



# The OM Triangle

**Glen M. Schmidt**

Georgetown University

**Abstract.** A key contribution of OR/MS models is to gain insights into trade-offs facing operations managers. One such trade-off involves capacity, inventory, and variability: while the firm would like to tolerate high levels of variability and run at full capacity utilization with virtually no inventory, this is not plausible. The standard G/G/1 queueing model is used to illustrate and gain insight into the trade-off between these three competing goals. In cases where better information can be used to reduce the variability in services or in arrivals, this insight can be expressed as an inter-relationship between capacity, inventory, and the third parameter of information (rather than variability). Adopting the terminology of Lovejoy (1998), this inter-relationship is referred to as the OM triangle. As discussed herein, it is the manager's job to find her firm's optimal position with regard to the OM triangle.

**Keywords:** queueing theory, capacity, inventory, waiting time, lead time, information, tradeoff.

## The OM Triangle: Instructor's Note

### 1. Introduction

Many popular texts in operations management lack simple examples that illustrate the tradeoff between capacity, inventory, and variability. This key insight is derived herein from queueing theory; more specifically, from the G/G/1 queueing model.

While queueing theory provides sound theoretical background for the derivation of the inter-relationship between these three parameters, the goal of this note is not to teach queueing theory. Instead, it is to bring out qualitative insights that can be used as a guide in managerial decision making. Students should not get lost in the calculations and thereby miss the insights that the theory has to offer. One of these insights is that capacity, inventory, and variability reduction are, in a sense, substitutes. For example, if you reduce variability, you can get by with less capacity and/or less inventory.

In a queueing setting, variability can be found in arrivals and/or in services. These two types of variability can be expressed, for example, by the coefficients of variation in interarrival times and in service times.

In some cases, variability in arrivals and in services can be reduced by the judicious acquisition and use of *information* (broadly defined). For example, rather than taking all patients as walk-ins, a doctor's office might get

information from patients as to their ailments, and then make appointments to smooth out the arrivals. Another example on the demand side is given by Womak et. al (1991), who discuss how new cars in Japan are sold by sales personnel who go door-to-door, thereby gaining the consumer pulse and allowing the firm to reduce the variability in demand.

Likewise, information may in some cases also be used to reduce variability in service times. A simple example might be a sophisticated Department of Motor Vehicles (DMV) office, which pre-screens customers as to the service needed and then routes each customer to the appropriate server (so that while there are differences in service times between servers, each server experiences little variation in service times). The example of automobile manufacturing again offers a more elaborate example. Spear and Bowen (1999, p. 98) describe the Toyota production system (TPS) as follows:

...all work is highly specified as to content, sequence, timing, and outcome...  
The requirement that every activity be specified is the first unstated rule of the system... most managers outside of Toyota and its partners don't take this approach to work design and execution – even when they think they do.

In other words, Toyota first generates precise information as to how every operation is to be performed, and then provides employees with that information and expects them to execute accordingly. If a process is executed the same way every time, the variability in the process is greatly reduced. In other words, information is key to the TPS, and has helped Toyota develop a lean manufacturing system with virtually no inventory and lesser capacity cushion.

In these types of situations, where information can be used to reduce variability, the tradeoff between capacity, inventory, and variability can equivalently be expressed as one between capacity, inventory, and information. Lovejoy (1998) calls the inter-relationship between these latter three parameters the OM triangle. The contribution of this teaching note is to more explicitly show how the OM triangle stems from queueing theory, and to get students to think about how they as (future) managers might themselves manage the tradeoff between capacity, inventory, and information.

## **2. Teaching Experience**

Following this instructor's note is a note that can be included in a course pack or handed out to students. Earlier versions of the note have been used successfully in core operations courses at both the undergraduate business and MBA levels. The note is flexible in that if a purely qualitative treatment is

preferred, then section 2, the “Theoretical Basis for the OM Triangle”, can be ignored.

Students using this note typically have no prior exposure to queueing theory, other than possibly a short teaching note on the topic. They have had prior exposure to the concept of a probability distribution, but typically do not remember terms such as the coefficient of variation. Thus these are reviewed during the discussion of variability.

There are a number of cases that work well to illustrate how the OM triangle applies in practice. Cases that have been used include Shouldice Hospital (Heskett, 2003) and Benihana (Ernst and Schmidt, 2004), which illustrate how firms that minimize variability can run at or near full capacity with low inventory. And be very profitable because of this achievement! The Southwest Airlines case (Frei, 2004) shows how low variability in the time it takes to turn an airplane around at the gate facilitates higher capacity utilization. At another end of the spectrum, the case of Zara (Ferdows, et. al, 2003) is an excellent example of a firm that uses a capacity *cushion*. This mix of cases shows that firms can be highly profitable using a variety of strategies. In other words, there is no one magic position that a firm must choose relative to the OM triangle. Rather, it should be emphasized that the firm’s position on the OM triangle should be consistent with its overall strategy.

### 3. Further Readings and Discussion

For more in-depth treatment of queueing theory itself, see Gross and Harris (1998) or Hopp and Spearman (2000). Alternately, the queueing theory section found in most operations texts can be a good source of further information.

There are many research-related references that discuss how information might be related to a reduction in variability and inventory. Not all of the articles discussed below apply directly to a queueing setting as addressed by the current note, but these references are listed to show more generally how information and variability might be linked or to show how information can in some settings substitute for inventory and/or capacity.

For example, Bourland et. al (1996) discuss the use of passing downstream sales information upstream in the supply chain to possibly dampen the bullwhip effect (defined as the magnification in variability of orders as one progresses up the supply chain). See also Lee et. al (1998). Ijiri and Itami (1971) model the cost penalty for a delay in receiving perfect demand information. Kekre and Mukhopadhyay (1990) empirically investigate the effects of electronic data interchange (EDI) involving a steel firm, finding EDI improves the firm’s inventory position and aids in synchronized manufacturing. In a more theoretical study, Milgrom and Roberts (1988)

consider the tradeoffs between building inventory to meet uncertain demand and surveying some portion of the market (i.e., getting information) to determine demand exactly. Zipkin (1991) discusses many of the tradeoffs in holding inventory.

Importantly, it is useful to note that while information may often be a necessary condition for reducing variability, it may not always be sufficient. To illustrate this, consider an analogy to the field of statistical process control. One might track the variability in interarrival times or in service times using something akin to SPC control charting, and then look for assignable causes for that variability. Information is needed in order to find these assignable causes – thus information may be a necessary condition for removing variability. At the same time, just because the firm knows the cause of variability doesn't mean it will be able to eliminate it. For example, an ambulance service may suspect that a terrorist will at some point create a huge spike in demand for services, but there is no (remotely practical) way the firm itself can remove this potential variability from the system and smooth out the demand.

**References:**

- Bourland, Karla E., Stephen G. Powell, David F. Pyke (1996), "Exploiting timely demand information to reduce inventories", *European Journal of Operational Research*, 92: 239-253.
- Corey, E. R. (1985), "The role of information and communications technology in industrial distribution", in: R.D. Buzzell (ed.), *Marketing in an Electronic Age*, Harvard Business School Press, Cambridge, MA, 29-51.
- Ernst, Ricardo and Glen Schmidt (2005), "Benihana: A New Look at an Old Classic", *Operations Management Education Review* 1(1), 5-28.
- Ferdows, Kasra, Michael Lewis, and Jose A.D. Machuca (2003), "Zara", *Supply Chain Forum*, 4 (2).
- Frei, Frances (2004), "Rapid Rewards at Southwest Airlines", Harvard Business Publishing Case 9-602-065.
- Gross, Donald, and Harris, Carl. (1998), *Fundamentals of Queueing Theory, 3rd Edition*, John Wiley and Sons.
- Heskett, James L. (2003), "Shouldice Hospital Limited", Harvard Business Publishing Case 9-683-068.
- Hilton, R.W. (1981), "The determinants of information value: Synthesizing some general results", *Management Science* 27/1, 57-64.
- Hopp, Wallace, and Mark L. Spearman (2000), *Factory Physics*, Second. Ed., Irwin.
- Ijiri, I., and Itami, H. (1973), "Quadratic cost-volume relationship and timing of demand information", *The Accounting Review*, October, 724-737.
- Kekre, S., and Mukhopadhyay, T. (1990), "Impacts of electronic data interchange on inventory, quality and performance: A field study", Working Paper, GSIA, Carnegie Mellon University.
- Lee, Hau V. Padmanabhan and Seungjung Whang (1998), "Information distortion in a supply chain: The bullwhip effect", *Management Science*, 43 (4).
- Lovejoy, William (1998), *Production and Operations Management*, 7 (2).
- Milgrom, P., and Roberts, J. (1988), "Communication and inventory as substitutes in organizing production", *Scandinavian Journal of Economics* 90/3, 275-289.
- Sasser, Earl W. Jr., and John R. Klug (2004), "Shouldice Hospital", Harvard Business Publishing Case 9-673-057.
- Spear, Steven and H. Kent Bowen (1999), "Decoding the DNA of the Toyota Production System", *Harvard Business Review*, Sep-Oct, 97-106.
- Zipkin, Paul H. (1991), "Does manufacturing need a JIT revolution?", *Harvard Business Review*.
- Womack, James P., Daniel T. Jones, Daniel Roos (1991), *The machine that changed the world*, Perennial.

**Acknowledgements**

The authors appreciate the many useful suggestions of the reviewers and editor, which led to significant improvements.

## The OM Triangle: Note for Students

### 1. Introduction

Say you are in charge of the popular dunking booth at the local amusement park. Customers walk up and throw baseballs at a target and if they hit it, the local celebrity who you have arranged to be in the booth gets dunked. People don't like waiting in line to play, so you are considering investing in a sophisticated "reservation system" like the FASTPASS system used at Disney. The system will give each customer a time when s/he should show up to play, but is heavily dependent on your acquisition of lots of information. It will take into account which celebrity is in the booth, how many people are currently in line, the time of day, how many people have entered (and are yet expected to enter) the park that day, and the profile of each potential customer (to predict whether each will buy several rounds of balls in trying to dunk the celebrity, thereby spending a long time at the booth, or only one round). Should you invest in the information system (i.e., the reservation system)?

Alternately, say you are managing the hamburger joint located inside the student union building. You know that impatient students arrive sporadically and at all times of the night (and day), so you need to stock some pre-made burgers. But stocking this inventory is costly, because you have to throw burgers out if these finished goods sit there too long. How many should you keep on the warming rack?

Or, say you are in charge of running a small hospital. You are currently looking at buying the hospital its first and only X-ray machine. Assume you know that on average a patient will arrive every 30 minutes or so, but not necessarily with perfect regularity. An inexpensive X-ray machine can examine a patient in 25 minutes on average, a moderately expensive one can do so in 20 minutes, and an expensive one only takes 15 minutes. You can only buy one – which one should it be?

In this last case involving the X-ray machine, you are making a decision on how much to *capacity* to acquire. The expensive machine has the capacity to examine four patients per hour, while the inexpensive one can examine only 2.4 per hour. In the hamburger example you are making a decision on how much *inventory* to hold (how many burgers). And in the amusement park example you are making a decision on whether to acquire *information* – the reservation system for the dunking booth uses information to improve customer service.

But if you look at each decision a little more deeply, and consider further possible alternatives associated with each situation, you might conclude that in

each case you could have improved the system by spending money on any one of these three things, capacity, inventory, or information. For example, in the case of the dunking booth, rather than spending money on the information system (i.e., on the reservation system), why not spend that money on capacity, by adding a second dunking booth? Your objective in adding the second booth would not necessarily be to increase the number of customers served but rather to simply reduce the wait time that customers currently experience. Or, why not rather spend your money on “inventory”? In this system, the people waiting in line at the dunking booth effectively represent raw materials waiting to be processed (the customers who have already had their turn represent finished products, and the customer currently throwing baseballs at the target is work-in-process or WIP). You might be able to tolerate higher inventory of raw materials (i.e., longer lines) if, for example, you install TV screens that play video clips of the celebrities in the booth, to help ameliorate the wait. Thus, in this amusement park example, you may really have a choice of spending money on any of these three things, information, capacity, and inventory.

Similarly, consider the hamburger example. With a little brainstorming you could possibly find a way that production capacity or information could be used to reduce the number of burgers that must be held on the warming rack. For example, if you added another worker and another grill (i.e., more capacity), you could quickly fry up a larger number of burgers when the place got busy. Or, you might hire students in various classes to give you a call exactly when a class is dismissed, and find out about how many students from each class comes to your joint, so that you can fry up burgers in anticipation of arriving customers (in effect, you are using this information to “take orders” for burgers even before the customers physically arrive). So again, you may really have a choice of spending money on any of three things, inventory, capacity, or information.

Or, consider the X-ray machine. Rather than paying for a higher capacity machine, you might instead simply accept the cost of holding more inventory. In this case waiting customers again represent raw materials, and patients having completed X-ray are finished goods. The “cost” of more inventory (longer waiting lines) may not be an out-of-pocket cost, but rather the less tangible cost of having disgruntled customers. Yet another alternative might be to spend your money on information. For example, if you knew what medical tests every patient in the clinic needed at every point in time, you might be able develop a scheduling and tracking system that minimizes each customer’s wait.

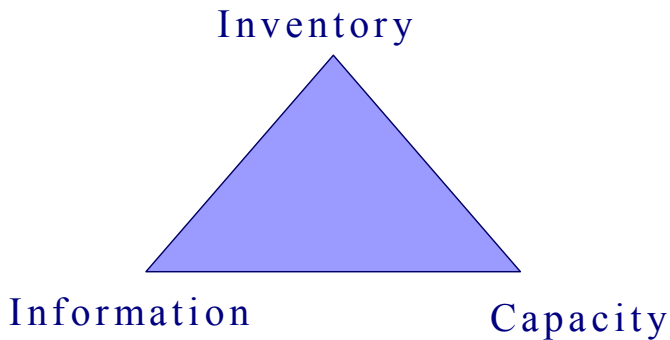
Thus we see that in each case, you have several alternatives for improving system performance. You can do so by spending money on either capacity, or inventory, or information (or some combination of these). Said another way, capacity and inventory and information are in a sense substitutes – if you hold

more of one you may get by with less of another. Of course, there are limitations as to how substitutable one is for another. Your job as a manager is to determine how the tradeoff applies to your unique situation. That is, you must decide how much to spend on capacity, how much to spend on inventory, and how much to spend on information.

The insight that *capacity, inventory, and information are substitutes* is portrayed in Figure 1 as the OM Triangle. In other words, a higher level of capacity can substitute for relatively lower levels of inventory and information, a higher level of inventory can substitute for relatively lower levels of capacity and information, and a higher level of information can substitute for relatively lower levels of capacity and inventory.

Said another way, *there is a tradeoff between capacity, inventory, and information*. The firm can trade off a relatively lower level of one parameter by holding a relatively higher level of another.

Figure 1: The OM Triangle Portrays the Substitutability of Inventory, Capacity, & Information.

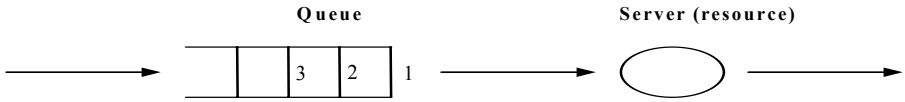


## 2. Theoretical Basis for the OM Triangle

The tradeoffs as described by the OM triangle are grounded in a sound analytical framework, the G/G/1 queueing model. In this section we derive the OM triangle using this model. (If a more qualitative approach is desired, this section may be omitted.)

A brief description of the G/G/1 and the notation used herein is depicted in Figure 2 below. A more comprehensive description of the G/G/1 is given in many Operations textbooks.

Figure 2: The G/G/1 Model.



<u>Arrivals</u>	<u>Services</u>
Arrival rate = $\lambda$	Processing rate = $\mu$
Mean interarrival time = $1 / \lambda$	Mean service (processing) time = $1 / \mu$
Standard deviation in interarrival time = $s_a$	Standard deviation in processing time = $s_s$
Coefficient of variation in interarrival time = $c_a = s_a / (1 / \lambda) = s_a \lambda$	Coefficient of variation in processing time = $c_s = s_s / (1 / \mu) = s_s \mu$
Expected waiting time in queue = $W_q$	Utilization = $\rho = \lambda / \mu$
Expected number in queue = $L_q$	Expected number in service = $\rho$
Expected total waiting time in the system (in the queue and in service) = $W = W_q + 1 / \mu$	
Expected total inventory in the system (in the queue and in service) = $L = L_q + \rho$	

In this model, customers (or jobs) arrive individually and join a queue, to be processed by a single resource (called a “server”). For example, the model can be applied to the dunking booth mentioned earlier. A customer arrives, she joins the queue, and when she gets to the front of the line and the previous customer finishes his turn, she is “served” (meaning she takes her turn at trying to dunk the celebrity).

The expected rate at which customers arrive will be denoted by  $\lambda$  (a Greek letter pronounced lambda), such that the expected time between arrivals is  $1/\lambda$  (if 15 customers arrive every hour, then customers arrive every  $1/15$  hour or every four minutes on average). The times between customer arrivals (i.e., the interarrival times) are not always the same, but vary, with a standard deviation of  $s_a$ . The coefficient of variation in interarrival times,  $c_a$ , is the standard deviation divided by the mean, or  $c_a = s_a / (1/\lambda) = s_a \lambda$ . The expected time that a customer waits in the queue before her turn is  $W_q$ , while the expected number of customers in the queue is  $L_q$ .

The maximum number of customers that can get served (i.e., processed) per unit of time is  $\mu$  (the Greek letter mu), which is the inverse of the average time that the time it takes for each service,  $1/\mu$  (in the dunking booth example, if each person spends two minutes throwing baseballs, then the maximum service rate is  $1/2 = 0.5$  customers per minute). The service times are not always the same (in the dunking booth example, some pick up the baseballs and throw quickly without buying another set, while others may joke around in-between

throws and may buy multiple rounds). The standard deviation in service times is denoted by  $s_s$ , yielding a coefficient of variation of  $c_s = s_s / (1/\mu) = s_s \mu$ .

The fraction of time that the service is busy is denoted by  $\rho$  (the Greek letter rho), and  $\rho$  is calculated as the time it takes for a service divided by the interarrival time,  $\rho = (1/\mu) / (1/\lambda) = \lambda / \mu$  (if a person spends two minutes at the dunking booth, and a person arrives every four minutes, then there is a person at the booth  $2/4 = 50\%$  of the time).

The expected time it takes for a customer (or job) to get through the system, denoted by  $W$ , is the expected time spent waiting in the queue, plus the expected time spent being served. That is,  $W = W_q + 1/\mu$ .

In this system, the items waiting in the queue (if any) are effectively raw materials, and the item in service (if any) represents work-in-process (WIP). There are no finished goods – an item leaves the system when service is completed. (This is typical of a service operation, where the item in service is a person. For example, if the service is a psychological counseling session, you don't find psychologists that have a finished counseling session ready on the shelf to give to a customer – it is infeasible to hold an inventory of finished goods.)

Thus the total expected inventory in the system (i.e., the total expected number of items in the system),  $L$ , is equal to the expected number in the queue,  $L_q$ , plus the expected number in service,  $\rho$ . To show that  $\rho$  is the expected number of items in service, the following logic is applied. There is always either one item in service or none (e.g., in the dunking booth, there is either someone throwing balls at the target, or nobody). When somebody is throwing baseballs, the server is utilized, and when there is no item in the server it is unutilized. Thus the fraction of time the server is busy,  $\rho$ , is also the expected number of items in service. This leads us to the formula  $L = L_q + \rho$ .

In deriving the OM triangle from this model, we will make use of a simple formula known as Little's Law. This Law applies to any queueing system in steady state, meaning that it applies if the number of arrivals into the system is, over the long run, equal to the number of departures from the system. Dr. Little showed that in such a system, the expected number of items in the system is equal to the rate of arrivals to the system multiplied by the expected time spent in the system. In the context of our model, this means that  $L_q = \lambda W_q$ . This means that we can now write our previous formula  $L = L_q + \rho$  as  $L = \lambda W_q + \rho$ . (We simply substituted  $\lambda W_q$  for  $L_q$ .)

Another useful formula gives the expected time that a customer (or job) must wait in the queue before being served. This is known as the modified Pollaczek-Khinchine (PK) formula:

$$W_q = \frac{1}{\mu} \left( \frac{\rho}{1 - \rho} \right) \left( \frac{c_a^2 + c_s^2}{2} \right)$$

Thus, again proceeding by substitution:

$$L = \lambda W_q + \rho = \lambda \left[ \frac{1}{\mu} \left( \frac{\rho}{1-\rho} \right) \left( \frac{c_a^2 + c_s^2}{2} \right) \right] + \rho$$

Simplifying algebraically, we find: 
$$L = \rho \left[ 1 + \left( \frac{\rho}{1-\rho} \right) \right] \left( \frac{c_a^2 + c_s^2}{2} \right)$$

Next define  $V$  as a variability factor,  $V = \frac{c_a^2 + c_s^2}{2}$ , such that

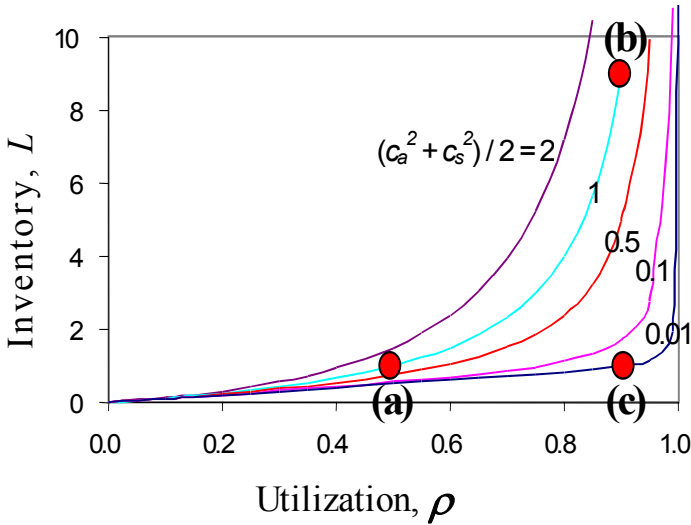
$$L = \rho \left[ 1 + \left( \frac{\rho}{1-\rho} \right) \right] V$$

Notice in the preceding equation that the inventory,  $L$ , is a function of only two variables, the utilization  $\rho$  and the variability factor,  $V$ . Thus we can generate a plot of  $L$  versus  $\rho$ , where  $\rho$  ranges from zero to one (utilization is greater than zero, and can't exceed one), making individual curves (isoquants) for specific levels of variability. Figure 3 shows the results for  $V=2$ ,  $V=1$ ,  $V=0.5$ ,  $V=0.1$ , and  $V=0.01$ .

Figure 3 effectively portrays the tradeoffs between utilization, inventory, and variability. In the upcoming section we will discuss these tradeoffs in the context of points (a), (b), and (c): point (a) is associated with  $\rho=0.5$ ,  $L=1$ , and  $V=1$ ; point (b) is associated with  $\rho=0.9$ ,  $L=9$ , and  $V=1$ ; and point (c) is associated with  $\rho=0.9$ ,  $L=1$ , and  $V=0.01$ .

Another equation following directly from Little's Law is  $L = \lambda W$ , which can be written as  $W = L (1 / \lambda)$ . Thus holding  $\lambda$  constant, the expected waiting time is simply some multiple of the expected number of items in the systems. This means that the curves for  $W$  versus  $\rho$  would look similar to (have the same shape as) those for  $L$  versus  $\rho$ . In other words, increasing variability and increasing utilization have the same detrimental effect on lead-time as they have on inventory level. (Here we think of lead-time as the expected total time it takes an item to get through the system – in other contexts, we may attach a slightly different meaning to lead-time, such as the time the firm *quotes* between receipt and delivery of an order, which may be different from the average time it actually takes.)

Figure 3: The Tradeoffs as Determined by the G/G/1



**3. The Capacity Position (a), Inventory Position (b), and (Low) Variability Position (c) Form the Triangle**

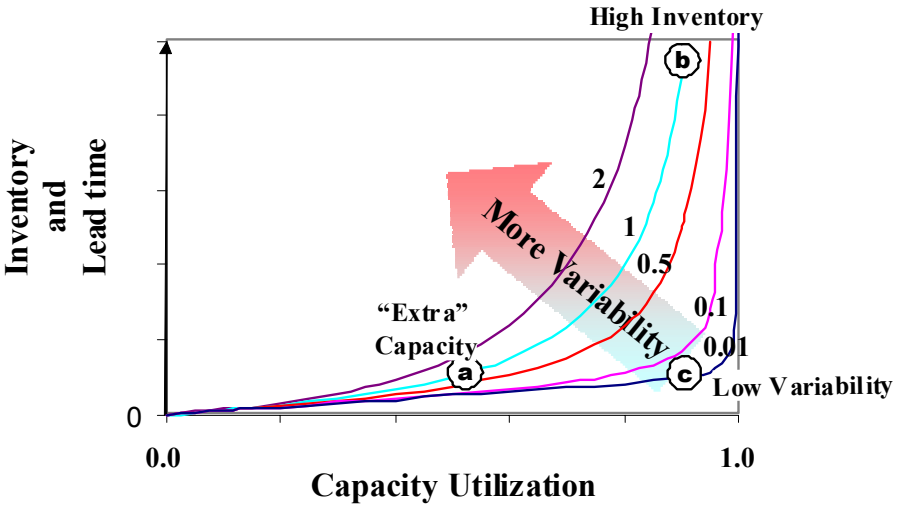
Figure 4 presents the results formally derived in the previous section. It is a very powerful figure, once it is understood.

Figure 4 shows the relationship between capacity utilization on the x axis, and inventory on the y axis (alternately, the y axis can be considered to be lead-time, since higher inventory is directly associated with higher lead-time). This relationship is shown as a series of curves, where each curve represents a specific level of variability, namely, variability levels of 0.01, 0.1, 0.5, 1, and 2. The manner in which these numeric variability factors are determined is described in the previous section – suffice it here to say that variability = 0.01 is extremely low and variability = 2 is quite high, and that this variability factor takes into account both the variability in demand (since customers don’t generally arrive with precise regularity), and the variability in service times (since there is generally some uncertainty in how long each job task will take).

Each possible point on the two-dimensional map given in Figure 4 represents a spot at which the firm might choose to position itself. Conceptually, every firm effectively positions itself *somewhere* on this map.<sup>1</sup> For example it may find itself at point (a), or at point (b), or at point (c). Each point describes a specific level of capacity utilization (graphed in the x dimension), a specific level of inventory/lead-time (graphed in the y

dimension), and a specific level of variability (determined by which curve the point lies on, ranging from variability = 0.01 to variability = 2). In effect, we see that the three factors of utilization, inventory/lead-time, and variability are inter-related.

Figure 4: The Firm Positions Itself at a Specific Point on this Map



The firm has some control over whether it is positioned at point (a), (b), or (c). (To reiterate, the firm is not limited to these three points – it could locate itself *anywhere* on the map. We simply refer to points (a), (b), and (c) to facilitate the discussion.) For example, say the firm operates in an environment with a level of variability equal to “one”. If it acquires a minimal amount of capacity, such that utilization = 0.9, it finds itself at point (b). By buying more capacity, it can reduce capacity utilization and move itself down along the curve representing a variability level of 1, and by purchasing enough additional capacity, it can reach point (a). By doing so, note that it will dramatically reduce its inventory and lead-time.

To make the insights more concrete, say we run a hospital. Assume we face a relatively high level of variability, equal to a value of one (this might represent the case where we run an emergency room, where inter-arrival and service times are random<sup>2</sup>). Assume the expected demand for our service is 2

1. Of course, it is unlikely that any firm rigorously fits the description of a single-server G/G/1 system of the type we have assumed in generating Figure 4. However, this simple model is still often used to lend some insight into the general behavior of more complex systems.

customers per hour.<sup>3</sup> If we choose to position our firm at point (a), then to achieve the capacity utilization of 50% that is associated with point (a) we need a throughput capacity equal to twice the customer demand, or a capacity of 4 customers / hour.<sup>4</sup> Our customers will not have to wait very long for service, and we will have very few customers waiting<sup>5</sup>, as indicated by our low vertical position on the  $y$  axis.

An alternate position for our hospital is point (b). Here we allow a much higher utilization of 0.9, indicating that we can get by with much less capacity while still accommodating the same demand of 2 customers per hour.<sup>6</sup> Since capacity costs money (more capacity means more equipment, a bigger facility, and more doctors and nurses, for example), this position may *appear* to be much better for us. But unfortunately, we can't "have our cake and eat it too". When we run at this high utilization level, given that we experience a high level of variability (equal to one), we must tolerate a high level of inventory<sup>7</sup>. Since in this example the inventory consists of patients (either in treatment or waiting for treatment), we need to consider whether this high inventory level is acceptable.<sup>8</sup>

In the third position, at point (c), our hospital is again able to accommodate the demand of 2 customers per hour with the lower investment in resources (i.e., the lower capacity). Furthermore, we do so without being forced to hold a high level of inventory (our vertical position on the  $y$  axis is relatively low<sup>9</sup>). But to locate itself at point (c), the hospital must reduce variability to an extremely low level of 0.01. How can the hospital reduce variability? Well, what if the doctors were better informed regarding the diagnosis and treatment of medical problems? What if the hospital had better information regarding the type of treatment the patients needed, even before the patients arrived? What if the hospital had better information regarding the minute-to-minute availability of, say, its X-ray machine, such that it could schedule patients in a more efficient sequence? All of these factors might help reduce variability, and all are associated with better INFORMATION. Since better information (or knowledge) is often a key factor in reducing variability, for purposes of this discussion *we will associate lower variability with better information.*<sup>10</sup>

- 
2. Footnotes will be used to quantify the analysis for those who wish to "see the numbers". In this reading we associate the term "random" with the exponential distribution, implying that for random arrivals we have  $c_a = 1$  and for random services we have  $c_s = 1$ , such that  $(c_a^2 + c_s^2) / 2 = 1$ .
  3. This means  $\lambda = 2$  customers / hour.
  4. This means  $\mu = 4$  customers / hour.
  5.  $L = 1$ : Can you calculate the expected waiting time in the queue?
  6. Specifically, we only need a throughput capacity of  $\mu = 2.22$  customers per hour, as calculated from  $\mu = \lambda / \rho = (2 \text{ customers/hr}) / 0.9 = 2.22 \text{ customers/hr}$ .
  7.  $L = 9$ .
  8. Since we can serve only one customer at a time, this means there are, on average, roughly 8 customers waiting to be treated (more precisely,  $L_q = L - \rho = 8.1$ ).
  9.  $L = 1$ .

#### 4. The OM Triangle: Capacity, Inventory, and Information

As alluded to in the above hospital example, the ability to reduce variability is often associated with more and/or better use of information. As another example, consider an airline company such as Southwest. When one of its flight arrives at the airport, it pulls into a gate that Southwest rents from the airport authority. Because Southwest pays a steep rental price for this gate, Southwest wants to rent as few gates as possible: to achieve this it must keep the gates that it does rent fully utilized. At the same time, it cannot tolerate much inventory – in this queueing system, inventory is a plane that must sit out on the runway and wait until a gate is open (or worse yet, circle in the air and wait). Given that Southwest wants to simultaneously achieve high capacity utilization and low inventory, low variability is required (see Figure 4). To achieve low variability with regard to “services” (i.e., with regard to “turning the plane around” and getting it airborne again), Southwest must have a regimented routine for doing so; a routine that does NOT involve unscheduled maintenance. To avoid unscheduled maintenance to the extent possible, and to maintain a reasonable turn-around time even when unscheduled maintenance is mandated, Southwest must have information. For example, it might be good to know that it is better to change all the light bulbs every weekend so that a bulb never burns out during a more heavily scheduled weekday. Further, in case unscheduled maintenance of some type is needed, it would be useful to know what parts to stock on hand and how to most quickly install these parts. In other words, low variability is associated with information.

By equating less variability with more and/or better information, we can express the tradeoff as one involving capacity, inventory and information (we simply replace the term variability with its equivalent, information). Thus, we see the firm can accomplish its goal in any one of three ways:

1. By holding a high level of *capacity*, i.e., by positioning itself at point (a).
2. By holding a high level of *inventory*, i.e., by positioning itself at point (b).
3. By holding a high level of *information*, i.e., by positioning itself at point (c).

In other words, *capacity, inventory, and information are substitutes*. Alternately, we can say *there is a tradeoff between capacity, inventory, and*

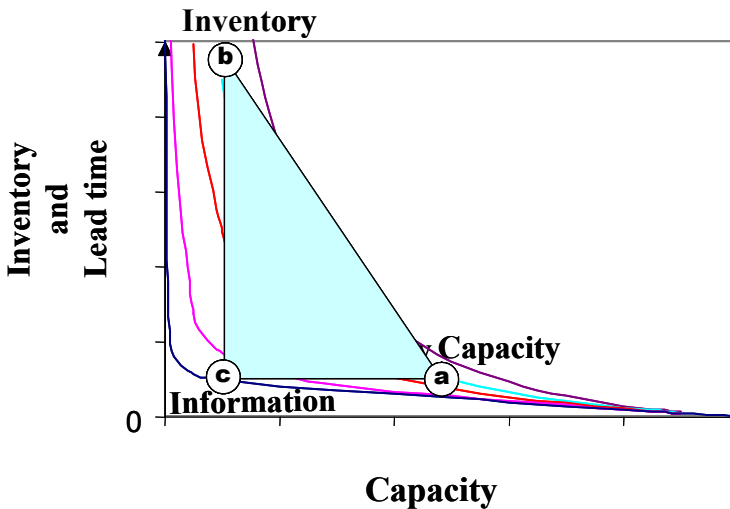
---

10. This is not meant to suggest that the ONLY way to achieve lower variability is to improve the level of information, just that this seems to be a COMMON way.

*information*: The firm can trade off a relatively lower level of one parameter by holding a relatively higher level of another.

This key insight (the substitutability of capacity, inventory, and information) is known as *the OM triangle*. This concept is illustrated in Figure 5. Note its similarity to Figures 1 and 4: in comparison to Figure 4, the *x*-axis is labeled as capacity rather than as utilization, and since utilization goes down as capacity goes up, Figure 5 is a mirror image of Figure 4. The apex of the triangle at point (a) is labeled as the *capacity* point, since at this point capacity is high, while inventory and information levels are relatively lower. The apex of the triangle at point (b) is labeled as the *inventory* point, since at this point inventory is high, while capacity and information levels are relatively lower. The apex of the triangle at point (c) is labeled as the *information* point, since at this point the level of information is high, while capacity and inventory levels are relatively lower. (To reiterate, the firm's choice is not limited to points (a), (b), and (c), or even within the boundaries of the triangle shown.)

Figure 5. The OM Triangle – Where Should the Firm Position Itself?



## 5. The Optimal Position on the OM Triangle

What is the optimal point for the firm? It is YOUR job, as a manager, or as a management consultant, or as a financial analyst, to figure out the best point for the firm. The answer will differ depending on the firm's strategy, the cost

structure of the industry, the availability of information, and many other factors. If the firm does not position itself in the *right* spot, it may be keeping customers waiting longer than it should. Or, its inventory carrying costs may be too high. Or, it may have spent too much on capacity acquisition. In other words, it will not be achieving as much profit as it could be making.

Another way of asking the positioning question is to ask, “where should the firm spend its money?” Should it spend it on acquiring capacity, or should it spend it on acquiring information, or should it spend it on inventory? (Spending it on inventory may in some cases be an opportunity cost. For example, if the inventory is represented by waiting customers, then the cost of holding too much inventory may be lost sales.)

Returning to our hospital example, in an emergency room situation it may be unacceptable to hold a high level of inventory (customers cannot wait for service). In this same emergency room, reducing the variability in arrivals and service times may be impractical, since the firm cannot get prior information regarding what treatment will be required. Thus, the best position may be one where the firm holds “excess” capacity, at point (a). (Here we put the term “excess” in quotes because it is *optimal* to have some capacity that generally sits idle. In other words, the unused capacity is not really “excess”.)

Another hospital may strategically choose to cater strictly to customers who need one specific type of treatment, say a hernia operation. This hospital may be able to get prior information regarding a patient’s needed treatment even before the patient arrives at the hospital (from a mailed questionnaire, for example). This information might allow the hospital to dramatically reduce its variability in demand and in service times, such that it might be optimal for *this* hospital to operate at point (c). Here, it can efficiently use its capacity, and at the same time minimize the wait time for patients. Shouldice Hospital seems to have been able to position itself at this operating point (reference Harvard case HBS 9-673-057).

A third hospital may choose to provide lower-cost, non-emergency service to a group of patients who they think are not particularly time-sensitive, and who seem to be more willing to wait for treatment. Such a hospital might choose to operate at point (b). Sometimes I think this is where the HMO that I go to is positioned! Maybe you go to a student clinic that seems to fit this description?

Thus we see the firm’s choice of an optimal position on the OM triangle is determined by a number of factors. First, it must position itself in concert with its strategic objective: does it strive to offer the quickest turn-around time, or does it serve a customer base where cost control is more important? Does it reduce the variety of product offerings to the extent that variability is minimal by nature? Or does it cater to market that demands immediate, customized service, and is willing to pay a premium for that service?

In addition to strategic objective, the firm must consider the relative costs of capacity, information, and inventory. Which of these is hard to get at any price, and which is relatively inexpensive? If capacity is relatively cheap, why spend money on information or inventory? If inventory is not perishable and readily storable, why spend money on capacity or information? If information can be readily gained, analyzed, and can be used to reduce variability, why not run at near-capacity levels and store virtually no inventory?

Examples of industries that operate at point (b), the inventory point, are those that have high fixed costs of capacity but some variability in demand and/or processing. These include many industries that make commodities, such as petrochemicals, paper, oil, steel, and cement. Firms in these industries try to keep their plants running at all times (at nearly 100% utilization), often operating round-the-clock (24 hours per day). The end products are relatively less perishable, so they use inventory stocks to smooth production in spite of some variability in demand.

Examples of industries that operate at point (a), with a large capacity cushion, are those that must provide emergency service. Can you imagine a fire station that continually fails to respond to an emergency call because all its trucks are already out on other calls? There is a reason that many firefighters have idle time at the station – they are “underutilized” for a large portion of the time because of the uncertainty of when the next emergency call will be received, but must be available to immediately respond to an emergency (that’s why it’s called an emergency – if they respond late, they may as well not respond at all).

Examples of industries and firms that operate at point (c), with low variability and nearly 100% capacity, are those that closely manage demand. For example, Toyota developed its lean production system in Japan, where the car buying experience was quite different than in the U.S. In Japan, car salespersons had a very close pulse on individual purchase decisions, such that they had good information as to what the demand rate would be. This allowed them to develop a system with steady production, close to 100% capacity utilization, and yet virtually no inventory. Other firms that can achieve this type of efficiency are firms that can make appointments and then schedule the appointments to closely fill out all the available appointment slots.