Why is Managing Capacity So Difficult? Main Challenges and Solutions

Melda Ormeci Matoglu
University of New Hampshire

Abstract

In today's supply chains, there are a lot of uncertainty. The proliferation of products, build-to-order, shorter order-to-delivery cycles, and increased personalization create highly variable demand patterns. There are tools with different scopes and foci to help manage this: from targeted tools to supply chain wide broader tools. In addressing this challenge, we focus on buffering variability with capacity and look at how to utilize production/service capacity as a lever to effectively operate under uncertainty. Many factors, varying from competitive priorities to costs, need to be considered and depending on the situation the right action should be taken. We present insights that tell us how this can be done under different conditions by effective capacity decisions.

Managing a business is no easy task and the challenges are exacerbated by the presence of uncertainty. Uncertainty presents itself in many forms: variability in demand, in supply, in quality, … From a supply chain point of view, many factors contribute to this: Proliferation of products which lead to line sharing and different product mixes being produced from day to day; build-to-order, shorter order-to-delivery cycles and increased personalization. This results in a large variation in the daily demand for parts, products and services. Most, if not all, businesses face the challenge of operating under variability. Hopp and Spearman suggest three levers for managing variability: capacity, inventory and time.[1] When there is variability in a system (as it's bound to happen) it needs to be buffered either by capacity, inventory, time or a mixture of these. Specifically, to address uncertainty businesses may choose to adjust capacity dynamically (keep extra capacity and use it as needed), hold extra inventory or quote longer due dates. The choice depends on the business and its competitive priorities. In the supply chain context, one can look at capacity in different dimensions:

Warehousing capacity, transportation capacity and production/service fulfillment capacity. Having ample capacity in these dimensions becomes crucial as firms try to fulfill consumer expectations. To meet customer service expectations, firms increase the footprint of their distribution networks by adding warehouse capacity and this has a positive relationship with transportation capacity.[2] Production/service capacity decisions are among the drivers behind warehouse and transportation capacity decisions, as the former creates the input for the latter processes.

We explore how to use production/service capacity as an effective lever in the face of uncertainty and provide some insights about how to best manage it.

In Operations Management literature capacity is defined as "the maximum rate of output of a process or system" or the throughput.[3] Capacity decisions are important decisions for businesses. In the long run, they determine capital requirements and have a big impact on fixed costs. In the short run, they determine whether demand will be met or lost. An important factor in making capacity decisions is the organization's competitive priorities and operations strategy. The decisions made by an organization that prioritizes low-cost will be different than one that prioritizes flexibility or time or quality. While a low-cost priority values efficiency and high utilization, the others may value having extra capacity on hand to handle potential fluctuations in demand or issues in quality or adopt strategies to adjust capacity without accumulation of inventory or excess capacity.[3]

Once a plant/facility is built the design capacity determines the maximum output and making changes to it, if even possible, typically requires big capital expenses. Similarly, in the service setting, the size of a facility, think of available seating and kitchen set up in a food service business, determines how many customers can be served in each time period. Yet, during day-to-day operations, managers have some control over capacity, and they may actively make adjustments to capacity. One way of doing this is by adding or removing shifts. For example, Toyota and Japanese manufacturers are known for scheduling for two shifts and using a third shift to catch up if needed.[1,4] While adding or removing shifts can mean relatively significant adjustments to capacity, managers may also utilize finer controls which allow them to adjust capacity within given limits (for example, by hiring temporary workers from external labor supply agencies, subcontracting, authorizing overtime, renting or shutting down work stations, and so on.) The availability of a flexible workforce makes such an adjustment a feasible option and makes capacity an important lever. Today a significant portion of labor is available in this flexible manner. For example, in 2017, the US Bureau of Labor Statistics reported 13.8% of the workforce

(over 20 million workers) as flexible workers (independent contractors, on-call workers, temporary help agency workers and workers provided by contract firms).[5] In 2017, in the Netherlands, 6.8% of the active workforce was composed of flexible workers (on-call and temporary agency workers) and in the EU 14.3% of the employment was reported as temporary.[6,7]

Many industries face the problem of managing capacity in the face of unpredictably varying demand. For example, consider a manufacturer that assembles products based on outstanding orders, like Dell assembling computer systems based on customer orders. When the order queue is long, the manufacturer may open available manufacturing lines by staffing them with personnel and stocking the lines with parts, and when the queue is short close some of the manufacturing lines by sending workers home. Call centers also face this challenge as they try to optimize their staffing levels. Having many staff members available to answer calls provides quick service to customers and increases customer satisfaction, but on the other hand, it means higher costs and potentially staff sitting idle when call volumes are low. Yet, if the call center chooses low staffing levels, then customers may have to wait a long time before receiving service, severely impacting service quality and satisfaction.[8,9,10] Similar challenges are also faced in managing warehouse operations. For example, in a survey about warehouse management excellence, 35% of survey participants list fluctuations in supply and demand and 31% of participants list the need for better utilization of underutilized resources among their top pressures.[11] Thus, determining the number of pickers to staff to handle the volume of orders that need to be picked is an important business decision. A high number of pickers means customer orders being picked and shipped quickly, which improves customer service, while a small number of pickers means slower fulfillments but potentially higher utilization of pickers. The challenge of managing capacity is not limited to manufacturing or service providers trying to determine the best staffing/production levels. Consider deploying webservers to handle internet traffic. The number of incoming requests can vary substantially over time and may depend on factors like time of day or day of the week and so on. As the incoming traffic varies, the available server capacity may be too low, sufficient, or more than necessary. One could choose to run just enough number of servers to handle the maximum anticipated load, but that would mean much of the server capacity will not be used most of the time and one would still have to pay for unused server capacity. Not the most efficient approach… Indeed, to run a website efficiently the recommended practice is to "match the number of servers to the current request volume".[12] That is to use more servers when the request volume is high and fewer servers when the request volume is low. For example, Amazon Web Services offers clients

three ways to manage the number of servers needed: 24/7 instances, where typically the capacity may be set to a baseload level that captures minimum anticipated demand, time-based instances, where the client sets a schedule for a varying number of server capacity based on expected varying demand, and load-based instances that are automatically started and stopped to handle traffic spikes, when a threshold is crossed and more server capacity is needed.

The main observation in all these settings is that, if capacity is set too high and the resulting outstanding orders/customers waiting to be served queue is low or non-existent, then economies of scale are lost and the investment is not being fully utilized. On the other hand, if the length of the outstanding orders queue is high, it means delivery dates are being missed, customer satisfaction is plummeting, and thus potential revenues are not being realized. Both cases are not desirable from the business owner's perspective. When service level is part of the competitive strategy, as is the case for most businesses, managing this outstanding order queue through effective capacity choices can be an important lever.

In the literature typically two approaches to capacity management are cited. The first is a "chase strategy" where capacity chases demand. The operations manager varies the workforce level by hiring or laying off staff or adjusting the output by overtime, part-time employees or subcontracting. The aim is to match capacity with demand. This strategy is typically adopted by high volume consumer services where rapid access to services is part of the competitive strategy.[13] The second strategy is called "level strategy" where capacity is kept at a constant level. The level strategy aims to maximize the utilization of expensive fixed resources. Both strategies have advantages and disadvantages. While a level strategy is much easier to manage as the workforce, and so capacity, is constant, it may result in unwanted inventory or workforce sitting idle, or, if demand is higher than the chosen level capacity, customers being turned down and leaving empty-handed. On the other hand, a chase strategy is able to meet demand by continuously updating capacity but requires constant attention, and adjustments to capacity/workforce levels may incur additional costs, such as setup, administrative and training costs, and service quality may be impacted. Companies typically utilize level strategy for jobs requiring a high level of skills, with high training requirements and high compensation, and chase strategy for jobs with low-level of skills, low training, and low compensation.[14] Many businesses utilize a combination of these methods. A third strategy, frequently used together with level and chase strategies, is demand management, where the organization aims to influence demand to

smooth out the load on resources. This typically involves using a mixture of pricing, promotion, and reservation systems.[15]

In our research, we look at ways to manage capacity, and thus the outstanding order queue, by choosing between available capacity rates, with the option of idling resources and rejecting new orders.[16,17] The choice of different available capacity rates can be thought of, for example, an assembly facility that has several production lines, but at any point in time, the decision is between staffing and operating only one of the lines, two of the lines or any number of available lines. Naturally, there is a cost to operate the different number of lines (due to different staffing level requirements based on the number of lines operated) as well as a cost to change the number of lines that are being utilized. This changeover cost reflects an additional cost similar to the setup cost that is incurred to get these lines up and running or turning them off, or the administrative and training costs to adjust the workforce levels. While it is hard to quantify, we also consider a cost for turning away customers. Businesses may choose to turn away customers, or customers may themselves choose not to place an order when service/delivery times are significantly long. Both cases will result in a loss of goodwill which will impact the business, and that is what we are aiming to capture with this cost. Having a line open, but idling may also incur a cost for the duration a line is idled. Besides these costs, a business also incurs a cost for the orders that are waiting to be processed, think of it as the opportunity cost of revenue waiting to be realized or just a holding cost. We share useful and practical insights to address the challenge of managing capacity under these considerations.

Insight 1. ***When level capacity rules:*** There are several instances when picking a single capacity rate and sticking with it pays off. When there are significant costs associated with changing the capacity rates it is best to pick the best level available and stick to it. Jobs requiring very skilled laborers with very high training costs and/or hiring/layoff costs typically fall under this category. If the changeover costs are above a certain threshold, it is best not to change the capacity rates, and idle if there are no jobs and possibly reject/lose customers if wait times get too long. This falls within the common arguments for level capacity. The more interesting, but less likely cases, that call for sticking to a single capacity rate happen when idling is cheaper than savings from not operating, or when turning away orders is cheaper than operating. The first case calls for utilizing the maximum available capacity and just idling resources when there is no demand, and the latter calls for picking the smallest capacity rate available and turning away/losing customers when the outstanding orders reach a certain level. As odd and

impossible as these two cases may sound, they indeed have real-life applications: Consider an emergency service like ambulance or firefighting. When designing these systems, one would want to have ample capacity available, even if they are idle some of the time, and ensure that there is sufficient capacity when an emergency arises. Also, note that for such services capacity is the only possible buffering choice as response time is a key performance measure and it is not possible to inventory such services. On the other hand, consider popular venues, where long lines and hard to get reservations are attributed as status signals. In this case the scarcity adds value to the product and service being offered.

Insight 2. ***The case of zero changeover cost: No middle capacity.*** Consider the case when operating is preferable to idling capacity and turning away customers, that is we are not operating under the special circumstances of scarcity or emergency. When changeover costs are zero, changing the operating capacity is effortless. One may think that this calls for the "chase demand" strategy where capacity is adjusted to follow demand and as many adjustments as needed are made. Surprisingly, a best policy is to only use the lowest and highest available capacity rates and switch from one to the other at a judiciously chosen outstanding order level. If the outstanding orders are below this critical level pick the minimum capacity rate, if it is above the critical level pick the highest available capacity rate. This is indeed a very easy to implement policy, the one possible downside is that just around the critical level there could be a lot of switching back and forth between the capacity rates arising from the fluctuation in outstanding orders. Under the assumption of zero changeover cost, such a policy makes sense, but in most real settings a switch is not so "frictionless" and cost-free. A possible application of such a policy is automated systems/machinery which may be turned on or off without any effort and cost.

Insight 3. ***In the presence of changeover costs, "chase" but not so quickly.*** In many cases changing capacity involves a changeover cost. This may include administrative, training and onboarding costs as well as setup costs due to setting up of tools, preparing and moving materials to the production line and testing initial output. When a changeover cost is incurred for changing from one capacity rate to another, depending on the magnitude of this cost, some available rates may not be employed at all. This is in line with insight 1, where we stated that if the changeover cost is greater than a certain threshold it is best to use only one rate. Yet, the choice is not one rate or all available rates, instead utilizing a subset of the available rates depending on the magnitude of the changeover cost provides best results.

The changeover cost, in some sense, makes the controller liable for past decisions and results in an optimal policy that depends not only on the length of the outstanding order queue, but also on the current capacity rate being employed. Indeed, we find that the best way to manage capacity is to define bands for each capacity rate that is employed, and stick to that rate while the order queue is within the limits of that band. When the outstanding orders reach the upper limit of the band switch to the next available higher capacity rate to quickly serve customers, and when the outstanding orders fall to the lower limit of the band switch to the next available smaller capacity rate. The upper limit of the highest capacity rate corresponds to the point where we start turning away customers (or customers balk and leave) as due dates are becoming unacceptably long, and the lowest limit of the slowest capacity rate, which is typically zero, corresponds to the point where we start idling due to lack of demand. These bands corresponding to different capacity rates will overlap and this is due to the existence of changeover costs: once we make an adjustment to capacity, we want that change to "payoff" before making another change, we don't want to switch back and forth between different levels continuously. This also makes managerial sense: once we incur start-up costs or wind-down costs we don't want to make an immediate change and reincur those costs, furthermore, if we are talking about adding/removing people from the workforce, frequent changes may impact the future availability of workforce and its cost. Higher changeover costs typically mean larger bands and less frequent changeovers. Higher changeover costs also imply that very high or very low capacity rates may not be employed at all. This policy resembles a "chase demand" strategy, but the presence of the changeover costs requires that as we monitor the outstanding orders we do not react too quickly to little changes, instead wait and observe and make changes only when a threshold (upper or lower limit of the band) in the number of outstanding orders is reached. When one considers that every hiring/layoff indeed results in some transaction cost (in terms of administrative costs, training costs, etc.) indeed this policy also makes a lot of intuitive sense: For low skilled jobs where new hires can be easily found and no significant training is needed, it is ok to reduce the workforce when a drop in demand is observed and increase when demand picks up. On the other hand, for jobs requiring high skills where hiring is a demanding and expensive process (think headhunter fees, long interviews, long training) we should be more conservative as we adjust the workforce based on demand and make changes to staffing levels only if the workload is significantly low or high.

Insight 4. ***Broader Supply Chain tools.*** While so far, we have focused on how to address variability by buffering it with capacity, managers have broader set of tools available to them to decrease and manage uncertainty and volatility within their supply chains. Reduced volatility not only makes managing capacity simpler, but it improves an organization's overall performance. One strategic tool that organizations can adopt is supply chain risk management, in which organizations identify, assess, and mitigate risk within their supply chains. Identifying the risks in each node of the supply chain and assessing the impact they could create on the business is an essential part of this. These risks include external risks like suppliers (any disruption to the flow of products/materials), environment (including social-economic, governmental, and environmental), demand (its unpredictability) and risks related to plants/facilities (yield, quality, business processes, people, regulatory, …), and transportation (from weather related to pilferage). Understanding these risks, their likelihood of their happening, the extent of the threats they bring and the organization's ability to deal with them will be key in developing preventative and reactive action plans to mitigate them. Monitoring risks and periodically reviewing and making sure mitigation plans are up to date will help improve the resilience of the supply chain. The fire in 2000, in the Philips semiconductor plant at New Mexico, is a great example of two companies facing similar capacity challenges but ending up with different results due to their different ways of managing supply chain risks. The fire at the Philips plant was put out in less than 10 minutes, but had, as it was discovered later, affected their cleanrooms and as a result impacted chip production for over two quarters. Nokia, one of Philips' big customers and a big cellphone manufacturer at the time, with its continuous communication with Philips was able to realize the big impact of the fire quickly and work together with Philips and through its deep relationships with its suppliers and knowledge of supply markets was able to overcome the resulting chip shortage and continue its cellphone production relatively without impact, and indeed increased its market share. Ericsson, another major customer of the New Mexico plant, on the other hand, had a more laid-back approach, was too slow to realize the impact of the fire, its impact on chip supply, and unable to obtain parts from Philips' other plants and other sources ended with a decrease in market share and Ericsson "exiting significant parts of the business."[18]

One important input for supply chain risk management and an important tool in decreasing and managing uncertainty within supply chains is end-to-end supply chain visibility. Traditionally organizations have relied on siloed data as each stage of the supply chain managed its own risks and capacity. This resulted with unconnected systems, manual data reconciliation and in

many instances, decision making with incomplete and/or erroneous data. End to end supply chain visibility requires a collaborative relationship with suppliers and downstream customers. It provides accurate and real time information on demand signals, inventory levels and progress of orders.

Visibility also feeds into better data analytics and demand forecasting. Having more accurate forecasts can enable bridging the gap between capacity and demand. Equipped with this information, organizations can better plan and allocate resources and quickly respond to any issues. Furthermore, combining visibility with demand management and marketing tactics, organizations can diminish uncertainty, increase predictability, and thus, enable better management of capacity.

To sum up, businesses must employ a broad range of tools to successfully produce and deliver goods and services to their customers: They need data and analytical tools to assess the supply chain as a whole and create good forecasts that connect economic indicators, current sales and other causal variables, and demand management and marketing tools to adjust prices, promotions and contract terms to address changing market conditions, and good capacity management tools to address the changes in demand.

Conclusions

The proliferation of products, build-to-order, shorter order-to-delivery cycles and increased personalization create highly variable demand patterns creating high uncertainties in supply chains. The disruptions in supply chains, due to pandemics, natural disasters, transportation failures, product problems, cyber-attacks, etc., compound this variability. Unfortunately, variability always comes with a cost. When there is variability in a system it has to be managed and one has to pay for it somehow. This can be thru wasted capacity, lost throughput, increased cycle times, higher inventory levels, long lead times and/or poor customer service.

We studied how to utilize capacity as a lever to effectively operate under uncertainty. Managing capacity effectively is not a simple task and there is not a universal rule that fits all situations. Many factors, varying from competitive priorities to costs, need to be considered and depending on the situation the right action should be taken. We have come up with insights that tell us how this can be done under different conditions. In the face of uncertainty, by monitoring the outstanding order queue, one can effectively manage capacity through a class of simple policies that are easy to implement. Furthermore, a broader set of tools are available to successfully steer a business and help bridge the gap between demand and capacity. Using these tools together will enable better management of capacity challenges.

Author

Melda Ormeci Matoglu is an Assistant Professor at Peter T. Paul College of Business and Economics at the University of New Hampshire. Her research interests span both fields of optimization and stochastic control, with applications mainly within the areas of supply chain management and logistics. Her work covers theoretical problems dealing with optimal stochastic control of Brownian motion along with practical applications of these problems. Her work on supply chain problems is mainly related to inventory and capacity management and varies from managing transatlantic supply chains to optimizing point of fit in the production line. Her research is published in Operations Research, Stochastic Systems, Annals of Operations Research, Journal of Operational Research, Advances in Applied Probability.
email: Melda.OrmeciMatoglu@unh.edu

Endnotes

1. Hopp, W. J., & Spearman, M. L. (2000). *Factory physics*. Boston: McGraw-Hill.
2. Rogers, Z. S., Rogers, D., & Leuschner, R. (2018). The logistics managers' index. *Rutgers Business Review*, *3*(1), 16-32.
3. Krajewski, L. J., Malhotra, M. K., & Ritzman, L. P. (2016). *Operations management: Processes and supply chains*. Boston: Pearson.
4. Schonberger, R. J. (1982). *Japanese manufacturing techniques: Nine hidden lessons in simplicity*. New York: Free Press.
5. U.S. Division of Labor Force Statistics. (2018, June 7). Contingent and alternative employment arrangements. Washington, D.C.: Bureau of Labor Statistics.
6. Big increase in flexible workers in Netherlands. (2017, April 11). *Industrial Safety and Hygiene News*.
7. Jepsen, M. (Ed.). (2017). *Benchmarking Working Europe 2017*. Brussels, Belgium: European Trade Union Institute.
8. Aktekin, T., & Soyer, R.(2014). Bayesian Analysis of abandonment in call center operations. *Applied Stochastic Models in Business & Industry*, *30*(2), 141-156.
9. Jagerman, D. L., & Melamed, Bn. (2003). Models and approximations for call center design. *Methodology & Computing in Applied Probability, 5*(2), 159-181.
10. Gans, N., Koole, G., & Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2), 79-141.
11. Warehouse management costs. (2011). *Controller's Report*, *5*, 6-9.
12. What is AWS OpsWorks? [Organization website]. (2021). *Amazon Web Services.*
13. Johnston, R., & Clark. G. (2005). *Service operations management: Improving service delivery*. Harlow: Pearson.
14. Sasser, W. E. (1976). Match supply and demand in service industries. *Harvard Business Review*, *54*(6), 133-140.
15. Klassen, K. J., & Rohleder, T. R. (2001). Combining operations and marketing to manage capacity and demand in services. *Service Industries Journal*, *21*(2), 1–30.
16. Ormeci Matoglu, M., Vande Vate, J. H., & Yu, H. (2019). The economic average cost Brownian Control problem. *Advances in Applied Probability*, *51*(1), 300-337.
17. Ormeci Matoglu, M., & Vande Vate, J. H. (2021). The economic average cost Brownian Control problem with proportional changeover costs: The multiple rate case. *Working Paper.*

18. Sheffi, Y. (2007). *The resilient enterprise: Overcoming vulnerability for competitive advantage*. Cambridge, MA: MIT Press.