a crowd of people, often in vehicles, waiting because the system is congested. Too many people (and very often too many cars) have arrived at the same place at the same time, and the system cannot cope. The people waiting, whether on foot or in vehicles, 'tail back' from the point where the flow is disrupted, and it is this 'tail-back' which gives a waiting line the name 'queue' – the French word for a tail.

8.7.2 Queues – waiting given shape

When customers flow through a service system, as they do when they travel, or check into a hotel, or arrive at a restaurant for a meal, then the need to wait often takes the shape of standing in a queue – 'waiting in line' as it is termed in the USA. There are three main causes of queues. They are:

1. The need to halt a flow of people or vehicles for a transaction. Transaction processes include selling tickets, having documents checked, passing through security and immigration checks, and being sorted into categories.
2. The slowing of a flow because there is a physical constraint which slows things down. This constraint is usually called a bottleneck, and if you up end a bottle of water you can see why. Bottlenecks occur because several flows converge (as at road junctions or transport interchanges) or because of some infrastructure limitation, like a narrow bridge crossing a river which divides a city.
3. Queues occur because people arrive before the service is open for business. Such queues include those waiting for a bus driver to arrive and open the doors, those outside shops waiting for opening time, and people waiting in areas of famine for food donations. This last example gives the reason why many service opening time queues occur. A queue ensures 'first come first served', which is particularly important if there is a belief that there will not be enough to serve everybody who waits.

The need to queue occurs frequently in service systems where customers do not order service in advance, as is true of commuter journeys by public transport, shopping in supermarkets, and driving down a city street at a busy time. In such situations the arrival rate of customers is often likely to exceed the capacity of the service system to handle the flow of people, especially at peak times. People who wait for service may find themselves in a crowd, or in a funnel-shape group moving towards a sales point, or in a panic-stricken stampede running towards an emergency exit. Forming an orderly queue is often the fairest and most efficient way to serve those who wait.

8.7.3 The reason for queues

Organizing a waiting group to form an orderly queue, or (as Americans put it) 'wait in line', has many advantages. The first is that it gives to those who wait the visible fairness of 'first come first served'. So long as there is no queue-jumping then the first person in the queue ought to be the first person to arrive, and the first person to be served. The importance which people in the queue give to this idea is very obvious if you look at the kind of queue that forms (often waiting all night) to get the best chance of bargains in the January sales of major department stores. Because there are, quite deliberately, not enough of the best bargains for everyone in the queue, who is first in the queue matters a great deal. When there are enough goods for everybody in the queue, then first come first served is a matter of equalizing the wait. In addition a queue lets people see 'where they stand' so they can make some sort of guess about how long they will wait. Queues have other advantages as a shape of waiting. They permit space to be used efficiently, as you can see when you observe two orderly lines of vehicles waiting to go through a 'bottleneck', such as a junction where two streams converge. If the vehicles stay in line, the space is used efficiently, and the 'throughput' is maximized. If drivers start jockeying for position, chaos develops, tempers rise and the speed of traffic falls.

Slow moving queues usually form for several 'supply side' reasons. The first is that the halt for processing takes a long time. This may be that the process is necessarily complicated or time-consuming, as is often the case when people are booking diverse travel tickets for future travel. Queues waiting for processing are often long because there are not enough servers to meet demand, even when peaks of demand are predictable. Two common solutions to processing queues are to encourage pre-booking (especially by purchasing on line) and adjusting supply by multi-skilling staff, so a customer service assistant can be switched from enquiries to check-in tasks. Processing often has to be careful to protect revenue or to prevent hazards, such as the ability to put bombs in airliner baggage. Waiting can be reduced by employing more service staff, by speeding or simplifying the processes (e.g. by using 'smart cards' for security), and by self-service systems, such as fast ticket machines. Bottlenecks are often very difficult to cure because the infrastructure limitation would be very expensive to remove, or because the very feature which impedes the flow is cherished for its charm. It would be possible, although costly, to build a four-lane motorway through the centre of the city of York, but to do so would destroy the appeal of the old city. Queues prior to service start times are often caused by the fear of shortages. In affluent societies they have most commonly been reduced by solving the problem of shortages. Another approach is by rationing, so that those who arrive first get no more than those who arrive later. We should note that some shortage queues are created deliberately, like those for bargain price goods in department store sales.

There are 'demand side' reasons why queues form. The most common reason is that services cannot be stored, but demand fluctuates, which is why there is often a price discount for booking early, especially on services that are likely to have unused capacity. We commonly find that the total demand for service in a working day is matched by the total capacity of the service, but that there are peak times when the number of customers exceeds the system's capacity, and queues form. Peaks occur because the demand for service is derived from other social and economic habits, such as the desire to fly from busy airports on the day before public holidays. Peaks may occur once or twice a day, or seasonally, such as at Christmas. Related to peak demands are surges. A surge is the term used to describe the sudden arrival of a large group of customers not related to peak times, but often having an identifiable cause. To take a simple example, the arrival of a coach party at a country inn will cause a surge of demand. This surge will probably have been recognized by the reservation of tables in the dining room, but the desire for a pre-lunch drink may cause a surge of demand at a bar designed to serve a local clientele. Two other demand side phenomena should be noted. The first is that the total demand for the service may exceed total supply, so there are queues at most times of day. An economist would describe this situation as one of disequilibrium and it often exists in services where there is no price mechanism to regulate demand.

Another demand problem which should be noted is that of panic – often amongst those fleeing from danger. In a smoke-filled nightclub or airliner the best way to get people out safely is to form orderly lines flowing through as many exits as possible. Reality shows us that in such situations people sometimes stampede, fight, freeze, or ignore routes to safety which could have been used.

8.8 Queuing: a behavioural perspective

A number of behavioural principles or propositions governing queuing and waiting have been proposed.[20, 21] They can be summarized as follows:

- Uncertainties – there is nothing worse from a customer's point of view than not knowing how long you will have to wait. People must be given an indication of how long they should expect to wait and it should err on the side of caution. For example, it is obviously better to say 'you will be seen in 20 minutes' or 'the take-off will be delayed by 30 minutes' where the wait is anticipated by the service provider to be less in both instances. Promising a wait of less than its predicted actual duration should be avoided at all costs. The worst thing to say would be 'We'll be with you soon' or 'The delay in departure will not be long'. In these cases the risk of customer dissatisfaction is increased because their expectations have been raised.
- The essential characteristic and apparent attractiveness of the appointment system are its complete lack of uncertainty. Unless you arrive early, there is no waiting involved as people expect to be taken at their allotted time. What is not always understood or tolerated is why a service does not adhere to the appointment schedule. A hairdresser will usually keep to the schedule while the local GP may not. The explanation lies in the greater degree of uncertainty and unpredictability surrounding patient needs in the doctor's surgery.

- Explanation – the length of a delay can be given meaning if people are told the reason(s). There are innumerable reasons for delay but the important point is that customers will make a judgement as to whether it is reasonable, acceptable, or justifiable. One might speculate that failure to inform customers is as much to do with avoiding unnecessary ridicule and censure. Train companies revelations that a particular type of snow or leaves on the line, or excessive heat, can cause severe disruption are usually met with incredulity and annoyance, particularly as there appears to be nothing that can be done about it. There is nothing more frustrating from the customer's viewpoint than serving points, e.g. in post offices, banks, supermarkets, suddenly closing without any explanation, or where service employees are seen to be 'sitting idly by' while the queues get longer. Ideally an explanation of the real cause of the delay should be given, but in the real world organizations may decide to give out either no reasons or ones deemed more acceptable to the customers.
- Anxiety – this feeling can often be the consequence of uncertainty and no explanation. It is the product of thinking 'I'll never be seen to', which with the advent of appointments and take-a-number systems is less frequent. However, it can be felt when standing far back in a very long queue entering a capacity-constrained facility, e.g. a pop concert, a football match. Organizers can eliminate it with the reassurance that 'everyone will get in'.
- Boredom – waiting can be incredibly boring. If organizations can offer some desirable distractions that take customers' minds off the time the response might well be one of 'how time flies'. There has not been a great deal of imagination generated in this area. Successfully filling unoccupied time is a difficult exercise and an area where there is enormous scope for experimentation and development. Quite often customers take it upon themselves to fill in time by befriending each other.
- Pre-process versus in-process – this is partially related to the previous point in that although customers have to wait they want to feel as soon as possible that progress is being made towards the service commencing. The obvious example is being given a menu on sitting down for a restaurant meal. The important point is that customers need to feel they are involved as quickly as possible. Anything the provider can do to fill in the customers' time before the core service begins will achieve that objective.
- Value – in general people value their time, so what they are waiting for has to be worth it. What customers define as 'valued' is as diverse as the reasons for delay mentioned earlier. It is not simply a matter of what is valued being seen as something highly priced. Long waits are endured to obtain an important prescription from a doctor. People camp out overnight or wait many hours to obtain entry to various forms of entertainment. The wait is very much seen in terms of 'it's worth it'.
- Equity – people correctly feel aggrieved if the first come first served (FCFS) and first in first out (FIFO) systems are not observed. The take-a-number procedure operated by many services is a good example of FIFO. It works well when all customers' transactions require about the same amount of time, but not where markedly different amounts of time are in evidence. There are instances where FIFO and FCFS would seem to be violated, e.g. an emergency arrival at a hospital, but this is an example where customers accept non-observance of the rule.

Sometimes what seems to be a breaking of the equity rule is understandable. For example, a person waiting in a restaurant for a seat witnesses a party of four entering and being given a table straight away. The single person may feel a sense of injustice but the restaurant owner cannot be expected perfectly to match his fixed capacity of seating arrangements with the unpredictability of customer demand patterns. Equally a customer in a department store may feel frustration when telephone callers receive priority service. An occurrence such as this can be avoided through the adoption of proper procedures.

8.8.1 Applying queuing theory

We have seen that many queues occur because of the need to halt a flow for processing. This is a problem to which queuing theory can usefully be applied. Queuing theory is the application of the mathematics of probability to the likelihood of queues. Those managing systems wish to ensure that the maximum duration of queues stays within given limits, because they do not wish to exceed the zone of tolerance of the customers, who may renege if queues are too long. The second desire is to ensure that serving staff (and their associated equipment) are efficiently utilized, so there is a wish to avoid situations where there is no waiting at all because staff have nothing to do. To solve this problem the managers of the system can employ mathematical models. These models may become quite complex, but they all start with two basic sets of data: the average arrival rate, denoted by λ (the Greek letter lambda) and the average service rate, denoted by μ (μm). From these two statistics a value for the traffic intensity, shown by ρ, can be derived. Traffic intensity can best be thought of as the average utilization of the service facility with the formula:

$$\rho = \lambda/\mu \text{ (note } \lambda < \mu\text{, otherwise we can end up with an infinite queue).}$$

Asking the question 'how many … ?' we have the two formulae:

$$M_q = \rho^2/(1 - \rho) \text{ and } M_s\rho/(1 - \rho)$$

where M_q is the average number of items in the queue and M_s is the average number of items in the system.

Asking the question 'how long to wait … ?' requires the following formula:

$$W_q = \rho/(\mu - \lambda) \text{ and } W_s = 1/(\mu - \lambda)$$

where W_q is the waiting time in the queue and W_s is the waiting time in the system.

Worked example: Customers in a supermarket join a single queue at an average rate of 30 per hour and are served at an average rate of 35 per hour. Find the average:

1 Number of customers waiting to be served (m_q)
2 Number of customers in the system (m_s)
3 Time spent queuing (w_q)
4 Time spent in the system (w_s)

Solutions: $\rho = 0.857$.

1 $(0.857)^2/(1 - 0.857) = 5.14$, i.e. 5 customers
2 $(0.857)/(1 - 0.857) = 5.99$, i.e. 6 customers

3 $(0.857)/(35 - 30) = 0.1714\,h$, i.e. 10.28 min
4 $1/(35 - 30) = 1/5\,h = 12\,min$

Summary

Service providers face a particular problem when it comes to demand management and capacity utilization. Unlike manufacturing, service organizations cannot stockpile their 'output' in a warehouse and wait for demand to materialize.

Demand for services can fluctuate in such an unpredictable way that capacity is either unable to cope, or grossly under-utilized. Achieving a match between demand and capacity is therefore a difficult goal to achieve. Services will vary in terms of how easily capacity can be adjusted. A number of options are available for making capacity more flexible, two of which are gathering interest, namely using customers as productive resources, and outsourcing. Demand represents more of a challenge. Among the options here, variable pricing remains attractive.

A technique attracting attention is that of yield or revenue management. By applying a number of key elements you will have gathered a basic understanding of how this technique works. No matter how successful yield management is, there still remains the residual matter of customers having to wait and queue. This is simply a feature of service for which there appears to be no available solution. However, attending to the behavioural principles governing queuing and waiting should focus your mind on how to make this phenomenon as comfortable and fair as possible for customers.

References

1 Lovelock, C (1994) *Product Plus*. New York: McGraw-Hill, ch. 16.
2 Heskett, L, Sasser, W E and Hart, C W L (1990) *Service Breakthroughs: Changing the Rules of the Game*. New York: Free Press, p. 139.
3 Kurtz, D L and Clow, K E (1998) *Services Marketing*. New York: John Wiley and Sons, p. 346.
4 Sasser, W E (1976) 'Match, supply and demand in service industries', *Harvard Business Review*, Nov.–Dec., 133–140.
5 Mills, P K and Morris, J H (1986) 'Clients as partial employees: role development in client participation', *Academy of Management Review*, **11** (4), 726–735.
6 Sasser, op. cit.
7 Zeithaml, V A and Bitner, M J (2003) *Service Marketing*. New York: McGraw-Hill, p. 421.
8 Crandall, R E and Markland, R E (1996) 'Demand management – today's challenge for service industries', *Production and Operations Management*, **5** (2), 106–120.
9 Sasser, op. cit.
10 Crandall and Markland, op. cit.
11 Ibid.
12 Ibid.
13 Smith, B C, Leimkuhler, J F and Darrow, R M (1992) 'Yield management at American Airlines', *Interfaces*, **22** (1), 8–31.
14 Wirtz, J, Kimes, S E, Theng, J H P and Patterson, P (2003) 'Revenue management: resolving potential customer conflicts', *Journal of Revenue and Pricing Management*, **2** (3), 216–226.
15 Duneavy, H and Westermann, D (2004) 'Future of airline revenue management', *Journal of Revenue and Pricing Management*, **3** (4), 380–383.

16 Kimes, S E, *The '4-C' Strategy for Yield Management*. Centre for Hospitality Research at Cornell University, Ithaca, NY.

17 Kimes, S (2002) 'Tourism Revenue Management Programme', Stirling Management Centre, 27–29 November.

18 Ingold, A and Huyton, J R (1997) 'Yield management and the airline industry', in Yeoman, I and Ingold, A (eds), *Yield Management.* London: Cassell.

19 Gleick, J (1999) *Faster*. London: Abacus, p. 1.

20 Larson, R C (1987) 'Perspectives on queues: social justice and the psychology of queuing', *Operations Research*, **35** (6), 895–905.

21 Maister, D H (1985) 'The psychology of waiting lines', in Czepiel, J A, Solomon, M R and Suprenant, C F (eds), *The Service Encounter*. Lexington, MA: Lexington Books, D C Heath and Company.