Note on Managing Capacity and Demand in Service Operations

Ensuring that service operations have the right amount of capacity to serve demand when it occurs is one of the most critical and interesting challenges in managing service operations. Not having sufficient capacity at the right time—over the course of the lunch hour, throughout the day, from week to week, over the course of the season, the year, or longer—will result in long lines, long waits, and customer departure. Having too much capacity, on the other hand, will result in excess worker idleness, low worker utilization and, therefore, unnecessarily high costs.

Having the right amount of capacity at the right time is complicated in service operations due to the inherent randomness in customer-facing operations, a phenomenon we have discussed earlier. Uncertainties in customer arrival times, customer expectations, service delivery times, etc. makes it difficult, sometimes impossible, to accurately predict the right amount of capacity needed ahead of the time.

Besides inherent randomness of demand for a service, another factor also complicates the task of capacity planning in service operations and that is the fundamental fluctuation—albeit reasonably predictable—of demand over a certain period of time. For example, restaurant operations experience pronounced peaks in demand during lunch and dinner times. Toll booths are overloaded during rush hours. Resorts typically experience peaks on weekends and during vacation seasons. While these “seasonal” fluctuations in demand may be reasonably predictable based on historical data, they still present the operations manager with the challenging task of raising and lowering the level of scheduled capacity (workers, equipment, facilities) continually to avoid overutilization or underutilization of capacity and the attendant problems of long waits during peak periods or loss of efficiency during slack times. Figure 1 graphically depicts random variations as well as average or predicted demand.

In summary, matching supply and demand in services involves handling the dual challenges of random fluctuations and relatively predictable fluctuations to create the right balance between the level of customer service, on the one hand, and operational costs, on the other. This trade-off between customer service level and service delivery

---

1 This note was written by Professor Farshad Rafii of Babson College for the purpose of class discussion in the Babson Fast Track MBA program. Copyright © 2007 Babson College.
2 Capacity is sometimes referred to as “supply” and demand as “load”. In this note, these terms will be used interchangeably.
4 While there are many elements that go into making the overall “customer service”, the waiting time to be served is probably the most critical one in most situations.
costs can only be made by taking into account the business strategy of the firm. As a general rule, operations that seek to differentiate themselves based on providing superior customer service will opt to schedule plenty of capacity and be prepared for even the highest possible level of demand. Those that seek to compete by providing low-price services will opt to economize on the level of scheduled capacity, thereby driving capacity utilization and efficiency levels up, even at the expense of longer customer wait times. The optimum decision thus depends on linking operations and strategy, a fundamental requirement of world-class operations we discussed at the outset of the Ops BBI.

The silver-lining in the complexity of the service capacity problem is that, if managed competently, it provides ample opportunity for creating competitive advantage.

In situations where multiple services are delivered as a package, such as in a hospital or resort operation, care must be taken to provide balanced capacities in all service delivery operations. Otherwise, a bottleneck in the provision of a single service such as amenities in a resort operation, might discourage customers from availing themselves of more profitable or strategically critical services such as renting rooms.

**Short-term Capacity Strategies**

At the simplest level, two alternative strategies are available for matching capacity and demand over any given period of time. They are referred to as *Chase Demand* and *Level Capacity*, or simply *Chase* and *Level*. As the names imply and as shown in Figure 2, the Chase strategy involves increasing and decreasing scheduled capacity to “hug” expected demand as much as possible. This is indeed the strategy Wal-Mart has recently tried to pursue as it has sought to schedule its store staff in short time slots that would correspond closely to expected store traffic during the day. The Level strategy, on the other hand, seeks to avoid the costs and hassles of frequent changes in staffing levels by scheduling a single level of capacity over the entire period of time under planning -- or at most a couple of different levels, one for the highest demand, or peak period, and another for the lowest demand, or slack period.

It is not hard to see that each strategy has advantages and disadvantages, and that the choice between the two requires making an intelligent trade-off. By tightly matching capacity to demand, the Chase strategy delivers a consistent level of service and results in a high level of utilization of the workforce (i.e., little idleness or frantic pace) and, therefore, a high level of efficiency. The primary downside of the chase strategy is the high costs of changing the capacity level. These could be the cost of setting up the work, hiring, firing, loss of learning, etc. Work environments following a Chase strategy are also generally more chaotic, experiencing a high level of turn-over and error rate.

---

The Level strategy avoids these problems by keeping staffing levels relatively steady. The price of this stability is that the pace of work will vary significantly over the course of day or season, resulting in very low utilization (efficiency) levels during slack periods, and a frenzied pace and potentially poor service during peak periods.

Because of these advantages and disadvantages, the Chase and Level strategies are appropriate for very different work environments. The Chase strategy is suitable for low-skill situations where hiring, training, and dismissal costs per worker are low and high turnover rates can be tolerated. Fast food operations, for example, generally follow this strategy. The Level strategy, on the other hand, works best in high-skilled or professional situations where hiring and training costs are high, turnover can be very costly, or operation of complex equipment is involved. In a stock brokerage firm, for example, traders and brokers are scheduled fairly evenly throughout the day, whereas the number of phone operators and customer service reps is tailored more closely to the call volume.

Table 1, adapted from a classic article on the subject, summarizes the advantages and disadvantages of these two strategies and the circumstances they are best suited for. Of course, in practice pure Chase or Level strategies are rarely used and the optimum strategy is often a mix of the two.

<table>
<thead>
<tr>
<th></th>
<th>Chase-Demand</th>
<th>Level Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labor skill required</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Job discretion</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Compensation rate</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Working conditions</td>
<td>Sweatshop</td>
<td>Pleasant</td>
</tr>
<tr>
<td>Training required per employee</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Labor turnover</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Hire-fire costs</td>
<td>High (in total)</td>
<td>Low</td>
</tr>
<tr>
<td>Error rates</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Amount of supervision required</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Type of budgeting &amp; forecasting req’d</td>
<td>Short-run</td>
<td>Long-run</td>
</tr>
</tbody>
</table>

Both Chase and Level strategies involve determining the appropriate level of slack to be planned for. Capacity in either situation can be provided generously or sparingly. For a given desired service level, the appropriate amount of slack depends on the randomness or volatility of fluctuations in demand: the higher the amplitude of fluctuations, the greater the amount of slack necessary to achieve a certain level of service, and the lower the resulting level of resource utilization. In other words, uncertainty costs money. For this reason, limiting demand randomness through such means as a reservation system or incentives can result in more efficient use of resources or higher level of service, or both.

---

Tactics for Matching Capacity and Demand

Demand for services can be met by influencing demand to reduce its variability and move it to when and where capacity exists, by providing capacity when and where demand occurs, or a combination of the two. In other words, you can manage demand, manage capacity, or manage both.

Demand can be managed in a number of ways, including:

**Peak shaving**—this approach involves shifting demand from peak to non-peak periods, through pricing, reservations, developing alternative services during peak, and other means. Peak shaving reduces the overall level of capacity needed and limits the frequency of costly changes in scheduled capacity. It is well-established in several industries, including the utilities industry.

**Developing non-peak demand**—this approach also results in leveling the demand curve. Addition of a breakfast menu by hamburger chains and weekend vacation packages by hotels that cater primarily to business travelers during the week are prime examples.

**Developing complementary services**—especially in situations where service delivery is constrained by a facilities bottleneck such as theatre seats or amusement rides, offering other attractions will reduce the pressure on the bottlenecked operation.

Tactics for managing capacity include:

**Using temporary or part-time employees**—stay-at-home parents during school hours, or students during summer time or school vacations are prime examples.

**Increasing self-service possibilities**—is another way to provide capacity demand during peak periods.

**Maximizing peak period availability**—the best available and most “efficient” capacity should be reserved for scheduling during peak periods. Workers with highest skill levels, motivation, and tolerance for pressure should, within limits, be used during peak hours. This applies also to equipment and facilities. Break downs are especially costly during peak periods so care should also be taken to ensure that all necessary equipment and tools are in excellent working order prior to the peak season and that work processes and supplies are properly set up ahead of the rush.

As mentioned before, random variability induced by the presence of customers in service operations is what makes delivering services particularly challenging. Matching supply and demand in services essentially involves either accommodating this variability as it occurs or taking steps to limit it at least to some extent. A fairly exhaustive collection of
tactics has been developed by Frances X. Frei\(^7\) for accommodating or reducing variability.

**Managing Customer’s Perceptions of Service Level**

So far we have discussed several actual ways of managing capacity and demand for services. Ultimately, however, service levels are as much a matter of perception as they are of reality. Perceptions, in turn, are often a function of expectations. It is therefore critical to set the right expectations for the level of service likely to be provided, especially during peak periods. Beyond setting the right expectations, here are several “truths” about people’s perceptions about waiting and their implications about how to manage these perceptions\(^8\):

**Uncertain or unexplained waits appear longer than known or explained waits.** This is why it is a good idea for airport check-in personnel to provide frequent updates of delayed flights, specify how long they believe the wait will be, and why the delay has occurred. (I am amused, though, how often the delay is due to factors that have nothing to do with the airline!)

**Idle waits appear longer than occupied waits.** Hence, restaurants hand out menus to waiting customers (which, by the way, probably speeds up selection and accelerates table turnover), doctors’ waiting rooms have plenty of reading materials, and there are sometimes mirrors near elevator doors.

**One long wait appears longer than several short waits.** Thus, interrupting a long wait by providing some attention to the customer (such as serving waiting restaurant customers in the bar) or by requesting customers to perform some action (such as filling out luggage tags) can reduce the perception of the wait time. Generally, the sooner the customer receives some attention, the better.

**Unfair waits appear longer than fair ones.** Any service layout that deviates from a clear first-in-first-out priority rule runs the risk of creating the perception of longer waits. Multiple lines in fast food restaurants and banks are a prime example where the perception that “the other line always moves faster” leads to quite a bit of frustration and line hopping. As a result, many establishments that have sufficient space typically opt for a single waiting line. Express check-out lanes in supermarkets and first-class check-in lines in airports provide a similar dilemma, although in these cases the operational and strategic benefits inherent in discriminating among different types of customers typically override the negative repercussions and the perceived unfairness.

---

\(^7\) Frances X. Frei, “Breaking the Trade-off Between Efficiency and Service”, Harvard Business review, November 2006, page 4. (This article was assigned for reading in the Ops BBL.)

Solo waits are longer than group waits. The sense of camaraderie that sometimes erupts among people who have been enduring a long wait actually serves to lessen the pain. It therefore behooves any service establishment to do what it can to promote the sense of group waiting and commiseration.

**Service Level vs. Capacity Utilization**

A very useful measure in analyzing service operations is the notion of “capacity utilization” or, simply, “utilization. It is a measure of how busy a resource (provider, equipment, facility) is and it is formally defined as:

\[ \text{Utilization} = \frac{\text{Actual Output}}{\text{Theoretical Capacity}}. \]

As mentioned before, a trade-off exists between the level of service (i.e., low waiting times) provided the customer and the level of utilization of the service provider’s capacity. This inverse relationship is shown in Figure 3. Unlike manufacturing operations, service delivery processes can rarely approach or exceed 100% utilization for sustained periods of time without resulting in very long waiting times. This is due to variability introduced by customer involvement in the process. For a given level of service, the higher the level of variability, the lower level of resource utilization that can be achieved.

Where on this trade-off curve to position your service operation is strictly a matter of strategic choice. Seeking differentiation through service requires maintaining a high level of service and low waiting times. This would invariably necessitate having plenty of service providers on hand. Alternatively, competing on the basis of low cost necessitates economizing on service providers, driving up provider utilization rates and customer waiting times.

Because utilization rates are usually easier to estimate and work with than service levels, this relationship between service levels and utilization rates can be used to simplify the task of capacity planning. Often, through historical performance data or benchmarking studies, a service operation can discover the worker utilization rates that result in the desirable level of service. This can be, say, only 65% in a service seeking differentiation or 90% in a cost-driven one. These benchmarks can be used in a simple process described below to arrive at the level of capacity needed.

However, before doing so, it is important to point out that the actual service level in a given set-up is both a function of strategic choice as well as astute managerial action. Because a typical operation is rarely at the most efficient frontier regardless of where on the trade-off curve you have positioned yourself, it is often possible to take managerial action along the lines suggested above to improve service levels without dropping utilization rates, or increasing utilization rates (and, therefore, efficiency) without dropping service levels. This is graphically indicated in Figure 4.
A Simple Step-by-Step Process for Determining Capacity Requirements

1. Forecast average demand for a given time period (hour, day, week, etc.)
2. Improve current service delivery system by better management of demand, current capacity, and the service delivery process (i.e., move to the most efficient frontier)
3. Based on strategic objectives, determine a target service level and the corresponding target utilization rate to be maintained during that time period, and finally
4. Estimate capacity requirements for that period by a variation of the above formula: capacity = demand / utilization.

In situations where a simulation model of the service delivery process and customer arrivals can be constructed, a variation of the above approach can be used:

1. Build a simulation model characterizing customer arrivals and service delivery rates for each time period
2. Improve the current system as much as possible
3. Based on strategic objectives, determine a target service level by specifying an average or a distribution of customer waiting times for each time period
4. Test various levels of capacity to arrive at the desired waiting times.

Long-term Capacity Strategies

Discussion in this note has concentrated mostly on managing the human delivery aspect of services in the short to medium term. In cases where service delivery has a significant facilities component, e.g., hospitals, hotels, etc., the creation of such facilities which often have very long lead times and come in big “chunks” has to be planned well in advance. As in the earlier discussion, there are two primary strategies for planning long-term capacity: Lead-Demand vs. Lag-Demand (Figure 5.) In a Lead-Demand strategy capacity is brought on-line in anticipation of demand developing down the road. Such a strategy is called for when the cost of running out of capacity is very high or a pre-emptive expansion move is advantageous from a competitive point of view. The downside, however, is significant underutilization of capacity for a period of time. As a consequence, a Lead-Demand approach is workable if capital and maintenance costs of such overcapacity are not prohibitive, as would be the case in, for example, creating store fronts or office locations. However, when capital and maintenance costs are high or when new capacity is “chunky”, (i.e., minimum economic size of new facilities is very large compared to the rate of growth in demand), it is prudent to allow demand to develop before new physical capacity is brought on line. A good example of such a situation would be wide-bodied planes for long-haul flights.

Under either strategy, it makes sense to build new service facilities with an awareness of future expansion needs. This would allow designing the facility in such a way as to make future expansions or additions quick and low cost.
Figure 1:
Random and Average Fluctuation Patterns in Demand

Demand (units)

Time (hours, days, weeks)

Predicted or Average Demand

Actual Demand
Figure 4: Managerial Action for Becoming Lean vs. Strategic Choice

Figure 5: Long-term Capacity Planning: Basic Strategies