11.2 What is capacity management?

Operations managers are concerned with ensuring that the service process has sufficient resources to deal with the anticipated levels of customer demand in such a way that quality of service meets pre-set targets in the most cost-effective manner. This is a particularly difficult task when managers are faced with very variable demand, not just in terms of volume but the variety of services required, as you can see in the New Zealand water taxi firm described in Case Example 11.1.

Capacity management is a delicate balancing act because both underutilised and overstretched resources can be disadvantageous. Underutilising resources has the potential to damage the long-term success of the organisation in a variety of ways:

- Expensive resources not earning revenue lead to poor financial results. The airline that fails to achieve a high load factor on its planes will struggle to survive.
- In many services, customers are suspicious of services that appear not to be busy. Banks and similar financial institutions find that customers are not happy about using an empty branch, and many diners prefer the 'buzz' of a busy restaurant.
- Service employees may become demotivated if underutilisation persists. Boredom and concern for their long-term employment may lead to poor service attitudes, which again lead to reduced customer satisfaction and lower profitability.

Conversely, resources that are overstretched also lead to problems for the success of the organisation:

- Overloaded resources mean that many aspects of service delivery suffer. A sudden surge of customers into a shop means that waiting times increase and staff cannot devote the amount of attention to customers that is desirable.
- Staff who are continually overloaded make more mistakes and, in the longer term, may decide to leave the organisation in search of less stressful employment.
- To deal with overload, staff may be drafted in to carry out tasks with which they are unfamiliar or for which they are only partially trained. The potential for increased error rate is high and, again, stress levels may be intolerable for some members of staff.

The task of capacity management is to try to achieve a balance between too much and too little resource utilisation, within the financial and operational constraints. Capacity management is concerned with putting a plan in place that makes the best use of resources to deal with the forecasted or expected demand for services.

Case Example 11.1 Pelorus Water Transport

The Marlborough Sounds covers an area of around 4,000 square kilometres at the north end of the South Island of New Zealand. The Sounds are a series of fjord-like water inlets connected to the Cook Strait which separates North and South Islands. The steep, wooded hills and small quiet bays of the Sounds are sparsely populated and access is difficult. Many of the small settlements and isolated houses and holiday homes are only accessible by boat. The main port of Picton sits at the foot of Queen Charlotte Sound. Serving the more isolated Pelorus and Kenepuru Sounds is the small town of Havelock, famous for its green-lipped mussels which are farmed in the Sounds.

John Beavon and Catherine Coates operate one of the few water taxi services from Havelock, Pelorus Water Transport. In what they describe as 'the best job in the world', John and Catherine ferry passengers and their small cargo around the Sounds. Passengers include residents with their shopping, holiday home occupiers and their belongings, backpackers and their rucksacks, builders working on the houses, contract workers, often loggers, and their smaller tools, and government environmental and building officers. They are sometimes involved in medical emergencies evacuating seriously ill people to the road in Havelock. They are also very happy to fill up a few places with tourists and take them along for the ride, doing their utmost to extend the trip for them and show them local features such as gannet and seal colonies and mussel farms or provide guided or unguided walks. They even serve afternoon tea on board!

The work is varied and as with any taxi not particularly predictable and so a degree of flexibility is required not only by the operators but also by their customers. John explained:

We try to get about 10 to 12 people on every trip, and we go out usually two or three times a day. The Sounds are very long and despite having a fast boat (a 24-foot custom-built 25-seater powered by twin 200-horsepower engines capable of cruising at 24 mph, 36 knots, which is pretty fast for water transport) it can take over an hour to reach the furthest point on the Pelorus Sound. Usually the first person to ring up and book the boat on a particular day has the say on what time we depart and where we go to. Then, as others ring or email, we will fit them in around that. We may also have to start making adjustments to the schedule to allow us time to pick the others up and load and unload their equipment. The vast majority are happy when we ring them up and ask them to adjust the times; I guess they realise it might be them we are trying to fit in next time. Obviously some timings are critical and can't change such as when we are bringing people back to Havelock to meet up with other transport.

The helpfulness of customers spills over to create a sense of camaraderie on board. The atmosphere feels more like a convivial bar than a taxi, with John and Catherine on first-name terms with many of their clients.

For exclusive use of the boat for a trip John and Catherine charge around NZ$274 per engine hour. If people are willing to share, as most are, they would aim to charge people 40, 60 or 75 dollars per head depending on distance travelled.

There is one other water taxi based in Havelock, and several other companies offering tours, such as the Mail Boat which covers three set routes three days a week. Access to the Sounds can also be by float plane, based in Picton and travelling between Wellington (in the North Island) and the Marlborough Sounds.

11.2.1 Defining service capacity

Before we discuss the methods available to try to balance capacity and demand we first need to define what capacity is and how it can be measured.

Service capacity is defined as the maximum level of value-added activity over a period of time that a service process can consistently achieve under normal operating conditions.[1] We can define and measure capacity relatively easily at the process level. For example:

- the number of calls a customer service agent can handle in the course of a shift
- the number of meals served by a restaurant during the lunchtime period
- the number of repair calls made by a computer service engineer during an eight-hour day.

It is important to note the words 'under normal operating conditions' and 'consistently'. It may be possible, in some cases, for an individual employee to exceed the throughput rate for a short period. If call centre employees handle 120 calls over 8 hours (15 calls per hour), it may be possible for them to achieve as many as 30 calls in one of these hours, but for this rate not to be sustainable over any length of time.

As we will see later, in Section 11.6, overloading resources may appear to increase output. Indeed, if analysed solely in terms of numbers of customer transactions completed in a given period, this may seem to be the case. However, there may be an impact on the nature of the service, the service concept and also the quality of the service provided. Service organisations must take particular care to ensure that the service concept is not changed in the search for

greater productivity. For example, a restaurant may decide to encourage customers to leave their table when they have completed their meal, in order to fit in a second sitting for dinner. On the face of it, the restaurant has doubled its capacity, but customers may feel that the nature of the service has changed and the level of service has deteriorated. Of course there are strategies that the restaurant can adopt to manage this sensitively, but the operations manager must be certain that in increasing productivity the desired service concept is maintained.

11.2.2 Measuring capacity

To manage capacity, we must be able to measure it. A simple measure is the amount of demand in a specified time period. The parcel delivery service may have an overall measure of capacity in terms of the total number of parcels that can be processed overnight; however, this overall measure is not very helpful in managing the day-to-day operation. It is necessary to develop a measure of capacity that is sufficiently detailed to give a 'good enough' estimate of capacity. For the parcel delivery company, some of the key aspects to be considered include:

- The size, weight and value of the parcels to be moved – a package that is small but valuable will provide more revenue per truck movement than something less valuable that takes up a large amount of space.
- The geographical locations served by the company – rural districts have greater travel times, and many inner city areas suffer from traffic congestion.

In determining capacity, a number of factors make the assessment of service capacity difficult:

- ***Service mix.*** If the service mix (the range of services provided) is made up of high volumes of 'runners' (see Chapter 8), the capacity calculation is relatively straightforward. However, once the service mix incorporates fluctuating volumes of 'repeaters' and 'strangers', the calculation becomes more complex. The customer service agent may be able to handle 120 'normal' calls in a shift; however, if some of the calls are complex enquiries or serious complaints, for example, the number of calls handled will drop significantly.
- ***The impact of location.*** At first sight, the measure of capacity for a computer service engineer or a telephone engineer would seem to be relatively straightforward. However, if we consider the difference between an engineer operating in a major city and another engineer dealing with rural communities, it can be seen that calculating capacity on the basis of calls completed alone would be inaccurate, since it would take no account of travel times.
- ***The extent of intangibility in the service.*** Services with low degrees of intangibility are relatively easy to deal with. The number of short transactions per hour in a fast-food restaurant is relatively consistent. However, the customer-facing staff in a gourmet restaurant may have greater discretion as to how to carry out their task. They may perhaps spend time with customers in 'building relationships', with the result that the individual's capacity becomes more difficult to define. This calculation becomes more complex when dealing with knowledge workers, who must combine short-term revenue-generating activities with long-term research and development. Capacity in this case is linked more to the individual service provider. It may be almost impossible for a manager to know when someone is working to capacity when output is so variable.
- ***The ease of identification of resource constraints.*** The capacity of a process is determined by resource constraints or bottlenecks. In the Karolinska Hospital example (see Case Example 11.3), it was clear that the resource constraint was the operating theatre. Finding ways to increase the effective utilisation of this space was relatively straightforward. For more complex systems, the identification of the key resource constraints may be rather harder. An information systems provider may require a wide range of technical skills relating to different applications and programming languages, but the precise requirements may not be known until part-way into the contract.

11.3 How can managers balance capacity and demand?

Major investments or disinvestments are required to deal with longer-term forecasted or anticipated fluctuations in demand. Alton Towers Resort (see Case Example 11.2) made significant investments every few years to install new or replacement rides to provide the capacity to deal with the increasing numbers of guests it wanted to attract to its theme park. It also invested in hotels and a water park to allow it to enter the year-round short break market. One key issue with such large-scale investments is that this capacity usually comes in large lumps, like a 180-room hotel, and so that capacity may not always be fully utilised when it is first installed/opened. Later in its life that capacity might become stretched and bookings might be turned away. As a result most operations managers spend a great deal of their time dealing with more short- to medium-term capacity management, trying to balance the day-to-day demand for their services with the available capacity. This section begins with a discussion of the longer-term capacity issues then focuses on short- to medium-term capacity management.

Case Example 11.2 Alton Towers Resort

Alton Towers Resort is Britain's answer to Disneyland. It is the UK's leading short break resort with a wide range of family entertainment. The resort, a division of Merlin Entertainments Group Limited, is perhaps best known for its Theme Park's white knuckle rides such as Thirteen, Air, Nemesis, Oblivion, Rita and Ripsaw. Oblivion, for example, is the world's first vertical drop roller coaster. The ride lasts 160 seconds and reaches speeds of up to 100 kph whilst pulling a G-force of 4.5 (astronauts only experience 3G at take-off!). Rita accelerates to 100 kph (62 mph) in 2.2 seconds. The Thirteen ride, opened in 2010, is billed as the scariest in the UK, combining physical and psychological fear. Some of the other less scary rides include Congo River Rapids, the Runaway Mine Train and Haunted Hollow. Opened in 2009 was Sharkbait Reef by Sea Life. This includes 'touch pools', where guests can interact with various underwater species, and a 10-metre ocean tunnel. In April 2010, a live webcam was installed to allow internet users to watch one of the tanks via the Alton Towers website. Younger guests can enjoy a range of special toddler and child play areas, and those who don't care for the rides or play areas can enjoy the gardens and floral displays, shops, restaurants and daily live shows. To help cope with the British weather about a third of all the rides are either indoors or in covered areas. The range of eating places includes McDonald's, KFC and Pizza Hut as well as Alton Towers' own branded family restaurants. Together they produce the 260 tonnes of chips and four million cans of drinks that the guests consume each year. While the Theme Park is open between March and November, the two hotels, Alton Towers Hotel and the Splash Landings Hotel with its water park, are open all year round.

There is parking for over 6,000 cars and 250 coaches at the Theme Park, which entertains around three million visitors a year. The cost of entry is currently around £30 for adults and £25 for children. Demand peaks

at about 50,000 visitors on Easter Bank Holiday Monday and usually runs at about 35,000 throughout the summer. The busiest times are usually during the week. Fridays and Saturdays during the peak season tend to be relatively quiet. The various activities in the Park reach peak demand at different times. The peak time at the gate is 10.30–11.00; for the restaurants it is 12.30–1.00. The major rides are very busy all day with queues reaching their longest in the early afternoon. Fast-track tickets can be obtained for the most popular rides, offering guests an allocated time slot. These tickets can be purchased in advance. The Theme Park employs around 350 full-time and 1,200 seasonal staff. The majority of staff live within a 20-mile radius and to help cope with unexpected fluctuations in demand a pool of staff is available for work at short notice. For flexibility most of the operators are trained to operate several rides.

Each year the company invests in new rides, attractions and infrastructure. For example, in 1994 £12 million was invested in Nemesis. In 1996 the 180-room Alton Towers Hotel was built at a cost of £20 million to allow the company to position itself as a major short break destination. More than 100,000 guests stayed in the hotel in its first year. In 1998 £12 million was invested in Oblivion. Air was built in 2002 at a cost of £12 million. Rita was built in 2005 for £8 million. Thirteen was added in 2010 at a cost of £15 million.

11.3.1 Long-term capacity management

There are four main considerations for making long-term capacity decisions: location, capacity, capability and resilience.

Location decisions

Location is the geographic positioning of a facility or facilities which is providing capacity. Location decisions are often expensive and may have a significant impact not only as an investment cost but also on operations costs, since location may be affected by local wage rates and business rates, for example. Location may also have an impact on revenues, particularly when the operation involves physical contact with customers. For operations that do not require direct physical contact with customers, for example call centres and internet-based service providers such as health or benefits advisory services, location decisions can be made to minimise the physical costs of the buildings and the running costs of the operation. For operations that need direct access to customers, expensive town-centre or out-of-town shopping malls may be essential.

Location decisions are a balancing act between supply-side factors and demand-side factors.[2] Supply-side factors are those that influence the costs and difficulties of a location decision. The demand-side factors are those that influence revenues. Not all the factors below will apply to every location decision but they are an indication of those factors that may need to be taken into account.

Supply-side factors include:

- land costs – the costs of acquiring the land
- labour costs – wage costs, employment taxes, welfare provisions etc.
- energy costs – the cost of energy or the availability or even the consistency of the supply of energy
- transportation costs – the costs of getting resources to the site and of transporting materials to customers
- government factors – local taxes, capital restrictions, financial assistance and political climate, and planning restrictions
- social factors – language and local amenities
- working environment – the history of labour relations and labour supply.

Demand-side factors include:

- convenience to customers – the site's accessibility for customers, including transport network, parking, distance from markets
- labour skills – the availability of particular talents, skills, accents and cultures
- characteristics of the site – the intrinsic and maybe aesthetic appeal of the site
- image – the reputation of the surrounding area and the extent to which there are complementary services in the vicinity.

Capacity decisions

Another key question is: how big should the facility be? For the package distribution operation, some estimate of volumes to be sorted in a relatively short time window each night will be required. Likewise, when deciding the size of a supermarket, call centre, airport, surgery or cinema, the costs need to be weighed against forecast demand – not only short-term demand, but also long-term demand, because the cost of changing facility size can be expensive and sometimes difficult.

The two interrelated issues for operations managers are:

- Facilities can usually only be added in large – and expensive – chunks.
- Capacity needs to match demand.

Adding new facilities usually requires the organisation to commit significant amounts of capital. This can be a risky business because long-term demand can rarely be predicted with any great certainty. If necessary break-even volumes are not met, the facility will not pay for itself. In some cases, such as a theme park where having sufficient customers creates the atmosphere, the service may not be as good as it should be if volume targets are not achieved. If volumes are exceeded, there may be significant localised problems for customers, resulting in customer dissatisfaction and lost business. Given that the majority of forecasts will be wrong (because they are only forecasts), operations managers will invariably suffer from the consequences of over- or under-capacity.

Many airports have suffered from the latter problem; furthermore, owing to the length of time it takes to design a new runway or terminal building, and to go through the planning process and build the facility, volumes may again exceed capacity as soon as the new facility is opened.

As with short- and medium-term capacity management (see Section 11.3.2), there are three main strategies for long-term capacity planning:

- ***Plan to exceed demand forecasts.*** This strategy is appropriate where there is an expanding market or the cost of building a new facility is inexpensive compared to the cost of, or problems that would be created by, running out, such as electricity or water supply, or air traffic control facilities.
- ***Build to forecast.*** This approach would balance the likelihood of not having enough and having too much capacity and is appropriate where the costs and consequences of exceeding demand are similar to those for not meeting demand.
- ***Plan not to meet forecast demand.*** This is an appropriate strategy where it is acceptable not to meet demand or where the cost of capital is very high compared to the costs and consequences of not meeting demand. Football clubs may be able to do this, using price premiums and revenues from television companies to balance the books and even set money aside for future expansion. The problem for some organisations that follow this strategy, such as supermarkets, is that it might give the competition time and income to pursue an aggressive expansionist strategy.

Developing a facility strategy involves steps that are easy to describe but difficult to implement:

- establish a measure of capacity
- develop demand forecasts, ideally several forecasts including optimistic and pessimistic ones, identifying the assumptions on which each is based
- identify alternative means of dealing with the forecasts
- undertake an assessment of the risk involved
- evaluate the alternatives.

Capability

It seems obvious that any new facility should be capable of doing what is required, but this is not as easy as it sounds. There are some airports whose runways are too short to accommodate some of the larger aircraft. A decision taken years ago on the length of a runway when planes were smaller creates constraints on operations now. Doctors' surgeries, too, have changed significantly over the last few years, as doctors form larger practices to share growing administrative costs and ease the burden of 24-hour cover. Surgeries also provide many more facilities than previously, such as well-person clinics and routine surgical operations, for example, putting stresses on facilities designed for a different way of working.

A key problem we face is in forecasting both demand and also the nature of that demand and thus the nature of the services that have to be provided in the future. It is little wonder that many operations management problems stem from the size and nature of the facilities available.

Resilience or flexibility

Although forecasting the size and nature of demand and future services is difficult, if not impossible, the only thing an operations manager can do, apart from keeping their finger on the industry's pulse, is to try to ensure that their facilities have some degree of resilience or flexibility.

Physical resilience can be created through either structural flexibility or developing the potential of the infrastructure. Building flexibility or resilience into a facility can be done in many ways:

- buying extra land to facilitate any possible future expansion
- having a flexible internal structure, with open-plan offices and movable walls
- using flexible equipment, such as cordless telephones or desk-sharing schemes
- adopting different methods of working, such as using more home-based workers
- developing contingencies – railway companies, for example, may plan to use different routes if one route fails.

11.3.2 Short- to medium-term capacity strategies

There are three basic short/medium-term capacity strategies, although, as we will discuss, many organisations employ a mixture of all three (see Case Examples 11.1 and 11.2). These strategies are:

- *Level capacity.* In this case scarce or expensive resources are maintained at a constant level, and the organisation must manage the consequential issues for customer satisfaction and operational service quality.
- *Chase capacity.* The service organisation attempts to match supply to demand as much as possible by building flexibility into the operation. The prime objective is to provide high levels of service availability or fast response, in the most efficient manner.

- ***Demand management.*** Rather than change the capacity of the service operation, the organisation influences the demand profile to 'smooth' the load on the resources.

Level capacity strategy

The prime objective of this strategy is to maximise utilisation of expensive fixed resources. An airline seeks to fly planes that are as full as possible with passengers paying the highest fares. The key operational measure is the 'load factor', with the airline knowing that if it is exceeding a certain figure (about 80 per cent for an international airline), it will be making profit.

To achieve this level of utilisation, the service organisation may have to make a number of trade-offs, most notably around customer perceived quality of service. Figure 11.1 illustrates the situation in a hospital clinic. Here, the task is to make the most of the medical consultant. The clinic has to solve the problem of always having enough patients for the consultant to work on, with the added difficulty of there being a high percentage of 'no-shows'. The clinic has chosen to overbook appointments, and believes that it is better to upset a few patients rather than lose valuable consultant time.

To deal with the no-shows problem, the clinic has made four appointments at the start of each 15-minute period, estimating that one in four patients do not arrive and that the consultant will require 5 minutes per patient. If all goes to plan, the first patient will be seen immediately, the second within 5 minutes and the third within 10 minutes, but it should be noted that they each have the same 2.00 p.m. appointment. In practice, some of the 2.00 p.m. appointments will still be waiting when the next 'batch' arrives for a (supposed) 2.15 appointment.

Some general principles and issues can be drawn from this example about the level capacity strategy:

- Resource utilisation goals are frequently achieved at the expense of customer satisfaction.
- Customers may receive inconsistent service levels (those with 2.00 p.m. appointments fare better than those with later appointments).
- Customers (patients) accept (or suffer) this poor level of service because the service is valuable to them and there may be no or few alternatives.
- There is a danger that the service provider may become complacent and not make attempts to cut the emotional cost of waiting for the customer, making it potentially vulnerable to competition (in this case private healthcare).

To overcome this problem of variable service levels, the service organisation may use yield management (see Section 11.7) or queue management approaches.

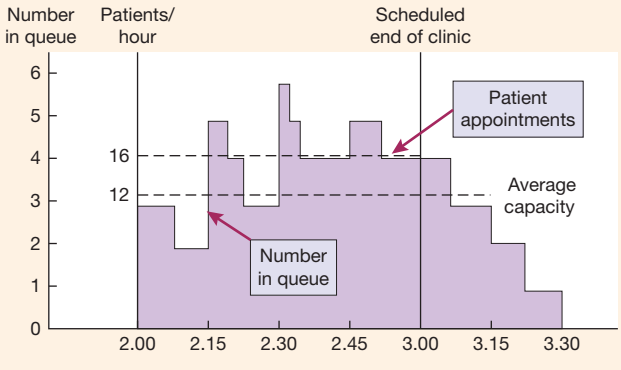

Figure 11.1 Level capacity strategy in an outpatients clinic

Examples of organisations that use the level capacity strategy as their dominant approach include:

- Airlines, which need to maximise the revenue from their most expensive resource (planes). The prime objective is to have planes flying as frequently as possible, preferably full of passengers. This may mean that passengers do not always receive the service they anticipated.
- Professional services, which may have a recognised expert in a specialised field. It is frequently the case that the overall workload will not sustain another professional, leaving clients with a choice as to whether they wait or find an alternative provider.
- Popular restaurants, which may intentionally not expand capacity in order to maintain exclusivity. Having to book days, sometimes months, ahead in order to ensure a table may enhance the service concept.

Examples of approaches adopted under the level capacity strategy include:

- ***Promoting off-peak demand.*** This is often combined with a pricing strategy to encourage customers to switch. The organisation must be careful that this does not bring about a change in service concept. A restaurant encouraging customers to move to less popular times may institute a 'happy hour' with cheap drinks, which may damage the restaurant's reputation with existing customers.
- ***Queue management.*** This is dealt with in more depth in Section 11.5, but it is important to point out here that making an assumption that customers will continue to queue can be dangerous. It is, after all, sending the message that their time is relatively worthless, and they are only prepared to wait because they anticipate that the service they receive will be valuable enough to make the wait (lost time) worthwhile.
- ***Booking systems.*** Making forward bookings is a form of queue management. It allows the organisation to schedule capacity ahead, and for customers to utilise queuing time for themselves. Supermarkets have successfully utilised this system for their delicatessen counters, issuing customers with numbered tickets to ensure that people are served in order, and allowing customers to judge whether they have time to continue shopping before their number is called. As with the physical queue, customers may not want to wait and so may go elsewhere. Indeed, if the organisation has the reputation that customers need to book ahead, it may lose potential sales if customers assume that there is no point in trying.

Chase capacity strategy

This strategy is usually adopted by high-volume consumer services, since a major aspect of their competitive strategy is the provision of ready and rapid access to service. For these services, capital resource utilisation is rarely a prime goal, although cost reduction will be very important. To explain this further, consider the following statements concerning a fast-food restaurant:

- A key objective is to maintain short queue lengths. This is managed by staffing tills and kitchen in line with expected demand.
- If the queues are too long, customers go to another fast-food outlet.
- The premises are not fully utilised: there are only about six hours out of the possible 24 hours when the facilities are 100 per cent utilised.

The challenge of these high-volume standard services is to develop volume flexibility (see Section 11.7.2). In other words, the operation must be able to cope with wide ranges of customer demand, providing consistent service standards at minimum cost. Figure 11.2 shows the demand pattern for a fast-food restaurant, with a crew roster to show how the restaurant manager schedules the staff to deal with the variation in demand.

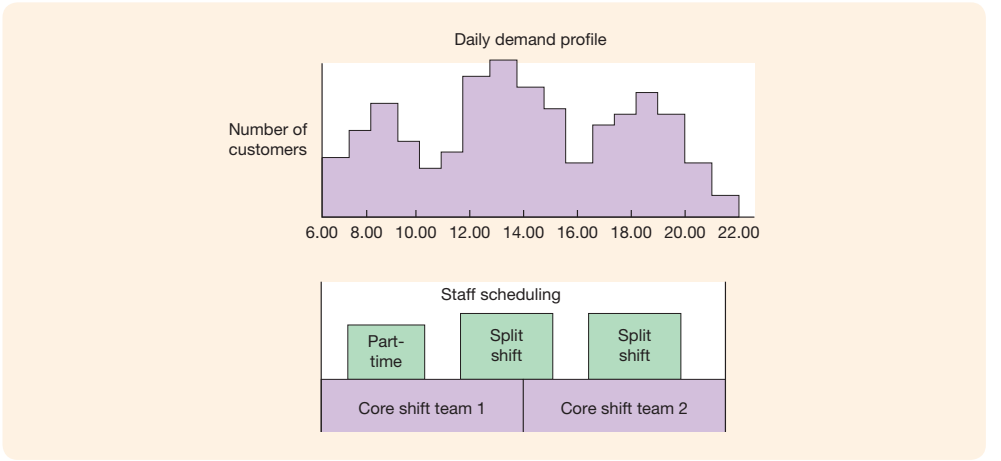

Figure 11.2 Chase capacity strategy for a fast-food restaurant

In this restaurant, the staff are organised into three categories: those on one of the core shift teams, those working split shifts, and those working part-time. The split shifts allow the manager to schedule staff for the forecast demand peaks. There will probably be a weekly rotation between the core shifts and the split-shift personnel. In this example, the early morning peak load is covered by employing part-time labour. In addition, the manager will have a pool of labour to be contacted at short notice to cover absenteeism or an unexpected rise in demand. A common strategy is to extend the length of the split shifts, with some organisations operating a 'compulsory overtime' policy as part of their conditions of employment.

General principles and issues for the chase capacity strategy are:

- Most organisations operating the chase strategy must develop a high degree of volume flexibility. In other words they must be able to respond to changing demand profiles. In most cases this is achieved through employing staff on flexible contracts, allowing the operations manager to decide working hours as required.
- Although a principal objective is to ensure that customer service targets such as availability or response times are achieved, many of these service organisations fall into the commodity category (high volume/low process variety). In contrast to the organisations employing the level capacity strategy, they frequently have relatively little means of differentiation and are therefore rather price sensitive. The challenge in adopting the chase capacity strategy, therefore, is to ensure that costs are strictly controlled and that flexibility is not achieved at any price.

Examples of organisations that employ the chase capacity strategy include:

- Retailers that need to deal with extremely high demand at weekends and after normal office hours.
- Direct insurance companies operating extended hours through call centres.
- Theme parks, which may open up more attractions as demand grows.

Typical approaches to the chase capacity strategy include:

- ***Flexible staffing levels.*** Some organisations use flexible employment contracts, allowing the operations manager to decide when staff will be working. In some cases staff will work a standard core time, but in many retail organisations staff may not know when they will be working beyond the next few shifts. Another approach is to employ part-time staff who must work 'compulsory' overtime as and when needed. Although this gives flexibility, the operations manager must be aware of the possibility of staff resentment at having to work inconvenient hours, and the knock-on effect of poor customer service.

- ***The use of subcontractors or temporary staff.*** Organisations may use temporary staff or subcontractors to deal with short-term overloads. Although these workers may be readily available, they may not be sufficiently trained or motivated to deliver service in the style of the organisation. However, some organisations report that their temporary staff may be more responsive and less complacent than long-service staff. Some call centres use organisations that specialise in what is called 'peak lopping'. Excess calls are automatically routed to the organisation and they are answered in such a manner that customers are unaware of the switch.
- ***Making use of customers.*** Many service operations may have the option of changing the service process to utilise customers as temporary employees. In effect this is again changing the service concept. Some regular customers may be very happy to be included in the service process – clearing tables, or even serving other customers as well as themselves.

Demand management strategy

Most organisations operate a mixed approach to capacity management. Whether adopting a principally chase or level capacity strategy, most service organisations also operate some degree of demand management. Examples of this approach include:

- ***Pricing strategies.*** This typically takes the form of offering price incentives to encourage customers to move to off-peak times. The 'happy hour' in the pub or wine bar is a good example.
- ***Restricted service at peak times.*** The philosophy here is similar, though taking the form of a disincentive. In this case the organisation may provide a limited service at peak times, again encouraging customers to move to less busy times. Some restaurants operate this policy, providing a limited menu at these times.
- ***Specialist service channels.*** Rather than provide a general service at all times, the provider may choose to segment the demand and to allocate specific times for special needs. Doctors' surgeries are a good example, with advertised times for services such as immunisations, mother and baby clinics, and counselling provision. This allows the surgery to schedule specialist resources to restricted times, often making better use of scarce resources.
- ***Advertising and promotion.*** Increasing public awareness of the service and informing customers of special offers will stimulate demand. Bookshops not only advertise, but will also stimulate demand by arranging sessions for authors to autograph their works. A particular problem with advertising is that it tends to increase the inaccuracy of any forecasting model used by the business. Although it is possible to track the effectiveness of advertising in stimulating demand, it is often difficult to pre-judge the likely impact of a new campaign.

Putting the strategies together

Most complex service organisations use all three of the capacity strategies in different parts of their operations, depending on the respective underlying cost models. Some examples are shown in Table 11.1.

The prime objective of the airline is to maximise 'load factor' on flights, the utilisation of its most expensive assets. It employs a number of strategies to ensure that it makes the maximum revenue on each flight, using sophisticated yield management techniques (see Section 11.7.1) to help adopt the optimum pricing strategy to sell unsold seats as departure time approaches. The airline may simply oversell seats in the belief that there will be a number of 'no-shows'. Passengers who are then not able to obtain the seat they thought they had booked need to be compensated in some way, although unless managed well that can significantly affect customer satisfaction.

To maximise the opportunity for customers to book seats, the airline employs a chase strategy in its sales department using relatively cheap resource (as compared with planes), scheduling staff to meet forecast demand patterns. It is better to suffer slightly reduced productivity here rather than lose potential seat revenue.

Table 11.1 Capacity strategies

	Level capacity strategy	Chase capacity strategy	Demand management strategy
International airline	Ensure that planes are flying with maximum payload as frequently as possible	Schedule staff reservations department to meet demand to ensure bookings can be made	Promote off-peak demand Try to maximise revenue from each flight (yield management)
Insurance company	Protect back-office experts (actuaries and investment specialists) from variations in customer demand	Schedule direct sales operation (call centre) to provide maximum access for customers	Influence selling cycle so as not to coincide with policy renewal peaks
Restaurant chain	Keep manufacturing of basic food materials as close to 'level' as possible Maintain high utilisation of process plant	Draw up staff rosters to reflect anticipated demand Use part-time staff to manage peaks Call in staff for demand surges	Use promotional activity to stimulate demand in quiet periods Devise special offers to allow for bulk-purchasing discounts

The insurance company uses a level strategy for its actuarial staff (back office), in part because they are relatively expensive, but more because they are often in short supply. The lead time to recruit and train an actuary is measured in years rather than weeks and therefore it makes no sense to attempt to chase demand.

Similarly the restaurant chain will operate a level strategy in its manufacturing function because it has relatively fixed capacity: although it can increase capacity marginally by overtime, significant increases can only be achieved by investing in another kitchen or restaurant.

11.4 How is day-to-day planning and control carried out?

In the previous section we discussed how operation managers decide a capacity strategy, or mix of strategies, thus creating a capacity plan. This sets the broad parameters within which this capacity may be allocated to specific customers and/or tasks in order to meet customer service and productivity targets. This section looks at the mechanisms that may be deployed to micro-manage this plan as effectively as possible. Day-to-day operations planning is concerned with creating a 'schedule' or timetable (often a daily or weekly schedule), based on the capacity plan, which

- allocates staff, customers, equipment and/or facilities to activities (often referred to as loading)
- decides what order things will be done in, e.g. which customers/orders to deal with first (sequencing)
- shows what time each activity will start and finish (scheduling).

Examples of 'schedules' include:

- A table plan for a restaurant – as bookings are received, tables are allocated, giving an instant picture of the loading for any given time period, and the likely start and finishing times. This implies that bookings will be accepted on a first come, first served basis and that customers without reservations may have to be turned away.
- A school timetable that shows where every student, and member of staff, should be at any time of the day, what they should be doing and when things take place.
- An appointment book at a car servicing workshop, providing space for a given number of standard services and more complex jobs, showing who will do which job and roughly how long they are expected to take.

Operations control is concerned with making adjustments to cope with changes as they happen.

11.4.1 Creating a schedule

Most operations have rules or policies regarding the allocation of capacity. Sometimes these rules are relatively informal, developed over time in such a way that most customers are satisfied. Other operations, usually those with more volume and/or complexity, tend to have more formal allocation systems. For example, most leisure centres have one large space or sports hall, which can be used for a number of activities. It might be configured for two five-a-side football pitches, four badminton courts or one tennis court. To create the schedule or timetable, the leisure centre manager must carry out the following tasks:

- Decide the proportion of time that the sports hall will be configured for each of the activities.
- Ascertain the optimal schedule for these activities based on customer preference. For example, badminton clubs may prefer Thursday evenings while five-a-side football may be more popular on Fridays.
- Check the schedule for ease of transition between activities. It may be relatively easy to move from tennis to circuit training because there are relatively few changes to make, but changing from badminton to football might require more staff to move and set up equipment.
- Create a booking schedule for the various slots to allocate capacity (times) to specific customers.

This schedule is required because the leisure centre does not have infinite resources. It is clearly impossible to satisfy all possible customer demands and remain a viable service organisation.

The way that operations deal with this issue is to create sequencing rules in order to manage the prioritisation of allocation. Here are six examples of sequencing rules:

- ***First in, first out (FIFO) or first come, first served.*** This is the approach used by many consumer services. The leisure centre may allow customers to book up to two weeks ahead for badminton for the Thursday evening time slots of 6.00–7.00 p.m., 7.00–8.00 p.m. and 8.00–9.00 p.m. and will take requests in order until all the courts are full. This scheduling rule is simple and has the advantage of being perceived to be fair to all potential customers.
- ***Last in, first out (LIFO).*** There are rare occasions when this rule makes sense. The sequence of loading a delivery truck would follow LIFO scheduling so that the first call would be to deliver the products or materials closest to the tailgate of the lorry.
- ***Most valuable customer first.*** The leisure centre may allow the local football club to have early access to certain parts of the forward schedule and to 'block book' capacity for training. If the club needs to have the use of the hall on every Friday for training, this represents steady income for the leisure centre, though it may be a source of annoyance for other, smaller organisations that are excluded from using the hall for single occasions.
- ***Most critical first.*** Emergency services adopt variants of this rule, grading the nature of each demand between critical and non-essential. Clearly, life-threatening situations call for immediate action and these demands tend to override all other activities.
- ***Least work content first.*** In situations where demand far outstrips supply, this rule allows more customers to be satisfied quickly. Airline check-in desks sometimes operate a version of this rule, providing faster service for those passengers with hand baggage only. This means that the total queue is reduced and fewer customers wait. The problem with this rule is that some customers wait far longer than they would under FIFO and are potentially more dissatisfied.
- ***Most work content first.*** With activities that have long process lead times it might seem prudent to schedule the tasks with the longest time requirement first. This may not always be helpful because, once started, these tasks may become lost in among other 'work-in-progress' and not progress as fast as they could. Using this rule alone does not usually produce the best performance against customer requirement dates.

Different rules may be applied in different circumstances. Where customers arrive together, they generally expect to be treated equally and therefore FIFO will tend to be applied. However, this is not always the case. At the Alton Towers theme park in the UK, it is possible to buy fast-track tickets that enable customers to bypass the queues of customers who pay standard prices. This obvious privilege can cause high degrees of ill feeling among 'standard' customers.

Back-office processes frequently deal with longer lead times than front-office processes, and may apply a number of scheduling rules. In the more complex situations it is possible to use a simple algorithm similar to that used in manufacturing shop floor control, called critical factor calculations. In this approach a comparison of estimated work content against requirement date allows for a calculation of an urgency factor. Each stage in the process is provided with a list of the most urgent jobs, enabling prioritisation. A spin-off from these calculations is a spot check of the performance of the process. If all tasks fall into the 'most urgent' category, clearly significant action is required to avoid major customer dissatisfaction.

11.4.2 Operational control

Systems for control, i.e. to enable the schedule to be adjusted, range from comprehensive, complex and expensive to extremely simple.

At the complex and expensive end of the spectrum are the ERP (enterprise resource planning) systems. These software-based systems, sold by companies such as SAP and Baan, offer the capability to integrate a number of functions across the organisation. For example, sales order processing systems can provide direct input into operations control, and then into supplier management/procurement systems.

At the other extreme, many control systems are basic, but effective nonetheless. Examples are the restaurant table plan, the school timetable and the appointment book, which act as both schedules and control systems. All of these 'systems' can be examined and assessed to see if, and how, changes can be accommodated. If a teacher reports in ill, the 'manager' will have to assess which lessons need to be covered and who might be free to cover them. A restaurant guest who extends the size of their party may be accommodated on a different table; the extra car service for an important client may be added as overtime at the end of the day. In all cases the existence of a detailed schedule/control system is critical to managing the day-to-day planning *and* control of the operation.

Having an effective control system can be a significant source of competitive advantage. The ability of the organisation to provide an immediate response to the question 'When can you do this?' is a major factor in building customer confidence. In recent years, retailers have been able to interrogate logistics systems so that customers can negotiate a delivery slot for their purchase of furniture or electrical products, with some confidence that this promise will be fulfilled.

The Karolinska Hospital operating theatre (see Case Example 11.3) provides a good example of the benefits of a good schedule/control system. Such a system should include:

- ***A clear customer flow.*** The schedule provides the opportunity for customers to be 'flowed' through the system to arrive at the right time and place. We might compare customers in this way to the work-in-progress (WIP) inventory in a manufacturing process. If WIP is minimised, customer delays are reduced, there is less disruption, and costs are kept to a minimum. An organisation that encourages customers to arrive early because it is unable to manage its schedules may have to invest in larger reception (customer-holding) areas, or suffer a major decline in customer satisfaction.
- ***Ensuring supporting resources are available to meet the schedule.*** Once the schedule has been established making best use of the availability of scarce resource, all other resources must be scheduled to meet this plan. For the hospital operating theatre this will involve the scheduling of people (operating theatre staff and surgeons), and of course physical inventories such as blood, bandages and surgical instruments. Restaurants, likewise, will match inventories of food to expected customer demand.

- *Creating schedules for interlinking activities.* The schedule allows the operations manager to create a realistic plan for the service as a whole and its many interlinking activities. In some organisations this is termed the master schedule. This then facilitates the production of supporting schedules for all the resources that feed this master schedule, using a variety of scheduling rules matched to each situation.
- *Creating schedules for suppliers.* Good information provides a sound basis for negotiation with key suppliers. This gives rise to the potential for managing day-to-day activity more accurately with less waste (see Chapter 12).

11.4.3 Managing short-term schedules and medium- and long-term capacity plans

The detail of schedules will increase as 'time now' approaches customer requirement dates. To return to the leisure centre example, tomorrow's schedule will tell us exactly which customers have booked a badminton court, and at what time. Moving further forward in time to, say, next month, the manager will be interested only in which evenings are allocated to specific activities which can then be booked by specific individuals as that week approaches. In the much longer term, say a year, the manager will need to know the size of the facility available and any plans to extend the sports hall.

It is essential to know the time period ahead within which it is impossible to change the schedule. A restaurant may be able to deal with changes in customer mix almost up to the last minute, whereas the operating theatre in Case Example 11.3 will require more notice of change in detail schedule. We will return to this topic later in this chapter when we consider resource flexibility, which aims to reduce the period ahead where the schedule must be fixed.

Case Example 11.3 Karolinska Hospital, Stockholm

Karolinska Hospital faced a crisis as the pressure on operating budgets was rising. This prompted an investigation as to how well expensive resources were being utilised. It was soon identified that operating theatres were not being used effectively. In fact, surgeons, operating theatre staff and of course the theatres themselves were idle for more than 50 per cent of the time. It soon became clear that the schedule of patients through the theatre needed to be managed more carefully. The scarce resource was time in the theatre itself, so the management looked at ways to reduce the time that patients spent in the theatre. A significant step forward created a separate patient preparation area allowing this activity to be carried out in parallel rather than in sequence with surgery.

Further investigation revealed that some delays were caused because anaesthetists were called away to other parts of the hospital. Adding anaesthetists formally to the operating room staff team and creating an anaesthesia clinic to evaluate pre-operative patients also improved the efficiency of the system.

Once the throughput through the bottleneck had been increased, more operations could be carried out in the same

Source: Pearson Education Ltd/Corbis/BrandX

timeframe, and waiting times were dramatically reduced. The unforeseen benefit was that it was now possible to create a much more reliable schedule because the lead time between diagnosis and surgery was dramatically reduced. Patients were happier because they were treated faster and there were fewer 'no-shows' than there had been when lead times were longer. A consistent theatre schedule meant that any tests required prior to surgery could be arranged with more certainty. Previously, these tests had been frequently repeated as surgery dates had been delayed.

For Karolinska, then, this approach to scheduling has paid off in a number of ways. The theatre carries out more operations per day, costs are reduced and patients are seen more quickly. Operating rooms were reduced from 15 to 13, while the number of operations per day was increased by 30 per cent.

Source: This illustration is based on material from the video *Time Based Competition* from Harvard Business School (1993), and from healthcare industry sources.

11.5 How do organisations manage bottlenecks and queues?

Bottlenecks (the parts of a process that constrain or restrict capacity), and the resulting queues of customers or their orders, are features of many service operations. Managing these well can have a significant impact on both capacity utilisation and customer satisfaction.

11.5.1 Bottleneck management

All organisations need to understand their key resource constraints. A clear understanding of these constraints or bottlenecks provides greater clarity as to what is a realistic estimate of capacity. To return to our example in Section 11.4, the key resource constraint of the leisure centre is likely to be the central sports hall. All other resources, such as other facilities and staff, are likely to be linked to the most effective use of this space.

Bottleneck management, or the theory of constraints, is generally well understood in manufacturing organisations. It is seen to be important to manage the bottleneck – the stage in any process with the lowest throughput rate and which therefore determines the effective capacity of the whole operation.[3]

In the same way, it is important for service operations managers to understand where the bottlenecks exist in service processes. For example, a company providing loan finance needed to increase the standard of service provided to its customers, while also increasing productivity of its risk assessment process. The management was given the task of meeting increasing demand without increasing resources. Initially, it was not clear how this could be achieved, but when the process was mapped it became obvious that a problem lay with the actuaries. Figure 11.3a shows a simplified version of the original process flow, with all proposals passing through an actuary's hands for sign-off.

As can be seen from Figure 11.3a, the original capacity was constrained by the throughput of the actuaries at 15 proposals per hour. It was recognised that many proposals did not require actuarial sign-off because the credit scores indicated a clear accept or reject decision, which could be taken by more junior, less expensive staff. The initial processing took slightly longer, but in the revised process (Figure 11.3b) only 50 per cent of the proposals needed to be seen by an actuary. The capacity of the process therefore rose by nearly 50 per cent to 22 proposals per hour.

This improvement was achieved simply by monitoring the activity levels within the process and then deciding whether the current resource constraints were really bottlenecks or not.

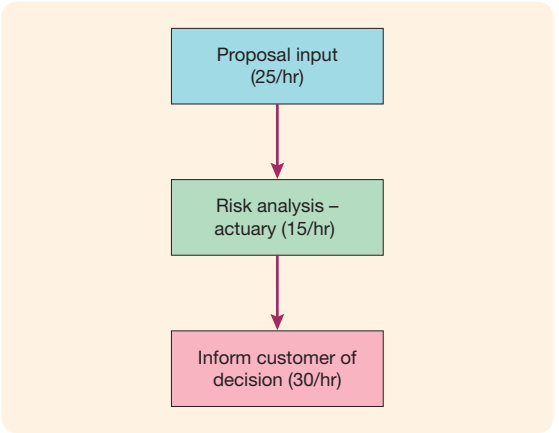

Figure 11.3a Original flow

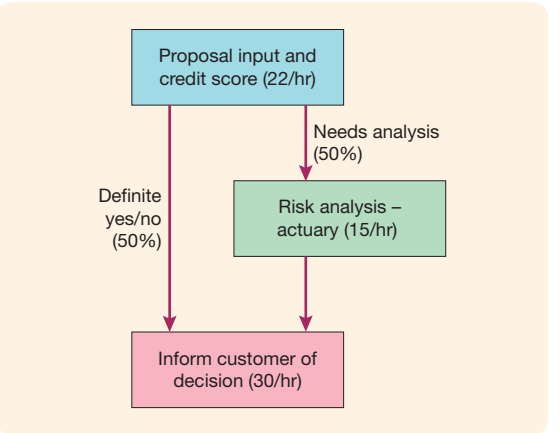

Figure 11.3b Revised flow

Once the assumption that all proposals must be seen by actuaries had been questioned, it became possible to improve response times and productivity simultaneously.

There are some general rules for managing bottlenecks:

- Ensure that only essential work passes through the bottleneck.
- Be ruthless in taking away non-essential activities from the bottleneck.
- Ensure that no substandard work passes through the bottleneck.
- Once you have established where the bottleneck is, devote proportionally more management attention to it to ensure maximum throughput and therefore maximum effectiveness for the process.

And finally, if you have a complex system, the best thing to do is not to try to move the bottleneck. It may be difficult to manage but at least you know where it is![4]

One way to identify a bottleneck is to observe where the queues of work or customers form in the process. In simple processes, this is probably as good a test as any, but it is as well to be wary. Queues may form much earlier in the process because people operating feeder activities may work at a slower rate than is theoretically possible because they believe there is no point in working flat out if the customer or the jobs will have to wait at the subsequent bottleneck. Indeed, service operations might well decide to limit the number of customers accepted into the process because of the potential dissatisfaction caused for customers who have to wait much longer than anticipated.

It is also important to distinguish between long-term and short-term bottlenecks. Long-term or fixed bottlenecks provide the best estimate of the capacity of the operation, and the basic capacity management approach can be determined as a result. However, short-term bottlenecks may frequently occur, giving rise to the need for immediate response. For example, a key member of a call centre may be unexpectedly absent, which may mean that capacity to deal with particular enquiries is dramatically reduced.

11.5.2 Queue management

Queues occur in most service activities. Indeed, for any operation using a level capacity strategy queues are 'designed-in'. Furthermore, no capacity strategy is perfect and queues are almost inevitable. Queues may be lines of people visible to both the customer and employee, or they may be invisible to one and/or the other, as with a queue of callers to a switchboard or a list of customers awaiting a repair engineer.

While queuing theory can be used to calculate the number of servers required to meet forecast demand, resource constraints and forecast inaccuracy invariably mean that operations managers need to look for other ways to minimise the impact of queuing on their customers. It has been shown that not only dissatisfaction with the wait increases with waiting time[5] but also dissatisfaction with the service as a whole.[6]

Given that perceived waiting time is usually greater than actual waiting time,[7] the answer is to try to reduce perceived waiting time, which can also be a great deal cheaper than employing more servers! Ten principles of waiting have been suggested:[8]

1. ***Unoccupied time feels longer than occupied time.*** It is a good idea to try to provide customers with something to do or forms of distraction so that the time passes more rapidly for them. Some services show promotional videos to people waiting in a physical queue. Waiting areas for lifts often have mirrors to enable customers to check their appearance. Telephone call centres or helpdesks frequently play music while 'on hold', although this is not universally welcomed.
2. ***Pre-process waits feel longer than in-process waits.*** Once customers feel that they have made a start inside the service process and that something, however trivial, is happening, they tend to feel happier. A simple acknowledgement by a server that they have been noticed can have a significant impact. Also, using pre-process time in some way, such as completing a form or making choices about the service, can reduce the perceived waiting time.
3. ***Anxiety makes the wait seem longer.*** Sometimes customers do not know whether they have been forgotten or not, which can be allayed by giving them numbered tickets to demonstrate that they are part of the system. Also, the nature of the service will have a significant impact. If the customer is worried about flying or going to the dentist the wait may seem interminable, possibly giving rise to some tense behaviour with service providers. Customer-facing employees should be trained to observe the effects of anxiety and to find ways of giving reassurance.
4. ***Uncertain waits feel longer than known, finite waits.*** Customers are generally more happy to wait if the expected duration is known, and if there is a good reason for it. If the duration is unknown, research suggests that customers become restless much more quickly. Theme parks frequently position markers at known points in the queue informing customers how long they should expect to wait. Of course, the real wait time is usually a little shorter than this, with customers pleased that they did better than expected!
5. ***Unexplained waits seem longer than explained waits.*** Being provided with a plausible explanation of a delay reduces uncertainty for the customer. It also gives the impression that the organisation knows it should not take the customer for granted.
6. ***Unfair waits seem longer than equitable waits.*** Generally, customers expect that those who arrive first should be seen first. Many organisations have replaced the multiple-queue/multiple-server system with a single-queue/multiple-server approach because of the perceived unfairness of being stuck in a slow-moving queue. This approach also eliminates the anxiety as to which queue to join. In some cases, such as a hospital casualty department, there may be a good reason why some customers are seen out of turn, but it still seems to be necessary for there to be an explanation rather than for the provider to assume that other customers will understand.
7. ***The more valuable the service, the longer customers will wait.*** The more complex the service, and the more it is customised to the needs of the individual, the more likely it is that customers may be prepared to wait. It should be noted, however, that this should not be assumed.
8. ***Solo waiting feels longer than group waiting.*** The realisation that others are also feeling the pain may reduce the customer's anxiety of thinking that they have made the wrong choice. If others think it is worth waiting, it confirms the customer's decision to wait. Also, people tend to talk to each other, providing a distraction from the length of the wait.

9 ***Uncomfortable waits feel longer than comfortable waits.*** By making queuing conditions as comfortable and indeed as distracting as possible, the wait time will be perceived to be much shorter. Uncomfortable conditions sensitise customers to the time and poor service.

10 ***New or infrequent users experience their wait as longer than frequent users do.*** Frequent users of a service may be attuned to a wait and they may be more relaxed because they know what to expect. New or infrequent users are likely to be more anxious and uncertain, so operations should consider trying to identify them and provide them with information and reassurance.

A booking system is a queue, with the advantage to customers that they do not have to physically queue for the service. The advantage for the service provider is that the operations manager is better able to manage resources to meet demand. Alton Towers (Case Example 11.2) has in effect created a 'virtual queue' through its fast-track system. Visitors to the park are given time slots to return to a pre-booked ride which allows them to use their time more effectively. Supermarkets that operate a ticket system at their delicatessen counters are using the same principle. In both cases, the service provider has found a way to ensure equity of treatment for its customers and has enabled customers to make better use of time otherwise spent queuing.

11.5.3 Queuing theory and simulation

Management scientists and mathematicians have studied the behaviour of queues, producing statistical models to predict queue length and so on. Fortunately, few of us need to understand the detail of these models, as there are a number of computer simulations available to help us predict the implications of operational decisions.

Simply put, there are three key parameters to queuing theory: the arrival rate for customers, the server rate, and the number of servers or serving positions available. The arrival rate and server rate must be further understood in terms of their variability. Even if the average server rate and arrival rate are the same, queues will still form if there is variability in these rates. This situation is illustrated in Figure 11.4, which demonstrates the impact of variability in arrival rates and process rates. In this case, the process time is reasonably constant – perhaps because it is a standard process, possibly determined to a large extent by automation or standard scripts – while customers arrive in a more random pattern. In this

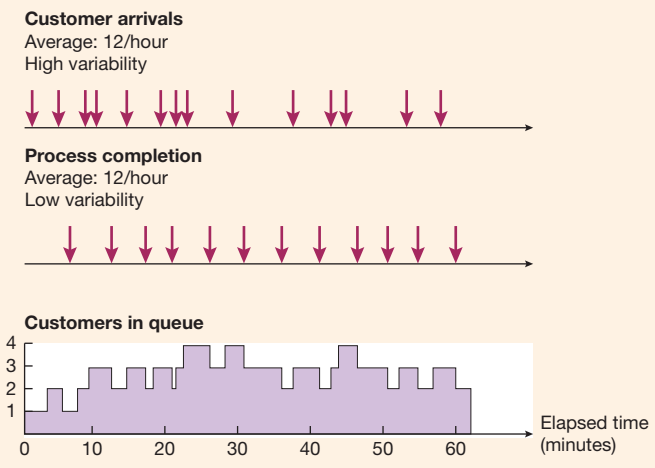

Figure 11.4 The impact of variability on queues

simple system, there may be up to four customers in the queue, the first being served and three waiting. The longest wait time for customers is therefore 20 minutes (4 × 5 minutes), which perhaps is surprising when we know that average process time matches the average arrival rate. The average queue length is about 2.5 customers, implying an average wait time of 12.5 minutes.

Computer simulations now provide invaluable information to the service operations manager. In a more complex situation than that described above, it would be impossible to model the likely outcomes, but a simulation can identify the impact of different queue designs, priority rules and so on. Whether the situation is complex or simple, the key question remains: 'How long is an acceptable waiting time?'

We do not provide a detailed analysis of queuing theory here. Fundamental texts cover this area well.[9] However, useful terminology includes:

- ***Calling population.*** This describes the customer base. In many cases, the calling population may consist of a variety of groups, each requiring different things. For example, customers contacting the computer company's call centre may enquire about billing, delivery, faults, purchase of service contracts, and so on. A key element of queuing theory is the size of the calling population. Consumer services have so many customers that the calling population is thought of as infinite and the probability of the arrival of a specific customer is unaffected by recent events. If, however, the calling population is relatively small, as exemplified by potential callers to an internal computer helpdesk, the probability of new callers is reduced if significant numbers have already called.
- ***Arrival process.*** It is clearly essential to understand the arrival pattern for customers. Many arrival patterns follow an exponential distribution. Intervals between customer arrivals in a retail store in a busy period might follow this distribution, with the majority of intervals being rather short, and long intervals being somewhat rare.
- ***Queue configuration.*** This describes the number of queues and their location. In many retail operations there has been a shift from multiple queues linked to multiple servers, because some queues seem to move faster than others, leading to customers moving between queues and possible ill feeling as customers believe they have been treated unfairly. A single queue leading to multiple servers has the advantage of demonstrating equity of treatment.
- ***Queue discipline.*** Management will choose the rule to determine who gets served next from the queue. The most common rules were discussed earlier (in Section 11.4.1), though first come, first served is the most popular with physical queues.
- ***Balking.*** One key measure for service systems is the number of customers who don't join a queue that they perceive as too long. This is referred to as balking.
- ***Reneging.*** This measures the number who join a queue, wait for a while then leave the service system due to the perceived intolerable delay.
- ***Jockeying.*** This is the term for customers who switch from one queue to another hoping to receive service more quickly.

11.6 What happens when managers can't cope with demand?

There is usually a point at which service managers find it difficult to cope with increasing demand (the break point: see Figure 11.5). This is when managers and staff enter the 'coping zone'. At these levels of capacity utilisation things are just too busy – staff become stressed, everything becomes a problem and, importantly, perceived quality, i.e. customer satisfaction, declines along with revenues per customer.[10]

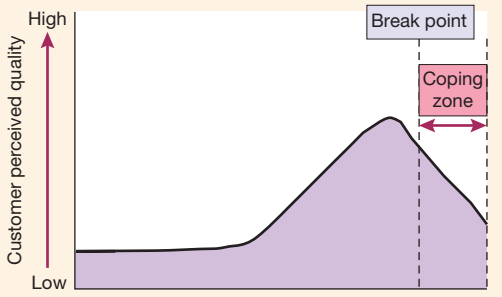

Figure 11.5 The coping zone in a high-quality restaurant

For example, in a restaurant operating at high levels of capacity utilisation, in the coping zone

- Customers have to wait a long time for service.
- There is increasing likelihood of 'stock-outs' (items removed from the menu).
- Customers feel rushed and under pressure not to ask too much from busy serving staff.
- Staff feel under pressure and are less likely to give courteous responses or the personalised service expected.

The break point is usually reached before full or 100 per cent utilisation for two reasons. First, it is often not possible to run any single resource at full capacity for any period of time; staff in a restaurant for example simply cannot work 'flat out' all the time. Second, several resources may be involved and while staff might be working 'flat out', only 80 per cent of the tables might be in use.

Interestingly there may be problems at times of low utilisation, too, which affect both staff and customers. In the case of the restaurant,

- The perception of the overall quality of service experience is low because the restaurant is 'dead'. There is no buzz of conversation; there are often prolonged silences.
- Service may be slow, because although there are not many customers, the kitchen may not be working at maximum effectiveness.
- In the same way, serving staff may be less attentive than might be expected, because again they may not be busy enough to be fully tuned in to customer needs.

Figure 11.5 illustrates the profile of customer perceived quality against capacity utilisation and illustrates how quality may suffer through both too many and too few customers in a high-quality restaurant.

The shape of the profile, the break point and the size of the coping zone will vary between organisations. Figure 11.6 illustrates the relationship between customer perceived quality (satisfaction) and capacity utilisation in a nightclub. Here customers may not enjoy the atmosphere until the place is crammed with bodies! Still, at some point the club's resources will start to struggle with the demand placed upon its resources – door staff, toilets and bar, for example – as it enters the coping zone.

11.6.1 How to manage the coping zone

There are seven steps in building up this profile and managing the coping zone.

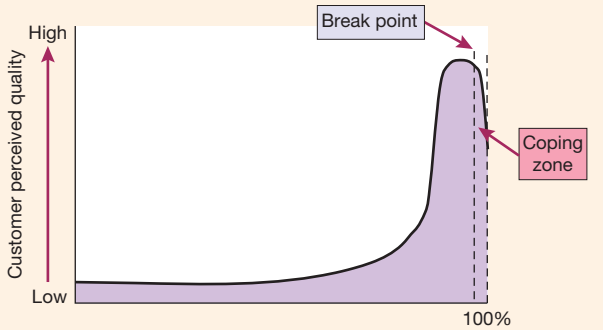

Figure 11.6 The coping zone in a nightclub

Step 1 Identify the service concept

Underpinning the service concept of a high-quality restaurant is the belief that customers may book a table for the whole evening. They will not be rushed to vacate their table since the restaurant has no intention of selling the space twice in an evening. It is intended that the service experience should be relaxed, with staff able to converse with customers and make recommendations about food and wine, where appropriate.

This is a very different concept from a restaurant wishing to create a high-energy situation, often with staff rushing around, and with the customers encouraged to eat up and leave. It is important to be clear as to what the designed or desired service concept is, particularly as the restaurant gets busy.

Step 2 Determine how capacity utilisation is to be measured

For the high-quality restaurant, the best measure of capacity utilisation is the number of tables, and also chairs, that are occupied during the evening. To some extent, other resources such as serving staff or kitchen capacity can be adjusted to the busyness of the restaurant area.

The unit of measure is often best taken at the lowest level that the business analyses or controls performance. A call centre might look at the average loading of a customer service agent on an hourly basis throughout the day, whereas a professional service firm might look at an individual's case load.

Step 3 Draw the outline profile

Figure 11.5 shows the relationship between customer perceived quality and capacity utilisation as it exists for the majority of customers on the majority of occasions. This does not represent all customers at all times. Some customers, for example, prefer the empty restaurant and would rank it as high quality at low utilisation. For others, the occasion and their mood will have a significant influence on where they would place themselves on the profile. This data can be captured from aggregating customer satisfaction indices and comparing it to utilisation at the time.

Step 4 Understand the nature and impact of the coping zone

While we accept that low utilisation is as much of a problem for quality of service as high utilisation is – indeed in some ways it is worse because revenue is also low – here we are focusing on operations working in the coping zone. It is important to recognise the signals

that suggest this zone has been reached and be sensitive to them. Tempers flaring, customers looking around or queues appearing may all signal that breaking point has been passed. It is also worth undertaking some financial analysis to try to demonstrate the impact between passing the break point and its effects on costs and revenues, for example.

Step 5 Determine the 'ideal' operating area

In Figure 11.5 we have identified the break point at 80 per cent capacity utilisation. At 100 per cent utilisation, in the case of the restaurant, it would be impossible to seat any more customers. There are two broad approaches that can be adopted:

- Operate at 80 per cent capacity utilisation. Operating at this point would suggest that the restaurant could be losing potential revenue. It is true that it might receive lower short-term revenues, but it may also upset longstanding customers by appearing to be greedy, squeezing as many people as possible into every available space. It is critical to understand the difference between customer satisfaction at 80 per cent and at 100 per cent, and to what extent this significantly reduces the customers' likelihood of returning.
- Operate at 100 per cent capacity utilisation. Generally speaking, this is a short-term cash-generating strategy. It is more appropriate to theme restaurants that are the 'place to be seen' for a period, before those customers that are concerned about fashion move on to the next 'in' place. This strategy might also be appropriate for restaurants in holiday resorts, which do not expect high levels of customer returns.

Our example restaurant depends on long-term customer retention and word-of-mouth advertising. As a result, it targets its operations at the 80 per cent point. To manage this, the owner has removed some of the tables to give a less crowded feel to the area, only replacing them on particularly busy occasions.

As a result, the owner has made the 80 per cent a 100 per cent effective capacity. In other words, at this point, the restaurant is making sufficient revenue to meet its short-term financial goals, and is giving a high level of customer satisfaction to safeguard its future business. It is worth noting that because there is a gap between 100 per cent effective capacity and 100 per cent potential capacity, it is possible for the restaurant to be working at greater than 100 per cent utilisation on some occasions. For many managers this is the norm in their businesses!

Step 6 Understand why coping happens

Clearly it is impossible to maintain the capacity balance on 80 per cent capacity utilisation at all times. Even if the restaurant has a booking policy, there is always the possibility that one of the most valuable customers will book at the last minute and the owner will be reluctant to turn this business away.

In other situations, the launch of a new service or periods of faster than anticipated growth may put parts of the business into 'coping' mode. This has been seen recently in the customer service departments of mobile phone network providers and banks following product launches. In some cases, some of the coping might have been avoided if the company had carried out some forecasting, or had simply communicated internally.

A key point here is to recognise that all but extremely resource-rich organisations will be in the coping zone sometimes. If the coping zone is never entered, the inevitable conclusion is that the organisation has too much resource.

Step 7 Develop coping strategies

Most organisations cope after a fashion. In the restaurant, all the diners are given food, but perhaps not with the greatest customer experience. Likewise, on the crowded flight, all

passengers get a meal and a drink, though those that are served last may have limited choice and little time to eat before the aircraft starts its descent.

Left to their own devices, customer-facing staff will find their own ways of coping. Some of these informal coping strategies will be entirely appropriate and innovative, using interpersonal skills and intuition to judge how to handle each customer. Others might be less satisfactory, typified by the following examples:

- Waiters who become overly focused on one task, making it impossible for customers to attract their attention to make yet more demands.
- Doctors' receptionists who, faced with a crowded waiting room, become extremely efficient in their dealings with patients, to the point of rudeness.
- Retail assistants who 'forget' to offer a customer a range of services, knowing that if the customer chooses one of these, their workload will increase.

Operations managers develop coping strategies based on one or more of the following:

- Giving more information to customers alerting them to possible difficulties. An example is an electricity company that after a major storm places a recorded message on its help line to say 'If you're calling about loss of power in this district, we should be able to restore it within two hours.' This reduces the load on overworked telephone lines and operators.
- Intentionally reducing the service on offer, perhaps using a limited menu at peak times in the restaurant.
- Being clear to staff about what really matters most for customers: concentrating on the 'must dos' rather than the 'nice to dos'.
- Building resource flexibility by bringing staff from a lightly loaded area to assist with the overload. Call centres manage this by switching calls to other centres, whereas Disney brings managers from back-office functions to assist with customer-facing operations on busy days. It is important to note that some of this resource may not be as efficient as the normal workforce.

There is a very strong link between prolonged overload and employee stress (see Chapter 10). It is relatively easy for providers to deal with short-term, predictable overloads. If we know we're going to be busy for a week or two, we can prepare for it, and many people get a 'buzz' from working together to cope with a crisis. The real problem with coping comes from protracted periods of overload, without hope of a let-up in the foreseeable future. Management support and appreciation becomes extremely important at this stage.

If the operation is in the coping zone for prolonged periods, it may be necessary for managers to give their staff 'licence to underperform'. For example, nurses in a busy accident and emergency department may not be able to carry out all their duties in the way in which they were trained. If this persists for any length of time, this will lead to stress and possible burnout. Part of the coping strategy, therefore, is to agree which bits of the service are 'must dos' and which bits can be safely left for the time being.

11.6.2 Coping: key questions

We have devoted a lot of space to coping because understanding how the organisation deals with this area may give clues as to where capacity management must be strengthened. The key questions to address are:

- What does the customer perceived quality/capacity utilisation profile look like for your service or services?
- How does this vary by service process and by customer group?

- What measures or early warning signals tell you that you are about to enter the coping zone (as opposed to measures like lost customers or increased complaints which tell you that you *were* in the coping zone)?
- What suffers for customers when you enter the coping zone?
- What suffers for employees when you enter the coping zone?
- How could you manage the coping zone better to reduce the impact on customers and employees?
- How could you avoid being in the coping zone as much?

Of course, coping will affect every part of the organisation, in areas where both chase and level strategies are operating, although coping is perhaps more obvious in operations that are employing a chase strategy. In effect, chase becomes level in the short term because the organisation is not capable of adding another unit of capacity quickly enough to deal with an unexpected surge in demand. Coping is perhaps more sensitive here because, as we have noted, organisations employ this strategy when fast response or high levels of availability to customers are particularly important. In such circumstances, customers are not usually prepared to wait, either because the service is not particularly valuable to them, or because there are alternatives available to them.

11.7 How can organisations improve their capacity utilisation?

There are four important additional ways of trying to improve capacity utilisation:

- yield management
- building flexibility
- reducing capacity leakage
- getting organisational support for capacity utilisation.

11.7.1 Yield management

Yield management is employed extensively by hotels and airlines to deal with the fact that their capacity is perishable (see Case Example 11.4). In other words, if the hotel room is not sold tonight, the contribution from that potential sale is lost for all time.

Yield management is focused on determining the maximum revenue to be obtained from the various segments served by the capacity at hand. Thus the airline estimates how many full-fare-paying (business-class) passengers will book for any given flight, and adjusts the remaining capacity for economy-class passengers and other discount, pre-booked customers. As departure time approaches, the airline may release some capacity to discount travel shops and, as a last resort at the very last minute, to stand-by passengers.

Service managers must be aware, however, of the potential damage to the service concept in using this approach. Full-fare-paying customers may be unhappy to discover that the person in the seat next to them is flying for a fraction of the price. This may give the impression that the airline is merely after every last dollar of revenue, with customer satisfaction of minor importance. The Kowloon Hotel appears to have overcome this particular objection by creating a completely new concept where the charging policy is clear and unambiguous. Customers can therefore make their choice of eating time, knowing that they will be treated equitably.

Case Example 11.4 The Kowloon Hotel, Hong Kong

Sheryl E. Kimes, Cornell University

Source: James Davies

Yield management, the notion of charging higher prices when demand is high and offering discounts at times of low demand, has traditionally been applied in reservations-based industries such as airlines, hotels and car rental agencies. Managers at the Kowloon Hotel in Hong Kong felt that it might offer them the solution to improving their restaurant revenues.

The Kowloon Hotel on Nathan Road in Hong Kong is well known for its sumptuous all-day buffet. The buffet, which includes a selection of sashimi, oysters, salads and desserts, is open from midday to midnight. As is typical with most restaurants, customers only wanted to dine at particular times of day, and the restaurant was often empty in the late afternoon and late evening. To deal with this problem, the Kowloon Hotel's managers decided to move away from a single price for its buffet and charge different prices depending on when customers arrive.

When guests arrive (check-in) they now receive a 'buffet zone pass'. The cost of the pass varies depending on their arrival time. At noon, the price is HK$118. It increases to $128 at 1.00 p.m., but then drops back to $118 at 2.00 p.m. The 3.00 p.m. price is even lower ($108), but then progressively increases from $128 at 4.00 p.m., to $168 at 5.00 p.m., $208 at 6.00 p.m. and $248 at 7.00 p.m. Following this peak, the price gradually decreases back to $138 at 10.00 p.m. and to only $98 at 11.00 p.m.

Not only has this new pricing system resulted in a 33 per cent increase in revenue – which was attributed to a fuller utilisation of the restaurant space, hence an increase in revenue per available seat hour (RevPASH) – it has also proved to be a hit with customers, with extremely positive customer reaction. As a result, the management has decided to continue the time-of-day pricing for an indefinite period.

11.7.2 Building flexibility

There are four basic forms of operational flexibility:

- ***New service flexibility.*** This is the requirement of the service operation to introduce new services into an existing mix. It will be necessary to define how frequently this might occur and the extent to which the operation will require new capabilities to achieve it. For example, house loan (mortgage) companies are continually introducing new 'products' with varying interest rates and repayment terms. In this case the frequency of new 'product' introduction is extremely high, but the requirement for new capability is low.
- ***Service mix flexibility.*** This is the ability of the operation to deliver more than one service. A hotel may provide a number of services simultaneously dealing with business people, holiday travellers, conferences and wedding celebrations.
- ***Delivery flexibility.*** This is the capability of the operation to change the timing of the activity. Courier organisations are increasing this form of flexibility, offering different speeds of delivery and a range of pick-up and delivery times.

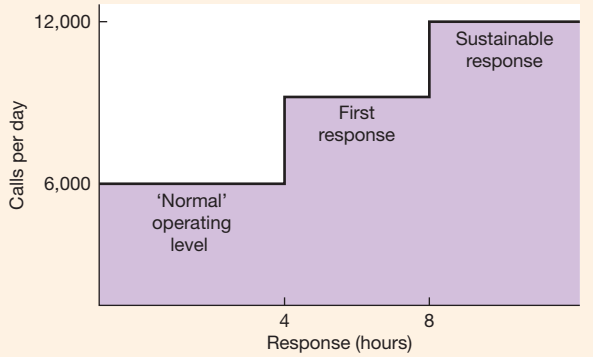

Figure 11.7 Minimum effective lead times for a call centre

- ***Volume flexibility.*** This form of flexibility is required by many consumer services operating a chase strategy. It refers to the ability of the organisation to change its level of output to cope with fluctuating demand. Thus the call centre may deal with 6,000 telephone transactions on a normal day, but may have to cope with twice that amount following an advertising campaign or a new product or service launch.

It is critical to define carefully the type and extent of flexibility required in order to develop effective capacity management plans.

Figure 11.7 demonstrates the notion of minimum effective lead times – the time it takes to respond to a change in demand. If an unexpected surge of demand occurs, the call centre manager can increase capacity by 50 per cent in four hours. This is accomplished by asking staff to stay on after their normal shift, by calling in off-duty staff, and by bringing in other staff from the organisation to man the phones. It should be noted that this is very much a 'first' or emergency response as some of the staff will not be fully trained and productivity levels will probably suffer as a result. Within eight hours the call centre manager can bring on line more capacity, perhaps from other call centres, if the increase in demand is sustained.

This is a valuable tool in assisting operations managers to plan for foreseeable contingencies. In addition to specifying the type of flexibility required, the service manager must also consider the following:

- ***Range.*** How much flexibility is required? Does the call centre need to move from 6,000 calls to 12,000 calls or only to 8,000 calls per day? How many new services will be introduced and how frequently?
- ***Response.*** How quickly must the change be made? Can the call centre change from 6,000 to 12,000 calls in four hours or will eight hours be good enough? Clearly the faster the response, the more expensive it is likely to be.
- ***Effectiveness across the range.*** Most processes have an optimal range. It is unlikely that they will be equally productive across the potential range; they are likely to strain at the extremes.
- ***Cost of providing the flexibility.*** What is the premium both for the change in output level itself, and for providing the capability in the first place? For example, providing training so that more staff are multi-skilled represents a significant investment.

There are a number of approaches to building flexibility. These include:

- ***Flexible employment contracts.*** Employees may be asked to work a given number of hours per week, month or year. In this way, staffing levels may be more easily matched to expected

demand. Flexibility in this area may extend to the requirement for staff to move between functions to cover fluctuating workloads. It is important to recognise that these staff may not be as effective when carrying out tasks that are less familiar to them.

- ***Overtime.*** Asking staff to work longer hours is a common short-term measure for building volume flexibility. However, it is generally accepted that using overtime in the longer term does not give the required increase in output as employees simply expand work to fit the extra time available. In this event, overtime is a poor approach, costing a great deal in extra salary and creating employees who are overtired through being at work for long periods.
- ***Short-term outsourcing.*** The development of relationships with other service providers who can deal with short-term peaks in demand is a common practice, particularly in call centre activities. This requires some pre-qualification of suppliers, but can provide flexibility at relatively low cost.
- ***Menu-driven service (standardisation).*** Standardisation of service offers provides opportunities for increased volumes of fewer activities. This creation of 'runner' activities (see Chapter 8) smooths the workload and provides opportunities to provide a degree of customisation or personalisation for customers. This is often referred to as the 'Dell' approach, whereby customers can create their own configuration of computer from a wide range of standard modules on offer. Education providers such as the Open University adopt a similar approach in allowing students to tailor their courses, within pre-determined constraints.
- ***Teamwork.*** The development of multifunctional teams allows a group of employees to deal with fluctuating workloads rather than simply operate on an individual basis. There are several benefits to this approach, including the support that group members give each other, and the ability to 'flex' capacity very rapidly.

11.7.3 Reducing capacity leakage

Many operations managers carefully plan their capacity in ways we have discussed in this chapter; however, sometimes managers are surprised to find they do not have as much capacity as expected. Here are some possible explanations:

- ***Labour sickness and absenteeism.*** Prolonged periods of overload and compulsory overtime usually prove counterproductive, with staff taking time off to recover. Alternatively, the organisation may need to look at its management style, placing more emphasis on team building rather than 'command and control' approaches.
- ***Labour underperformance.*** It is extremely common to find that call centres may have the right number of 'heads' but they may be ineffective because there has been insufficient investment in training. Alternatively, employee churn means that experienced staff leave just at the point when they are becoming effective.
- ***Scheduling losses.*** There are times when staff are idle with too much capacity for the demand, whereas at other times there is too much demand for the capacity to deal with. This often arises because demand profiles are not understood or are too volatile, or where staff preferences for work patterns do not fit with the business need.
- ***Costs of complexity.*** The more the organisation deals with a broad range of services, the greater the possibility that staff deal with a greater percentage of tasks that may not be part of daily routine. This potential 'relearning' may give rise to inefficiencies and rework.
- ***Quality failures.*** The need to deal with quality failures is clearly lost capacity. Of course, part of the role of the call centre may be to deal with poor quality generated elsewhere in the organisation. It is essential that the extent of this rework is understood and charged to the appropriate location.

11.7.4 Organisational support for capacity utilisation

The challenge for operations managers is to understand capacity utilisation in the context of a changing world. Many of the issues need to be resolved by the organisation as a whole, rather than simply confined to the management of service delivery processes. Aspects of this organisational support include:

- ***How is the service concept changing?*** To what extent do operations managers have 'visibility' in the future strategic direction of the organisation? Without this inclusion in the strategic development process it is unlikely that capacity planning and capacity utilisation and development will be effectively carried out. Often resource managers are left to develop capacity plans that have little relevance because the concept has been changed. For example, many service companies providing repair and maintenance support for information technology have changed the emphasis of their service concept away from servicing equipment towards supporting customers. The nature and length of customer transactions have changed beyond recognition, and a link is required between this change in strategic direction and resource planning for it to be implemented.
- ***How well are the internal interfaces managed?*** A key role for the operations manager is to manage the internal relationships as well as customer relationships. Co-ordination of marketing promotions and new service introductions is vital, as is getting to the root of quality failures and long-term quality costs. Successful organisations are often those that manage the internal relationships well. This does not necessarily mean that everyone always agrees, but that there is valuable internal debate. In fact, it could be asserted that an organisation without some degree of conflict will not learn and move forward.
- ***How important is resource management in the culture of the company?*** In some sectors, resource management is seen as a low-level task. The 'stars' of the company are often seen as those who deliver the latest deal or solve the latest crisis. Resource management needs a different type of 'hero' who is able to plan for the longer term and persuade the organisation to think differently about resource management.

11.8 Summary

What is capacity management?

- Capacity management is concerned with putting a plan in place that makes the best use of resources to deal with the forecasted or expected demand for services.
- Service capacity is defined as the maximum level of value-added activity over a period of time that the service process can consistently achieve under normal operating conditions.
- Capacity is influenced by a range of factors including service mix, location, intangibility and resource constraints.

How can managers balance capacity and demand?

- Most organisations adopt a mixture of capacity strategies: level, chase and demand management. The mixture should reflect the strategy of the operation.

How is day-to-day planning and control carried out?

- Day-to-day operations planning is concerned with creating a 'schedule' or timetable based on the capacity plan, which allocates people, customers, equipment or facilities,