

A Note on Managing Waiting Lines¹

Professor Edward Anderson
UT McCombs School of Business

Waiting Lines—The Big Picture

Managing waiting lines create a great dilemma for managers seeking to improve the return on investment of their operations. On the one hand, customers dislike waiting intensely. If they feel they are waiting too long at your firm for service, they will either leave the line prematurely or not return to your firm the next time they need service. This will reduce customer demand and eventually revenue and profit. Furthermore, longer waiting times increase costs because longer waiting times equal more customers in a firm's building. Hence, a firm will need more space for the customers to wait in, which increases rent.

On the other hand, as we shall see, managers primarily reduce waiting times by increasing capacity, which is itself quite expensive and will reduce profit. Finding a waiting time that customers find acceptable while keeping utilization reasonably high is thus of critical operational importance but relatively unintuitive, for it turns out that average waiting times can be quite long even when capacity is significantly greater than demand.

Why are we always waiting?

When the demand for a service exceeds the capacity of that service, waiting is unsurprising and inevitable. Surprisingly, however, even when process utilization is less than 100%, there can still be waiting on average. This can occur for a number of reasons. One classic situation occurred at a health care clinic in which the bulk of patients arrived when the clinic opened at 8:00 am, but only 2 of the 6 physicians arrived at that time. Those 4 arrived at 10:00 am. Hence, because the patients had a two-hour "head-start," the line necessarily built up and waiting times increased during the first two hours.

A more common problem, however, is when capacity remains constant, but demand fluctuates. Consider Figure 1. Capacity is greater than the average customer arrival rate throughout the figure, but firms cannot perform services such as haircuts, medical appointments, and car repairs "in advance" during periods A and C, when the capacity exceeds the arrival rate. Hence, they are unable to create "inventories" of services to await future customer arrivals.

¹ Special thanks to Mary Ann Anderson, Wonsuk Doh, and Loutfallah Farah for proofreading this document and making suggestions for its improvement. Any remaining errors are the author's responsibility.

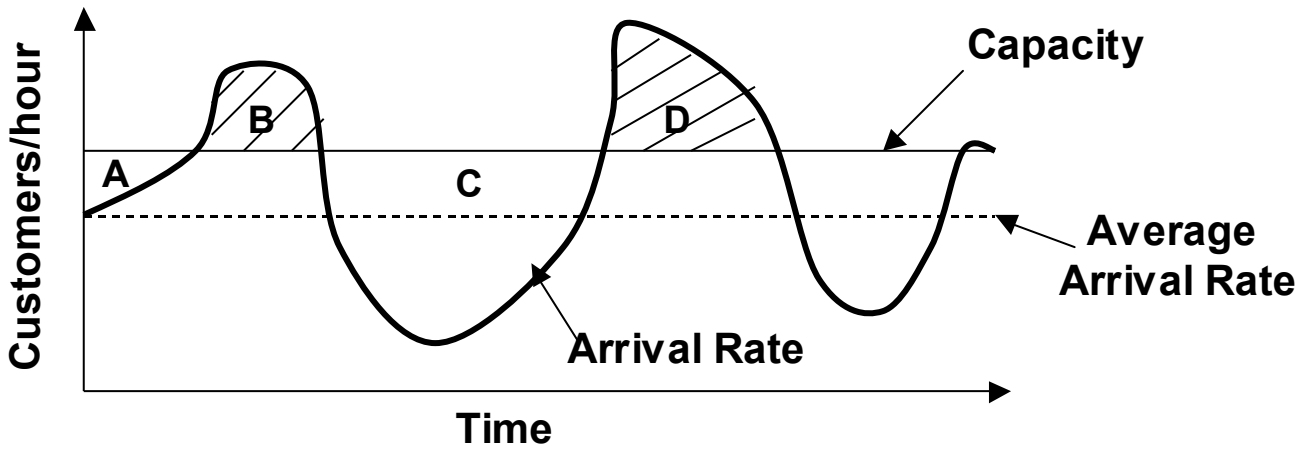


Figure 1

So any customer arriving after the first customer to be served during intervals **B** and **D**, when the arrival rate temporarily exceeds capacity, will increase the length of the line and the average wait. Thus, while *on average* utilization is less than 100% (because the average arrival rate is less than the process capacity), the average customer will still experience a waiting time greater than zero because of variability of “lumpiness” in the arrival rate.

Estimating Waiting Time and Line Lengths

It would be very helpful if we could figure out a way to answer questions like the following: “If the tellers at my bank have an individual capacity of 12 customers per hour and I want my customers to wait no more than an average of five minutes before being served by a teller, how many tellers do I need on duty?” Mathematicians and engineers beginning with A. K. Erlang, a Danish telephone engineer, have developed a methodology, called queuing theory, to answer these sorts of questions. (Queue is the British word for a waiting line.)

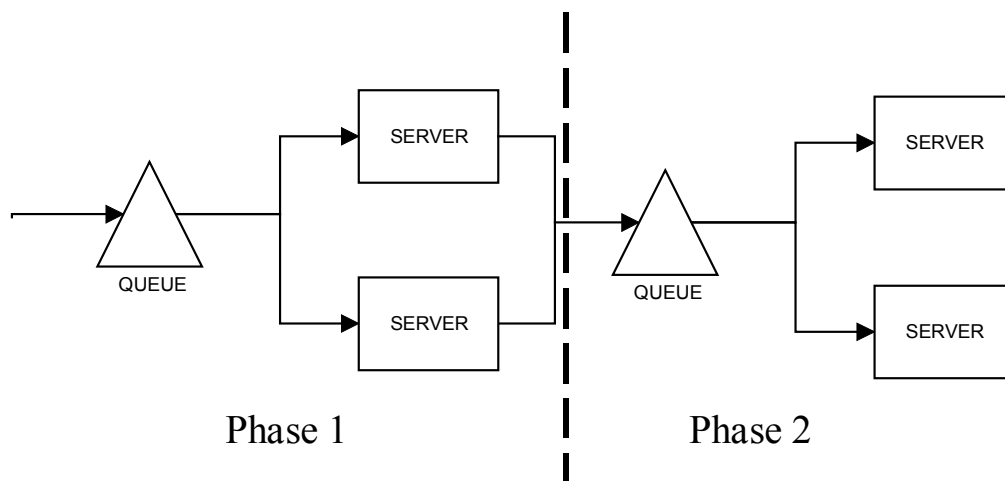


Figure 2

Before we can learn how queueing theory answers these questions, we need to learn some vocabulary. These terms are summarized below, and their common abbreviations are in parentheses. Referring to Figure 2 will help clarify their definition.

Queue:	A line (or buffer or inventory) feeding a number of servers
Server:	An operation fed by a queue.
Arrival rate (λ):	Mean number of arrivals per unit time (usually per hour or day).
Service rate (μ):	Mean number of customers that can be served at 100% utilization <i>by each individual server</i> per unit time (usually per hour or day). At the individual workstation level, the service rate will equal capacity.
Channels (M):	The number of parallel operations connected to an individual queue. In Figure 2, each queue has 2 operations and hence two channels.
Utilization (u):	A measure of how “busy” the system is. It is generally defined as the ratio of thruput to capacity. Note that $u = \lambda/(M\mu)$ if $\lambda < M\mu$, i.e. the utilization is less than 100%. (Also, note that while the Greek letter μ —or mu—looks a bit like u , they are in fact two different variables.)
Phase:	A queue and its connected servers. The portion of the flowchart to the left of the dashed line in Figure 2 is one phase; the portion to the right is another phase. Taken together, they make a two phase system.
Balking:	When a person, who would otherwise have entered a line, decides not to enter it.
Reneging:	When a person, who has entered a line, later decides to leave it without being served.
Interarrival Time:	The time between when one customer arrives at a queue and when the next customer arrives.
Service Time:	The time it takes for one particular server to complete a customer’s service. The average service time will be the same as the cycle time.
CV :	The coefficient of variation. This is a measure of a random variable’s variability. For a random variable x , CV_x is defined as $CV_x = \frac{\text{Std. Deviation}(x)}{\text{mean}(x)}$.
CV_{LAT} :	The coefficient of variation of the interarrival time. The greater the CV_{LAT} , the “lumpier” the arrival rate.
CV_{ST} :	The coefficient of variation of the service time. The smaller the CV_{ST} , the more “consistent” a server is.
L_q :	The average number of people in a line awaiting service.
W_q :	The average length of time a customer waits before being served.

Line Lengths and Waiting Times (Little’s Law)

Actually, we can answer some questions about waiting times without knowing anything other than the average line length and the average customer arrival rate. To see this, look at Figure 3. If a customer joins the line just after a customer begins to be served at the vending machine, then intuitively one would expect the newly arriving customer to wait (Line Length)(Cycle Time) = (8 customers)(1 min/customer) = 8 minutes. If the line length is doubled to 16 people, then the waiting time should be (16 customers)(1 min/customer) = 16 minutes. Similarly, doubling the cycle time to 2 minutes should also raise the waiting time to 16 minutes.

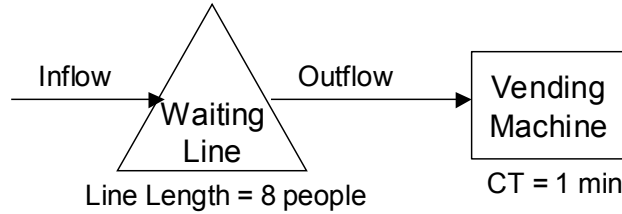


Figure 3

In fact, this intuition is not only correct, but only part of an even deeper relationship. For example, if the cycle time were 0.5 minutes half the time and 1.5 minutes the other half, then the expected waiting time would be $(8 \text{ people})(50\% * 0.5 \text{ min/person} + 50\% * 1.5 \text{ min/person}) = (8 \text{ people})(1 \text{ min/person}) = 8 \text{ minutes}$. This suggests that, if there are no starvations or blockages

$$W_q = L_q * CT . \tag{1.1}$$

And because the cycle time is inverse of capacity

$$W_q = L_q * \frac{1}{Capacity} = \frac{L_q}{Capacity} . \tag{1.2}$$

We have almost arrived at Little’s Law, but given that there are often starvations and blockages of the queue, one needs to substitute throughput (or average outflow from the queue) for capacity. Then,

Little’s Law
$$W_q = \frac{L_q}{Thruput} . \tag{1.3}$$

Formula (1.3) above is known as Little’s Law and can be applied in any system in which the mean waiting time, mean line length (or inventory size), and mean thruput (outflow) remain constant. Recall that the thruput is always measured as the number of items leaving the system—not entering it. To some extent this is an arbitrary decision, but in most real-world situations, measuring the outflow of a queue is easier than measuring its inflow.

Another interesting point is the generality of this formula. For one thing, this relation will hold no matter what the distribution of interarrival times or processing times is. Even more amazingly, Little’s law is not restricted to simple systems with one line and a number of servers. *It will hold no matter what the internal structure of a system is.* So if we were asked what the average time in the system for a patient at a hospital, with all the multiple phases that entails, and were told that the average number of patients was 102.5 and the average discharge rate was 67.5 patients/day, we would not need any more information. We could simply answer (the q subscripts are left off of L_q and W_q because we are looking at a complex system with multiple queues and servers).

:

$$W = \frac{L}{\text{Thruput}} \Rightarrow \text{Avg. Time in Hospital} = \frac{\text{Avg. \# Patients}}{\text{Avg. Discharge Rate}} = \frac{102.5 \text{ patients}}{67.2 \text{ patients / day}} = 1.53 \text{ days.} \quad (1.4)$$

A final point on using Little's Law with queueing problems. When using queueing theory, to get any usable results whatsoever, process utilization must be less than 100%. Otherwise, the line would grow forever. Hence, what flows into the queue must eventually flow out. So over time the average inflow will equal the average outflow. Using this relation, the variation of the Little's Law most often used in queueing problems is:

$$W_q = \frac{L_q}{\lambda} \quad (\text{so long as } \textit{utilization} < 100\%). \quad (1.5)$$

The Big Queueing Formula

But what happens if we don't know either the line length or the waiting time, but only the capacity of the system? It turns out that this is enough to calculate average line length and waiting time—if we make certain assumptions.² An empirically derived formula (Sakesagawa 1982) can estimate line length quite accurately as long as we know the arrival rate, the service rate, the number of servers, and the coefficients of interarrival and service times. As long as utilization is less than 100%, no balking or renegeing occurs, and there is no “ceiling” or maximum on line length at any particular moment, then the length of the waiting line is

$$L_q \approx \frac{u^{\sqrt{2(M+1)}}}{1-u} * \frac{(CV_{IAT})^2 + (CV_{ST})^2}{2}. \quad (1.6)$$

Examining this powerful formula, we can see a number of interesting managerial relationships:

- The line length (and by Little's Law waiting time) increases with utilization, and will “blow up” if utilization reaches 100%.
- Increasing capacity by increasing the number of channels (M) helps, but there is less of a return in line length reduction for every additional channel added.
- The “lumpier” the arrival rate, the longer the line.
- The more consistent the process (in terms of service time), the shorter the line.

Some of the results of these relationships can be seen graphically in Figure 4 below.

² In most MBA textbooks, finding line lengths with this data is typically done by using a complex table that gives results for utilizations such as 85%, 90%, etc. The Sakasegawa assumption, however, estimates line length results quite accurately without resort to a table except when line lengths are so small as to be of little managerial relevance. More importantly, it will allow us to deal with utilizations of 96%, 97%, and others that are typically not listed in queueing tables.

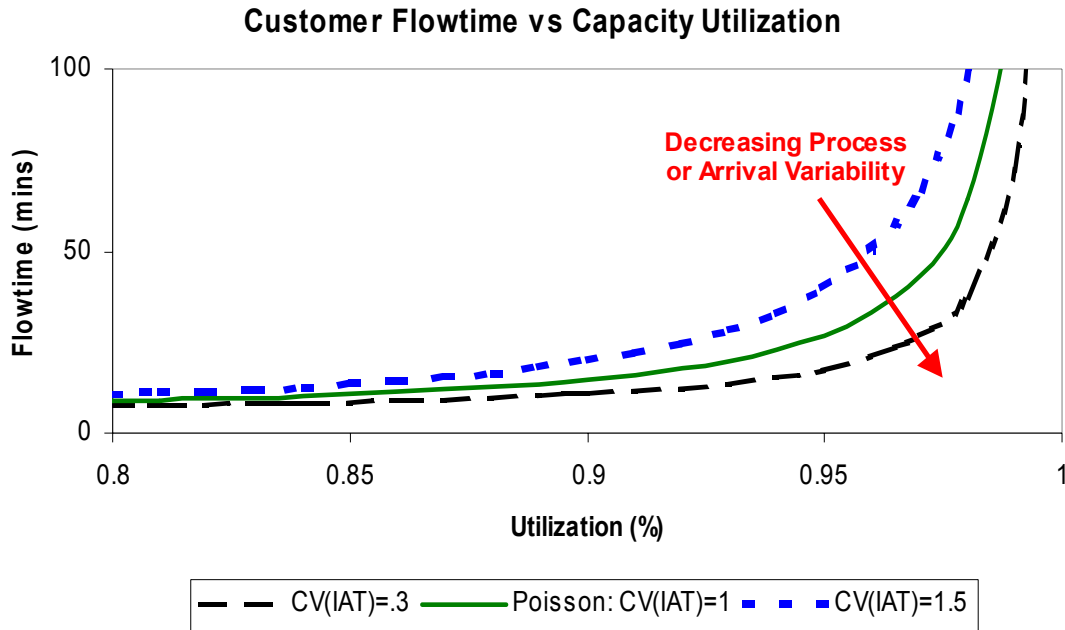


Figure 4

Figure 4 shows the typical tradeoff between capacity utilization and customer waiting times. However, it also shows that a manager can improve this tradeoff by reducing the variability in the system. This can be done by either reducing the variability in interarrival times (through appointments, scheduling, etc.) or service times (through process improvement) or both.

The Small Queuing Formula

Often the hardest problem for a manager who wants to use the formula above is finding out the two coefficients of variation used in the formula. In practice, however, two assumptions are commonly made that simplify the issue. The typical assumption for the arrival process is that no two customers coordinate their arrival plans. From empirical studies, this turns out to be a fairly accurate approximation under many common circumstances. This assumption results in a so-called “Poisson” arrival process and implies that interarrival times have an exponential distribution, which in turn means that the coefficient of variation of customer interarrival times is 100%. Intuitively, what this means is that typical interarrival times are much more variable than anything that can be described by a normal “bell-curve” distribution, so the arrivals will not appear to have much “rhythm.”

The most typical assumption for the distribution of service times is that they are also exponentially distributed and, hence, also have a CV of 100%. From empirical studies, this turns out to be a bit on the conservative side, but not unreasonable. This assumption in combination with the Poisson arrival assumption does provide some benefits, however. For one, as we will discover later, the two assumptions permit the manager to study complex, multiple line systems relatively easily. More importantly, they also simplify formula (1.6) drastically to:

$$L_q \approx \frac{u \sqrt{2(M+1)}}{1-u} \tag{1.7}$$

(For math whizzes, the approximation above is, in fact, exact for lines that feed one server.) Because of its simplicity and usefulness, we will most often use this formula to solve queueing problems in this class.

Multiple Line Systems

If we are analyzing a system that is characterized throughout by Poisson arrival processes and exponentially distributed service times, there are two other additional facts that can help us analyze systems with multiple lines.

- If two Poisson processes merge, the result is a single Poisson process whose mean arrival rate is simply the sum of the mean arrival rates of the two merged processes. Similarly if some percentage of arrivals (say $x\%$) are randomly diverted from a Poisson process, the diversions will themselves also be a Poisson process with a mean arrival rate that is $x\%$ of the original arrival process mean.
- Jackson's Law: If the arrival rate to a line is a Poisson process and the line is served by workstations with exponentially distributed service times, the outflow from each server will also be a Poisson process.

To get a feel for how these facts help us, let's look at a barbecue restaurant's flowchart in Figure 5.

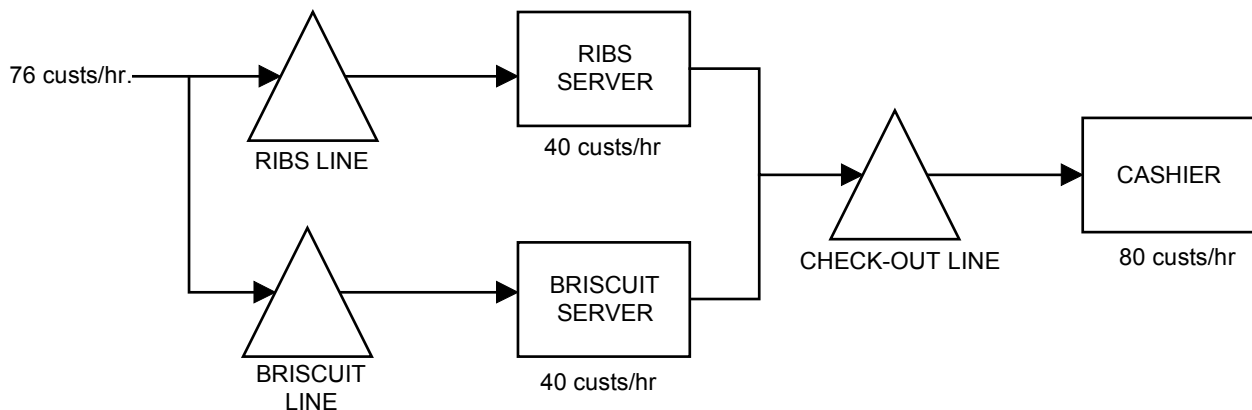


Figure 5

76 customers per hour arrive at a barbecue restaurant that serves briscuit and ribs. There is a separate line for each. We'll assume that half of the restaurant's customers decide to get ribs and half get briscuit, but that they all go to single cashier before leaving. None of the customers coordinate their arrivals with each other, implying a Poisson arrival process, and no-one balks or reneges. Service times at all workstations are exponentially distributed, and line lengths never get long enough to run into any physical limitations. What are the waiting times in each line, and what is the flowtime for the entire system?

Realizing that the utilization for the ribs line is $\lambda/\mu = [(50\%)(76 \text{ custs/hr})]/[40 \text{ custs/hr}] = 95\%$, the line length is

$$L_{q,briscuit} \approx \frac{u^{\sqrt{2(M+1)}}}{1-u} = \frac{(.95)^{\sqrt{2(1+1)}}}{1-.95} = 18.05 \text{ customers} \quad (1.8)$$

From Little's Law, this in turn implies that the average waiting time before a customer is served briscuit is:

$$W_{q,briscuit} = \frac{L_q}{\lambda} = \frac{18.05 \text{ customers}}{38 \text{ custs/hr}} = 0.475 \text{ hrs or } 28.5 \text{ minutes.} \quad (1.9)$$

Because the arrival and service rates are the same for the ribs line, its average line length and customer waiting time will be the same as for the briscuit line. Both the ribs and briscuit lines' arrival rates are Poisson and their service times are exponential. So Jackson's Law tells us that the outflows from the two servers are also Poisson. Hence, the cashier is fed two Poisson arrival rates of 38 customers/hr, which will imply a single Poisson arrival rate at the cashier line of $(38 \text{ custs/hr} + 38 \text{ custs/hr}) = 76 \text{ custs/hr}$. Utilization at the cashier is $(76 \text{ custs/hr})/(80 \text{ custs/hr}) = 95\%$, the same as at the preceding two lines. Using our formula, this means that the line length is:

$$L_{q,cashier} \approx \frac{u^{\sqrt{2(M+1)}}}{1-u} = \frac{(.95)^{\sqrt{2(1+1)}}}{1-.95} = 18.05 \text{ customers.} \quad (1.10)$$

This is the same as in front of either ribs or briscuit. However, the average waiting time for the cashier is much shorter as we would expect from our previous discussion.

$$W_{q,cashier} = \frac{L_q}{\lambda} = \frac{18.05 \text{ customers}}{76 \text{ custs/hr}} = 0.238 \text{ hrs or } 14.3 \text{ minutes} \quad (1.11)$$

In fact, this is a general insight from queueing theory, that average waiting times for a given number of servers will be less if they are fed by one combined line instead of separate individual lines.

Now, we need to find the flowtime for the entire restaurant. There are two ways to do this. One is to find the individual flowtimes of each phase of the process, multiply them by the probability of a customer entering them, and then sum up the results.

$$\begin{aligned} FT_{rest} &= 50\% (W_{q,briscuit} + CT_{briscuit}) + 50\% (W_{q,ribs} + CT_{ribs}) + 100\% (W_{q,cashier} + CT_{cashier}) \\ &= 50\% \left(W_{q,briscuit} + \frac{1}{\mu_{briscuit}} \right) + 50\% \left(W_{q,ribs} + \frac{1}{\mu_{ribs}} \right) + 100\% \left(W_{q,cashier} + \frac{1}{\mu_{cashier}} \right) \\ &= 50\% \left(0.475 \text{ hrs} + \frac{1}{40 \text{ custs/hr}} \right) + 50\% \left(0.475 \text{ hrs} + \frac{1}{40 \text{ custs/hr}} \right) + 100\% \left(0.238 \text{ hrs} + \frac{1}{80 \text{ custs/hr}} \right) \\ &= 0.75 \text{ hrs} \end{aligned} \quad (1.12)$$

Another possibility is first to add up all the people in the restaurant and then exploit Little's Law. To do this, we need to realize that the average number of people at any server will equal its utilization. (To see this consider that a server with 75% utilization will be serving 1 person 75% of the time and 0 people 25% of the time. Then $75\%(1 \text{ person}) + 25\%(0 \text{ people}) = 0.75 \text{ people}$.) Then adding up the people at each queue and each server:

$$\begin{aligned} L_{rest} &= (L_{q,briscuit} + M_{briscuit} * u_{briscuit}) + (L_{q,ribs} + M_{ribs} * u_{ribs}) + (L_{q,cashier} + M_{cashier} * u_{cashier}) \\ &= (18.05 + 1 * 0.95) + (18.05 + 1 * 0.95) + (18.05 + 1 * 0.95) \\ &= 57 \text{ customers} \end{aligned} \quad (1.13)$$

Now, we can use Little's Law to find the flowtime through the restaurant because the mean arrival rate at the restaurant from Figure 5 is 76 customers/hour.

$$FT_{rest} = W = \frac{L}{\lambda} = \frac{57 \text{ people}}{76 \text{ custs/hr}} = 0.75 \text{ hrs} \quad (1.14)$$

Note that the answers from both approaches are identical.

The Psychology of Waiting

One final note on waiting line management: What we have studied so far is only the “physics” of waiting line management. In reality, there are always two sides to any customer experience: the reality and the perception. For example, in skyscrapers, it has been shown that the presence of mirrors next to elevators will cause people waiting for elevators to consistently reduce their estimate of their waiting time. There are numerous other psychological approaches that have been proposed. The evidence on the effectiveness of most of them is mixed. However, research has clearly shown that customers tolerate more waiting when they perceive a service as being more valuable. Hence, many service firms now provide express lines, such as those seen in grocery stores, for people requiring minimal service. However, there may be other approaches to managing waiting line expectations. One barber shop with a typical waiting time of 45 minutes has changed the focus of their service from just giving a haircut to providing 45 minutes of relaxation from the hustle and bustle of normal life. To this end, they continuously run first-run DVDs on an expensive entertainment system in the middle of the barbershop, provide a large variety of new magazines, and provide a shoulder massage after every haircut. The moral of this story is that when the “physics” of the system leave you with an unsatisfactory wait time for your customers, perhaps you can change the rules of the game by managing their expectations.