Lecture 11: Matching Supply and Demand: Impact of Uncertainty

Class Outline:
Dynamic Systems:
- Impact of uncertainty?
- How to deal with it?
- Barber
- Trucking
- Airline
- Catalog
- Semiconductor

Notes on Upcoming Readings
- Just skim reading #40 (Katz, Larson, and Larson, "Prescription for the Waiting in Line Blues: Entertain, Enlighten, and Engage.") – it's less technical than the version of this paper published in *Operations Research.*

- Reading #41 (Maister, "Note on the Management of Queues") shows a glimpse of the mathematics of queuing. Only read it if you're seeking some analytical models of queues; it won't make a difference in course.

Review of Previous Class (Merrill Lynch Case Study):
The Merrill Lynch case is very important because its content transcends our discussion that day into other parts of the course. My views of the Merrill Lynch case: they are trying to change the interface with their customers in many ways and in many places simultaneously. Their financial advisors (FAs) are very successful and they have low turnover among their FA population. Still, the FAs had to worry about their mortgage, kids, etc.; maybe they think "the company should do more to alleviate for us, the uncertainties and the risks surrounding the new program".

One opportunity for improvement is their current approach, where it is possibly insufficient to show FA's some examples of success with the program and expect greater participation in the Supernova program. From the case analysis, it is clear that FA's are working hard and are generally happy with their financial performance; with this new program, their question is likely to be "What has the company done to share the risk with them? They did not invest anything in software, for example, saying that the systems would come later. The company was also not willing for example, to guarantee the FAs their previous year's salary".

Structure induces behavior – like in the "Beer Game". Informal structure could say to the FA's "you're on your own"; which structure in turn induces behavior that necessitates the Supernova program.

Behavior of Dynamic Systems Under Stress

We were talking about one of most fundamental problems in services – the behavior of dynamic systems under stress (under congestion). We tried to explain last time the variability that we observe in the outputs of dynamic processes. What is the magnitude of impact of uncertainty in the inputs and in the environment and how do we deal with it? It's not intuitive, so it needs some careful analysis.

I left you with a question in the last session:

**Waiting Time /
Unsatisfied
customers /
Operating costs**

Utilization rate (r)
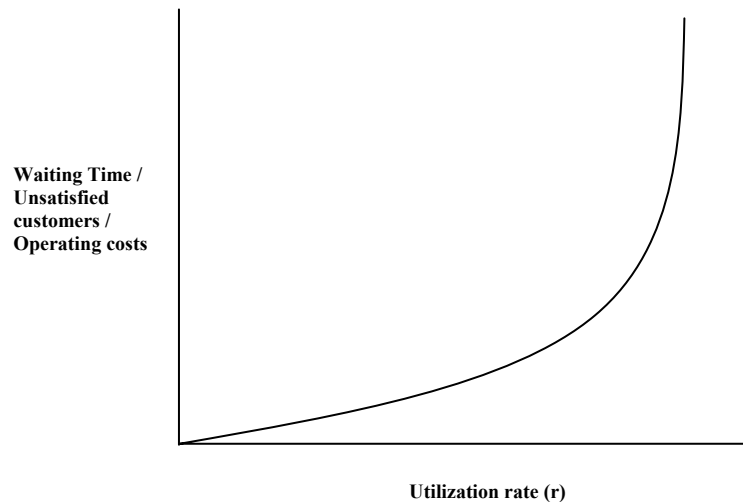
Examples:
- At a cafeteria, you see there is a long line to get lunch. At a later time, the line is much shorter.
- Driving to the airport through a small tunnel and there is serious congestion. It takes a long time to drive through the tunnel, yet at the end, there is only a broken down car. Why was there such a delay for this small event?
- At end of the quarter in manufacturing, you get all kinds of crises. It's often called the "the hockey stick", where we ship much of the merchandise for quarter in the last 3 days.

Why does this happen? How could a manufacturing plant owner not predict it? Actually, he could not because as we will see, dynamic systems with a lot of uncertainty are very unpredictable.

$$P = \frac{\lambda}{\mu} = \frac{1}{2} = 0.5$$

The diagram above represents the length of the queue for different congestions. What kind of curve would we have for increasing λ? It would be an exponential (approximately) curve because we have less ability to recover.

Why does the curve accelerate going up? Remember this is the average queue length: any idle time (when there are no queues) is time that is lost forever because of the perishabilty attribute of services. This leaves very little time to recover. The problem gets more serious as we approach 100 percent utilization: a system at very high utilization has a small probability of becoming idle.

As the system becomes more congested, it becomes loses its ability to recover. If there is 50 percent utilization, most of time there is light traffic. And there are times with no traffic. So even if we have high congestion for 10 minutes, we can clear backlogs and recover in time to meet the next surge in demand. At 75 percent utilization, there are more queues, there is more business, and less idle time for servers. As idle times occur, capacity is lost and the system has less width to recover. At 100 percent, we have no such flexibility or width, so the time in queue increases at a rapid rate.

The left portion of the curve (from 0-50 percent utilization) is almost a straight line. From 25 to 50 percent the rate of increase in waiting costs is almost double. We can use proportionality in low utilization levels, but this does not hold for high utilization (which is most systems), where a small increase in utilization means a large increase in waiting.
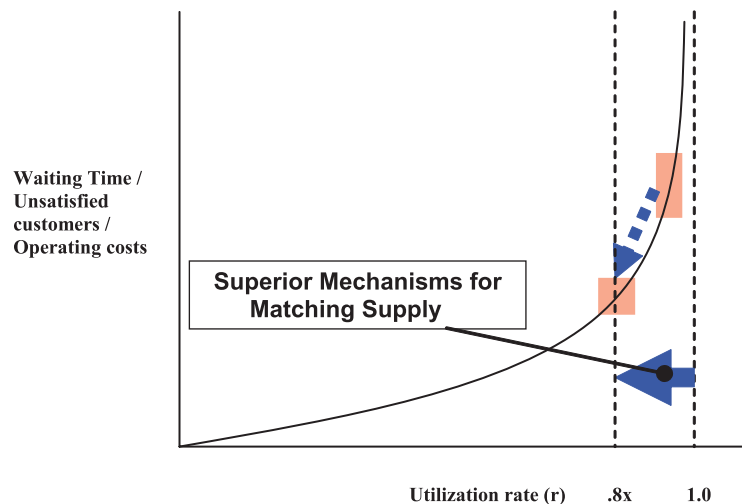
Say a machine breaks down for a few minutes, so we lose a little capacity and utilization goes from 50 to 55 percent. It will create a small delta and increase the waiting line. So, the inventory will be larger by small amount. Suppose instead of being at 50 percent, we were operating at 80 percent and lose the same 5 percent. We're now at 85 percent utilization and have a larger delta. Imagine if at 92 percent utilization, we increase 5 percent – it will become a huge delta and we would have a crisis on our hands. This happens at a factory in the last few days of a quarter when a machine breaks or a supplier cannot supply you, etc. It's beyond our capability to control, so we cannot therefore predict its behavior. An entire supply chain can now be disrupted because of one random event – no wonder companies have problems delivering on time given any complexity in operations.

It's also true with people – you wait a long time at that barber shop if one person wants to color their hair and was not in the schedule. An airline will have phone lines jammed. A catalog company will not have enough customer service representatives. Wafers pile up waiting to be processed in front of an expensive machine for the semiconductor company.

This is all explained by this figure. The villain here is uncertainty. Since the system response to such uncertainty and variability is not linear, it is not part of our intuition.

What Can We Do?
- Increase capacity, which will bring us back to low utilization, but we cannot operate at low utilization because we cannot afford it.
- Manage demand better to spread it out over time – send out a catalog in stages, for example. You can also use incentives to change behavior (have people order earlier). The bottom line is you want to reduce variability –this need and phenomenon is present in marketing, operations, finance, etc.
- You can share risk and variability with competitors, like the airlines do. This is similar in spirit to reducing variability.
- Simplify the system
- Make the system more predictable
- Seek out more information: get rid of uncertainty instead of living with it.

Waiting Time / Unsatisfied customers / Operating costs

Superior Mechanisms for Matching Supply

Utilization rate (r)   .8x   1.0

Suppose I reduce uncertainty. How would the curve look? If I reduce enough uncertainty, the system becomes more stable at higher level of utilization – we typically get the same delta at 50 percent as we do at say 80 percent if we reduce uncertainty and variability in the operations.
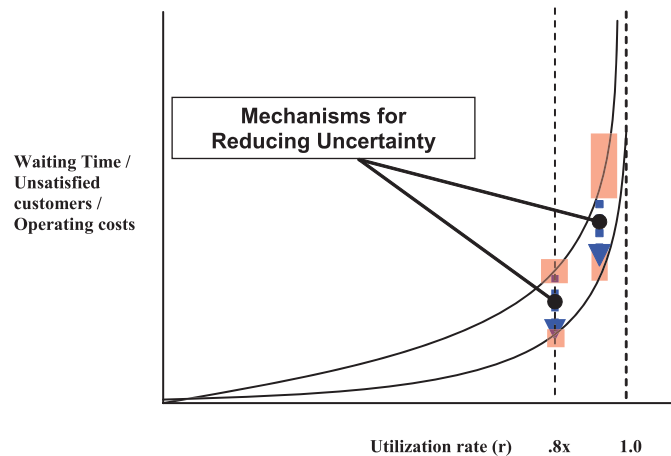
"Just in Time"

What happens if I continue to reduce uncertainty? At the limit, the curve will collapse on the x-axis. At the limit – 100 percent utilization and 0 queue – what do you call it? "Just in time" (JIT).

JIT is the behavior of a dynamic system with very high utilization once we have removed all uncertainty. We need to remove from the whole chain all uncertainty to have JIT – it is very hard to get there since it is a limit condition. So JIT operations typically keep a little inventory on hand just in case.

We would need to do a lot of infrastructure work and identify and eliminate all sources of uncertainty to do JIT. Can we be JIT in services? In principle, yes, but the hard thing is management of uncertainty.

JIT comes at a cost, so we need to balance costs with benefits of JIT. However, the savings from reducing eliminating uncertainty are often significant, they might justify those costs. Notice the recurring theme: if I improve something, it will not necessarily cost more. It is a logic disconnected from practice – we just need do business in a different way, to reinvent ourselves. Our new system can be less expensive and more profitable. We need to stop thinking in a limited way. If we try to reach the stars, at least we won't be caught with our hands in the mud.

Do we want JIT? With all of the uncertainty today (strike, terrorism, natural disasters, etc), JIT operations can suffer if they are implemented without some "just in case" buffers.

Uncertainty

Uncertainty is associated with us human beings – we're not machines. Let's think about some sources of uncertainty:
- suppliers
- machine breakdown
- demand variability/seasonality/consumer behavior

- strike
- war
- weather
- transportation problems/ trucking
- technology forces me to produce big lots
- competition
- regulation/government action
- cash flow problem
- internal systems/communication breakdown
- emerging disruptive technology
- flexibility in production capacity
- etc…

The point is that in a majority of these topics, I can handle most of uncertainty. It's a matter of prioritizing uncertainty, seeing which one has the greater impact. Factories today are much more predictable today, much more under control, which leads to higher quality products. Technology allows smaller lots, preventative maintenance, better training, etc.

Periodic or continuous review systems apply to just a few items. It assumes demand is constant or we order continually. Today, we don't do that any more. These practices existed in the past – we had to make big lots; technology has since come a long way to reducing the need for large lot sizes. This is not the way we do it anymore, though. Historically, that is how we talked about inventory – we used to be unable to handle uncertainty, so we built a lot of inventory. It would take care of some things, but also created other problems: it prevented us from having good data because the orders were all fudged. It was not based on forecasts, but rather we were just anticipating problems.