CORE CURRICULUM

Operations Management

Roy D. Shapiro, Series Editor

HARVARD BUSINESS PUBLISHING

READING + INTERACTIVE ILLUSTRATIONS

Managing Queues

ELLIOTT N. WEISS
DARDEN SCHOOL OF BUSINESS

8047 | Published: July 28, 2014

Table of Contents

This reading contains links to online interactive exercises, denoted by a 🔍. In order to access these exercises, you will need to have a broadband Internet connection. Verify that your browser meets the minimum technical requirements by visiting http://hbsp.harvard.edu/list/tech-specs.

1 INTRODUCTION

"If this is a service economy, why am I still on hold?" ask Frances Frei and Anne Morriss in their book *Uncommon Service: How to Win by Putting Customers at the Core of Your Business.*[1] This is just one example of waiting that you may have experienced. Have you recently waited in a checkout line at the supermarket? At a "fast"-food restaurant? For a web page to download? All of these are examples of the phenomenon of queuing: the act of waiting for a service to begin. Queuing is an important concern for operating managers in both service and manufacturing environments. They often need to make decisions regarding capacity levels, staffing, and technology that affect the time customers or goods spend waiting for a service.

Waiting lines, or *queues*, most often form because of variability in a system. In most service businesses, as well as some manufacturing settings, the operating manager's job is complicated by an inability to specify either the time of arrival of a customer request or the work content of the service to be provided. In a fast-food business, for instance, customers arrive at unpredictable, random times during the day and expect to be able to order from a wide variety of items. The challenge for the restaurant manager is to fulfill customers' expectations while consuming as few resources as possible. The manager must determine the appropriate staffing level, the proper technology for preparing the food, and the target for the customer's **waiting time**. For example, too few personnel and a grill that takes an excessive amount of time to cook burgers may result in a waiting time that's too long. Too many personnel and a grill that cooks the burgers more quickly (and is more expensive) may provide a short waiting time but decrease profitability.

Queues can also form in a manufacturing environment, where, for example, physical parts may have to wait for a machine to become available. In this reading, we primarily use service examples to explain queuing phenomena. Those interested in manufacturing can easily apply the frameworks and models discussed here.

Waiting lines can result in delays. The consequences of these delays depend on the context. A shorter delay in a hospital emergency room may prevent serious medical complications for patients with heart attacks. Quicker turnaround times for loan processing may increase the number of loan applications and the lender's profitability. Long waits at a restaurant may cause potential customers to go elsewhere for a meal.[a] Shorter delays result, in many cases, in better use of labor, equipment, and facilities, and a greater capability for providing customers with an excellent level of service.

In this reading, we present frameworks and tools that managers can use to understand the characteristics of waiting-line systems and improve operations where queues may form. We first discuss queuing-related questions that managers ask in designing a service operation and the nature of waiting-line systems. We then lay out some basic assumptions about those systems and describe various characteristics of them.

---

[a] We are reminded of the quote by the famous American baseball player and "philosopher" Yogi Berra about a popular restaurant: "Nobody goes there anymore; it's too crowded."

We also present a formula for approximating *line length* in a system, some of the insights into queuing systems that arise from this formula, and some further applications of the formula. In the Supplemental Reading, we address ways to manage customers' perceptions of waiting.

2 ESSENTIAL READING

2.1 Managerial Objectives and the Nature of Waiting Lines

The objective for managers in applying the calculations in this reading is to understand and estimate the various costs and performance characteristics of waiting-line systems in specific service processes so that they can make appropriate design choices. Costs might be related to the number and type of servers; performance characteristics include considerations such as customer waiting time and line length.

Thus, in designing a service, managers might ask the following queue-related questions:

- On average, how many customers will be waiting for service? How long, on average, will a customer wait?

- How many servers are necessary to achieve a specified limit to customer waiting time?

- What number of servers (in a particular situation) leads to an average waiting time for the customers and an average *idle time* for the servers that minimize the total cost of the service system?

- Should servers be combined into one or more centralized areas?

- Which process improvements should be implemented given different service rates, operating costs, and target cost and service objectives?

- What is the impact of a reduction in the variability of either the time between customer arrivals or the service times?

The answers clearly depend on the relative costs (in a particular situation) of customers and servers. As the above questions imply, a waiting-line system in a service setting has three basic components:

1. Servers

2. Customers, who may arrive at random or unscheduled times

3. Service encounters, which typically vary in their completion times

At times, a customer's waiting time and a server's idle time incur easily measurable out-of-pocket costs. For example, every minute a prospective customer waits on the telephone for a customer service representative costs the organization long-distance telephone charges. But at other times, there are no clear costs associated with a waiting customer. Customers queued in restaurants, post offices, banks, airports, and so forth, do not impose direct costs to the service provider, but if the lines become too long, some customers may leave before completing the transaction or choose a competitor the next time they need that service. Thus, an operations manager must plan for waiting lines and make staffing decisions accordingly—all the while striving to manage the ongoing trade-off between excessive service capacity and needless waiting.

It is important for managers to understand that a line can form even if the capacity of the server(s) exceeds customer demand. Waiting time cannot be entirely eliminated

because of the variability we mentioned earlier. A large number of customers may arrive over a short period of time, or there might be a few closely spaced service encounters that take an excessive amount of time. Conversely, even if few servers are present and waiting lines prevail, the servers may be idle some of the time if there is a long gap between arrivals or there are some successive short service encounters.

Interactive Illustration 1 shows a graphical representation of a queuing system. There is one server, and service time is always exactly 4 minutes per customer. Therefore, during a 25-minute interval, the server could serve 5 customers and have 5 minutes of idle time. Likewise, it is possible for all 5 customers to be served without having to wait for service. However, that situation occurs only if customer arrivals are evenly spaced during the 25-minute interval.

To run the simulation, click "Start" and then click "Customer Arrival" four times, once for each customer. When the 25 simulated minutes end, you can see data on the length of queues that form the waiting time per customer. Try different arrival patterns to see the effect on performance metrics.

Interactive Illustration 1 Cumulative Arrivals and Departures
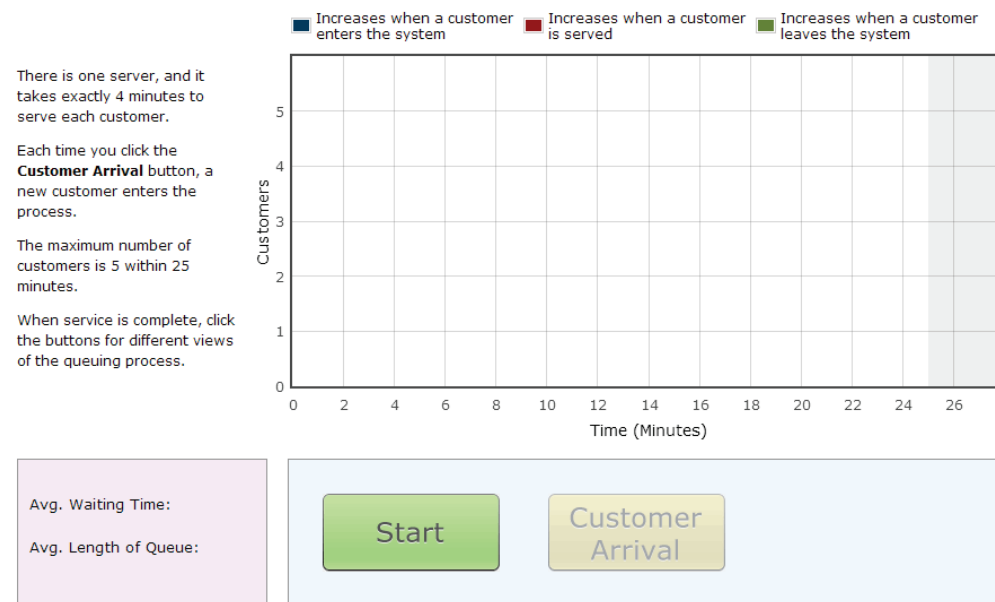
2.2 Basic Assumptions, Symbols, and Definitions

The calculations discussed in this reading depend on the following assumptions about waiting-line situations:

1 **First come, first served.** Customers form single queues while waiting for service and are taken for service on a first-come, first-served basis. Customers move into a server channel (for instance, a teller's window at a bank) as it becomes available.

2   **Servers perform identical services.** Servers can also be referred to as *channels* performing services; thus, the number of servers equals the number of channels. There may be multiple servers operating in parallel to service one line.

3   **Random arrivals and service.** Customer arrivals occur randomly at an average rate, called the *arrival rate*. Services are performed at an average rate, called the *service rate*, at each of the servers. Note that the units of average arrival rate and average service rate are customers per unit time. The reciprocal value of the service rate, *service time*, is the average time for serving a customer and is measured in time units. Similarly, the reciprocal value of the average arrival rate is the average time between arrivals, called the *inter-arrival time*, and is also measured in time units.

4   **Statistical equilibrium (steady state).** On average, the capability to process customers must be greater than the rate at which they arrive; otherwise, arriving customers might have to wait in line[b] indefinitely. The average capability of the system to process customers per unit time is given by the average total service rate, which equals (number of servers) · (service rate per server). The average number of customers requiring service per unit time is given by the arrival rate. In transient situations, such as system startup—for example, when an amusement park first opens in the morning—the number in the queue and in service depends on the number of customers who waited for the system to go into operation and on how long the system has been in operation. Over time, the impact of the initial conditions tends to dampen out, and the number of customers in the system and in line is independent of time. The calculations described in this reading are not appropriate for transient situations.

5   **Infinite source of potential customers.** We assume an infinite source of potential customers. We may also use these calculations when the population of the queuing system is not greater than 1% of the population of the source of potential customers.

6   **Appropriate queue behavior.** We assume the arrival rate equals the demand rate. This assumption implies no instances of:

   a.   balking (potential recipients refusing to join the queue)

   b.   reneging (customers in the queue leaving before being served)

   c.   cycling (recipients of service returning to the queue following service)

7   **Uniform service effectiveness.** The effectiveness of service at each server channel is uniform over time and throughout the system.

---

[b] The use of *in line* and *on line* to describe a person in a queue varies regionally in the United States. The author is from Philadelphia, so he has chosen to use *in line*.

The calculations in this reading also use the symbols and definitions outlined in **Table 1**.

Table 1 Definitions of Queuing Parameters

|	
Facility	A total servicing unit; a facility may be composed of several channels (as defined below) or as few as one
Channel	A servicing point within a facility
A	Arrival rate; average number of arrivals into the system per unit of time
S	Service rate (per server); average number of services per unit of time per channel
T_s	Average service time ($1/S$)
m	Number of servers or number of channels in the facility
mS	Average service rate of the facility ($m \cdot$ service rate per server)
u	A/mS = (arrival rate)/($m \cdot$ service rate per server) = utilization or utilization factor
N_s	Average number of customers in service
L	Average number of customers in the system; those being served plus the waiting line
L_q	Average number of customers waiting in line; this number may be approximated using a formula provided later in this reading
W	Total average time in the system
W_q	Average time spent waiting

2.3 Performance Characteristics of Waiting-Line Systems

Now that we have established the assumptions and definitions provided in the preceding section, we can describe the performance characteristics of waiting-line systems (utilization factor, average number of customers in service, average number of customers waiting in the queue, average number of customers in the system, expected waiting time in the queue, expected total average time in the system, and expected costs) and explore the relationships among them.

Utilization factor (u): The **utilization factor** measures the percentage of time that servers are busy with customers.

$$u = \frac{A}{mS}$$

Note that u is a percentage and is a unit-less number. Consider a system with an arrival rate of 10 people per hour, where each server can service 11 per hour. For $m = 1$ server, the utilization is $u = 10/11 = 91\%$. This indicates that the server is busy 91% of the time. As you can see in **Figure 1**, the utilization factor decreases as the number of servers increases.

Figure 1 Utilization Factor
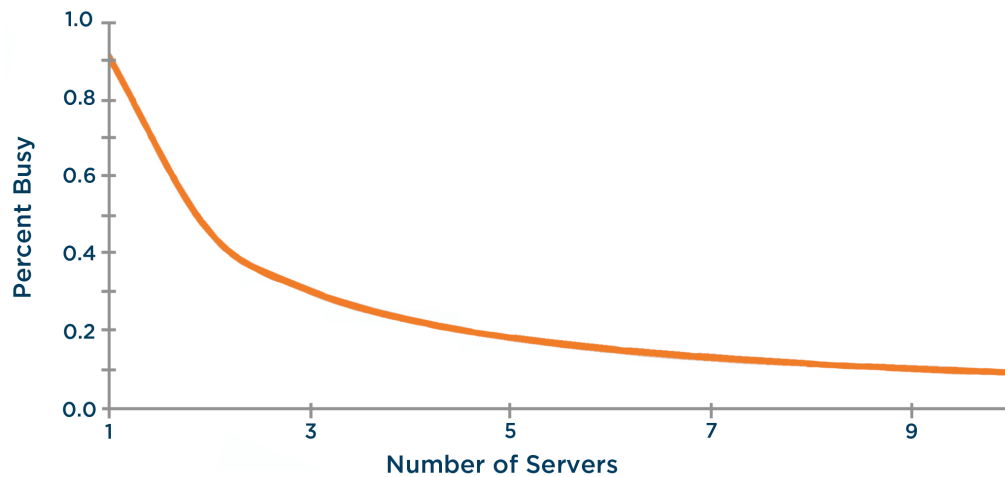
Average number of customers in service (N_s): The average number of customers in service is A/S, derived by multiplying the utilization factor (u) times the number of servers (m). (This does not include the number of customers waiting in line.)

$$N_s = m \cdot u = m \cdot \left(\frac{A}{mS} \right) = \frac{A}{S}$$

N_s will vary as a direction function of the utilization, which is a function of the arrival and service rates of the system, as shown above.

Average number of customers waiting in the queue (L_q): The average number of customers waiting in the queue excludes the customers that are being served. For most systems, L_q cannot be calculated directly because of the complex interactions caused by the probability distributions describing the arrival and service processes; its value must be approximated. An approximation formula (or **queuing approximation**) for this performance characteristic is presented in the next section. We will show that the queue length is a function of the number of servers, the utilization factor, and the arrival and service variability.

The remaining performance characteristics of waiting-line systems all depend on the value of L_q—either observed empirically or estimated from the approximation formula described in the next section.

Average number of customers in the system (L): By definition, the average number of customers in the system is equal to the average number of customers in line plus the average number of customers in service. That is:

$$L = L_q + N_s$$

Expected waiting time in the queue[c] (W_q): The expected waiting time can be determined by dividing the average queue length by the average arrival rate:
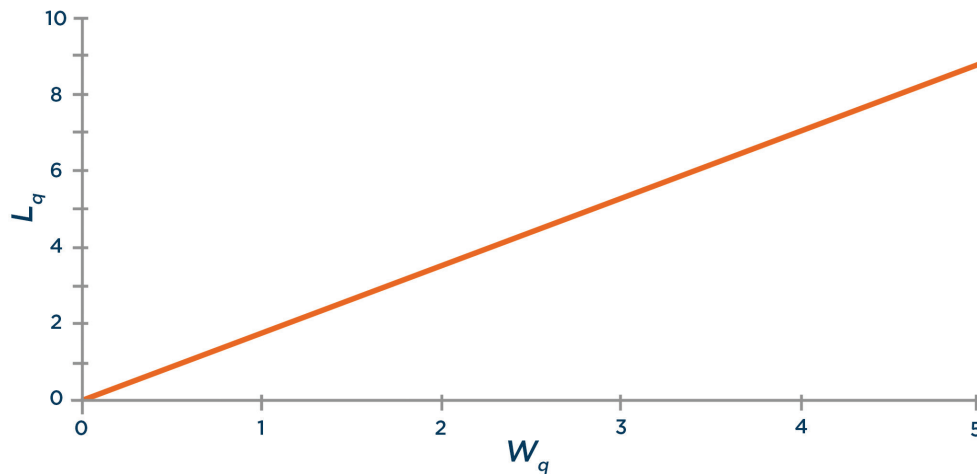
$$W_q = \frac{L_q}{A}$$

---

[c] There is often a large difference between the actual waiting time and the perceived waiting time. Methods for managing customers' perceptions are discussed in the Supplemental Reading.

The formula that relates the waiting time and line length is commonly referred to as *Little's Law.* This relationship, $L_q = A \cdot W_q$ for the queue and $L = A \cdot W$ for the entire system, was initially proven by John D. C. Little.[2] The formula states that the length of a line is directly related to the time spent in the line (or that the number of customers or items in the system is directly related to the total time spent in the system). Given any two of the parameters, the third can be determined. For example, if we know that the average length of the line at a doughnut store is 3.5 customers, and we know that the average time spent in line is two minutes, we can infer that the arrival rate to the store is given by:

$$A = \frac{L_q}{W_q} = \frac{3.5}{2} = 1.75 \text{ customers/minute} = 105 \text{ customers/hour}$$

Figure 2 shows the impact of changes in the queue length on the average waiting time. Obviously, the longer the line, the longer the wait. The slope of the line is the arrival rate. The slope is *not* the service rate because Little's Law deals with average wait time and queue lengths at steady state. The departure rate from a steady state system must equal the arrival rate. Thus, the relationship between line length and queue time is a function of the average rate at which the line decreases. This average rate must weight busy and nonbusy times.

Figure 2 Little's Law

Expected total average time in the system (W): Because the average service time per person or item is equal to the quantity $1/S$, the total average time expected in the system is:

$$W = W_q + \left(\frac{1}{S}\right)$$

The above relationships among W_q, L_q, A, and S are *not* dependent on any assumptions regarding the distribution of arrival and service times but instead are true in general.

Expected costs: The expected costs of a queuing system to the organization are usually determined by some function of the number of servers and the expected customers' waiting time or the number of customers in the system. The cost of waiting

can be either an out-of-pocket cost or an opportunity cost. When the cost of waiting is known, the cost of providing service may be added to the cost of waiting to obtain a total system cost. For example, consider a queuing system with three channels and with a cost per channel of $25 per hour. The total channel cost is $75 per hour. If the average waiting time per customer is 15 minutes (0.25 hours), and the opportunity cost of waiting to the firm is $15 per hour per customer, each customer arrival would cost (0.25 hours) · $15 = $3.75 per customer. If 10 customers arrived per hour, the cost of waiting would be $37.50 per hour. Added to the $75 cost of service, the total cost would be $112.50 per hour.

Adding an extra server would cost $25 per hour. The server would be added if the total system cost dropped to less than $112.50 per hour. Given that the cost of servers is now 4 · $25 = $100, four servers would be optimal if the total opportunity cost of waiting is now less than $12.50 per hour because the new total system cost is $100 + the cost of waiting. This implies that the average waiting time for a customer would need to be less than 0.0833 hours ($12.50/[$15 waiting cost per hour per customer · 10 customers per hour]). Note that at 0.0833 hours per customer, the waiting cost per customer would be (0.0833 hours) · $15 = $1.25 per customer. Multiplying by the 10 customers per hour yields a total cost of waiting of $12.50 per hour.

Often, the waiting cost for customers is not known. Although there may be an opportunity cost to the individuals or items in line, there may be no direct cost to the firm. In these cases, managers may set targets for the average waiting time or for probabilities of waiting no more than a specified period of time. They may also consider some trade-offs. Consider the previous example. If we did not know or could not estimate the $15/hour waiting time cost, we might ask, "Would you spend $25/hour to reduce the average waiting time from 0.25 hours to x hours?" (where x is the estimated average waiting time for customers in the new system). Note that rather than have an explicit numerical answer to this question, we must rely on managerial judgment. One could also factor in the lost revenue per customer and the probability of losing customers if the lines are excessively long.

We have ignored the cost of the customer(s) in service (that is, customers who are being served and so are not waiting in a queue) in our calculations. This may be included but will not affect how many servers that managers decide to include in the system because the time in service is not a function of the number of servers in these examples. (Recall that $N_s = A/S$, and neither A nor S is a function of the number of servers.) If managers had to decide the actual service rate at each channel, it would be necessary to include the time spent in service as part of the total system cost.

2.4 An Approximation Formula for L_q

In this section, we present an approximation formula for calculating the value of L_q. Recall that once we have this value for the line length, other important performance characteristics, such as W_q, W, and L, can be determined. The *relationships* among W_q, L_q, W, L, A, and S, described in the preceding section, are not dependent on the nature or shape of the probability distributions describing the arrival and service processes. The *values* of the measures W_q, L_q, W, and L depend highly on the assumptions regarding these distributions, as we will see.

The L_q approximation formula below can be used for approximating the line length of a queuing system. It is often referred to as the Sakasegawa approximation.[3] Although many approximation formulas are available for queuing systems, we present this

particular formula for its relative ease of use, its use for developing insight regarding queuing systems, and the relative quality of its approximation to "true" values for systems with significant waiting lines.

$$L_q \approx \frac{u^{\sqrt{2(m+1)}}}{(1-u)} \cdot \frac{\left(\left(CV_{IAT}\right)^2 + \left(CV_{ST}\right)^2\right)}{2}$$

u = the utilization factor, as defined above

m = the number of servers

CV_{IAT} = the coefficient of variation of the inter-arrival times

CV_{ST} = the coefficient of variation of the service time

The coefficient of variation of a probability distribution provides a relative measure of its variability and is defined as the standard deviation of a distribution divided by its mean. It is used in recognition of the fact that the magnitude of the standard deviation of a distribution is more relevant with respect to the size of its mean than as an absolute measure. Thus, a service time distribution with a standard deviation of 10 minutes is much more significant if the mean is 10 minutes than if the mean is 50 minutes.[d]

Whereas the general principles of queue management—such as Little's Law, the effects of increasing variability, and the nature of pooling or the combination of queues (which we'll discuss later)—are applicable in a wide variety of situations, the L_q approximation formula is more limited and should be used only in situations with first come, first served arrivals; a single type of customer; no limit on the waiting room; and stationary, steady state conditions. For systems with finite waiting areas, priority queues, different customer types, and overflow queues, more-sophisticated models and simulation tools are available. For example, the Erlang loss formula can be used for systems where customers are turned away when the system is full.[4] Many Excel add-ins that handle specific situations can be downloaded from the Internet (see, for example, Queuing ToolPak from the University of Alberta School of Business). The Extend simulation package is also often used to model complex queuing situations.

In addition, the L_q approximation is useful for calculating the *average* line length and waiting time. In many instances, we are interested in the probability that line lengths or wait times will exceed a certain threshold. For example, a call center's goal may be to have an average hold time of 30 seconds, with no more than a 5% chance of waiting 45 seconds or longer. In this situation, more-detailed results on the shape of the waiting time distribution would be needed to determine staffing requirements.

---

[d] In general, distributions with a $CV < 0.75$ are considered low variability, $0.75 < CV < 1.33$ are considered moderate variability, and $CV > 1.33$ are considered high variability. The actual values of CV can be measured from empirical data, as in the previous example. Often, for convenience and with no other information, we assume that $CV = 1.0$. If the probability distribution describing the number of arrivals in a given period is Poisson, a common occurrence, then $CV_{IAT} = 1.0$; if the probability distribution describing the service time is exponential, a common assumption, then $CV_{IAT} = 1.0$.

2.5   Some Insights into Waiting-Line Systems

At first glance, the L_q formula appears intimidating; however, its form provides some general insights into queuing systems. Here we discuss these and illustrate them using the L_q formula.

Increasing Utilization Degrades Performance at an Increasing Rate

Looking at the L_q approximation formula, we see the first term has the factor $(1 - u)$ in the denominator. As the system gets busier, the term $(1 - u)$ (the percentage of idle time) gets smaller; therefore, the line length grows at an increasing rate. Indeed, as u approaches 1 and $(1 - u)$ approaches 0, the line length grows without bounds.

To illustrate the effects of changes in utilization in a waiting-line system, consider a facility with one channel, an average arrival rate of 15 customers per hour, and an average service rate of 20 customers per hour. The average service time, T_s = 1/service rate = 1/20 of an hour (3 minutes). Thus, arrival rate = 15/hour; m = number of service channels = 1; and service rate = 20/hour. Assume that the coefficient of variation of both the service time distribution and the arrival time distribution is 1.0.

- The utilization ratio, u = arrival rate / ($m \cdot$ service rate) = $A/(m \cdot S)$, is thus [(15 / (1 · 20)] or 0.75.

- The average number of customers in service is given by $N_s = A/S = 15/20 = 0.75$.

- The average number waiting (L_q), excluding those being served, is 2.25 using the formula above.

- The average number of customers in the system (L) is 3.00 ($L_q + N_s$). The average waiting time (W_q) is 0.15 hours, or 9 minutes using Little's Law, $W_q = L_q/A$.

- The average system time (W) is 0.2 hours, or 12 minutes ($W_q + [1/S]$).

Excel data tables can be used to measure the sensitivity of the performance measures to the input parameters. **Interactive Illustration 2** can also be used to test the sensitivity to the arrival rate and utilization. Move the slider for the arrival rate to see how the average time in the queue changes. W_q grows exponentially as the arrival rate increases and the utilization, u, approaches the value 1. Line lengths grow so drastically because the impact of variability is so much greater at higher utilization rates. A longer service time or arrivals that occur closer together intuitively have a larger effect on the average line length when systems are more crowded. The marginal benefit of each additional server decreases as servers are added.
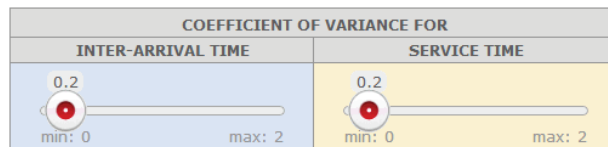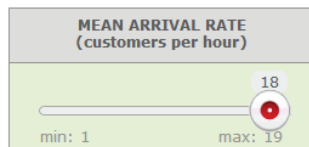
Interactive Illustration 2 Average Time in Queue
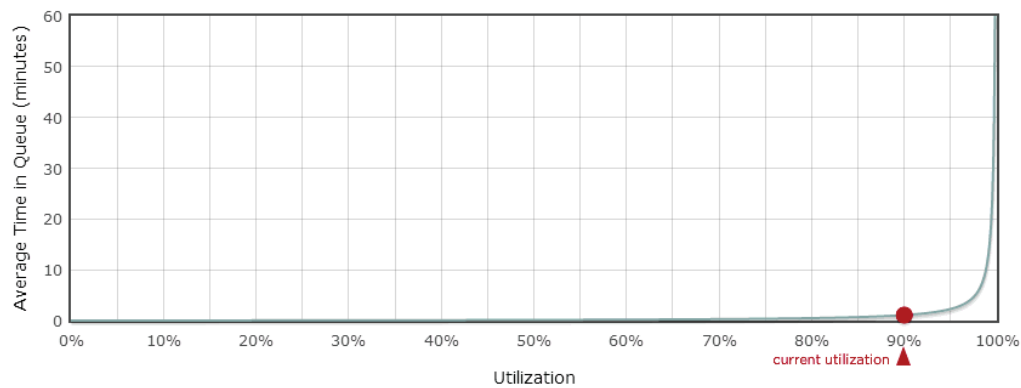
| An employee can serve, on average, 20 customers per hour, so the mean service time is 3 minutes. | Standard deviation of inter-arrival time | 0.67 | Standard deviation of service time | 0.60 |
| | Mean inter-arrival time | 3.33 min | Average time in queue | 1.08 min |

MEAN ARRIVAL RATE (customers per hour)	COEFFICIENT OF VARIANCE FOR	
	INTER-ARRIVAL TIME	SERVICE TIME
18	0.2	0.2
min: 1 max: 19	min: 0 max: 2	min: 0 max: 2

Increasing Variability Degrades Performance at an Increasing Rate

Looking at the second term of the L_q formula, we see that when the coefficients of variation of both the service rate and the arrival rate are equal to 1, the second term has no effect (because the second term is equal to 1 in that case). This is a common situation because the exponential probability distribution accurately describes the empirically observed inter-arrival and service times of many queuing systems. The exponential distribution has a CV of 1.

The line length goes up with relative variability in either the inter-arrival time or the service time. Indeed, as we've said, one of the causes of lines in systems is the inherent variability of the service time and the inter-arrival time. When we observe the second term, systems with lower coefficients of variation have shorter queues, all other things being equal, as the second term becomes less than 1; when the coefficients of variation exceed 1, the queue is longer, all other things being equal. In both cases, the magnitude of the change is affected by the squaring factor. So a 25% reduction in the CV reduces the queue by less than 25%, but a 25% increase in the CV increases the queue by more than 25%.

To illustrate the effects of changes in variability, look again at **Interactive Illustration 2** and see what happens as you change the coefficients of variation of the inter-arrival time and the service time. For any given utilization factor, as determined by the mean arrival rate and depicted on the graph by the large red dot, the average time in queue (and by Little's Law, the length of the queue) increases with an increase in the system variability, as measured by the coefficients of variation. The chart also shows the values for the standard deviations of the inter-arrival times and service times as points of reference.
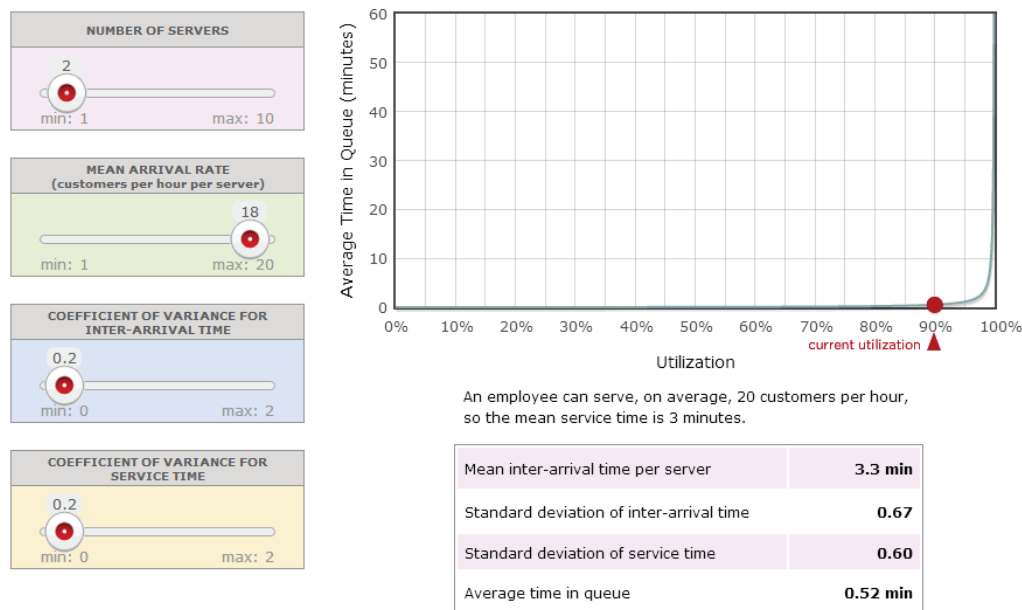
All Other Things Being Equal, Bigger Is Better

For the same utilization and variability values, a bigger system (that is, one with more servers) is better able to handle the variability. We know this intuitively because a larger system can "pool" the variability across multiple servers and can better "absorb" periods with a large number of arrivals in a short time. Thus, in larger systems, a long service time at one or more of the channels does not have as large an impact because the other channels can serve the additional customers. This effect is apparent from the formula by again inspecting the first term. Notice that m, the number of channels, enters in the exponent of u, which is a value less than 1, so as m gets larger, the first term gets smaller (proportional to the square root of m), and the line length is smaller for larger systems with the same utilization. Note the implications of this: If two smaller systems can be pooled (combined) into one larger system, the average line length decreases while maintaining the same utilization. Alternatively, the pooled system can be run at a higher utilization than the individual smaller systems while maintaining the same average line length as the smaller system.

    Interactive Illustration 3 demonstrates the effects of changes in system size. The graphic depicts a system with a service rate of 20 customers per hour and allows you to change the number of servers, the mean arrival rate, and the coefficients of variation for the inter-arrival time and the service time. Note the effect of changing the number of servers. For any given utilization, the average length of the queue is smaller for a larger numbers of servers. As the number of servers increases, the queue time graph moves down and to the right. This will be investigated in more detail in an example below.

Interactive Illustration 3 Average Time in Queue with Multiple Servers

NUMBER OF SERVERS
2
min: 1   max: 10

MEAN ARRIVAL RATE (customers per hour per server)
18
min: 1   max: 20

COEFFICIENT OF VARIANCE FOR INTER-ARRIVAL TIME
0.2
min: 0   max: 2

COEFFICIENT OF VARIANCE FOR SERVICE TIME
0.2
min: 0   max: 2

An employee can serve, on average, 20 customers per hour, so the mean service time is 3 minutes.

Mean inter-arrival time per server	3.3 min
Standard deviation of inter-arrival time	0.67
Standard deviation of service time	0.60
Average time in queue	0.52 min

2.6 Sample Applications of the L_q Approximation Formula

The previous section described insights that can be developed from the L_q formula regarding utilization, variability, and pooling. In this section we provide examples of how the formula can be used with sample data, and how it can be used to evaluate a consolidation decision at a bank and a staffing decision for a telephone call center.

Using the Formula with Sample Data

The key to applying the L_q approximation formula is understanding the difference between the probability distribution that describes the *number* of customers arriving in a time period (arrival rate) and the probability distribution of the time *between* customers arriving (inter-arrival time). Calculating utilization requires information regarding the average arrival rate and the average service rate, measured in customers per time period. The coefficients of variation in the L_q formula are determined by looking at the inter-arrival time and the actual service times. While the arrival rate can be determined from the inter-arrival times, it is not necessarily true that the inter-arrival distribution can be derived from the arrival rates. Likewise, if the service time distribution is known, then the service rate and the coefficient of variation of the service times can be calculated. Knowing the number of people served in a given time period does not necessarily yield the service time distribution.

Table 2 Sample Call Center Data

Customer Number	Arrival Time	Time from Previous Customer (minutes)	Service Time (minutes)
1	8:03		4
2	8:06	3	3
3	8:09	3	1
4	8:12	3	3
5	8:16	4	3
6	8:20	4	2
7	8:23	3	4
8	8:25	2	4
9	8:28	3	1
10	8:32	4	1
11	8:36	4	4
12	8:41	5	4
13	8:45	4	3
14	8:49	4	1
15	8:53	4	3
16	8:55	2	4
17	8:58	3	3
18	9:00	2	2

Customer Number	Arrival Time	Time from Previous Customer (minutes)	Service Time (minutes)
19	9:04	4	4
20	9:09	5	3
21	9:13	4	3
22	9:17	4	4
23	9:20	3	3
24	9:24	4	2
25	9:26	2	3
26	9:30	4	3
27	9:35	5	3
28	9:38	3	3
29	9:43	5	2
30	9:46	3	4
31	9:48	2	3
32	9:52	4	4
33	9:54	2	3
34	9:58	4	3
35	10:00	2	1
Mean		**3.43**	**2.89**
Standard Deviation		**0.95**	**0.99**

Consider the call center data presented in **Table 2,** which shows the arrival and service times for the customers who called between 8 a.m. and 10 a.m. one morning. Thirty-five customers arrived in the two-hour interval. The mean time between arrivals was 3.43 minutes with a standard deviation of 0.95. Because 35 customers arrived in a two-hour interval, the arrival rate is 35/2 = 17.5 customers per hour. The average service time was 2.89 minutes, so the service rate is 60/2.89 = 20.76 per hour.

- If there is one server, the utilization is given by 17.5/20.76 = 0.84, or 84%.

- The coefficient of variation of the inter-arrival times is given by $CV_{IAT} = 0.95/3.43 = 0.28$.

Because the mean service time was 2.89 minutes with a standard deviation of 0.99 minutes, the coefficient of variation of the service time is given by

$$CV_{ST} = \frac{0.99}{2.89} = 0.43$$

The average line length can be predicted to be equal to 0.58 customers.

Effects of Pooling

In the previous section, we claimed that bigger is better. If so, a system that is a combination of a number of smaller queues should have smaller line lengths and waiting times than the individual queues themselves. The following example illustrates this phenomenon, known as pooling.

Consider a bank that is evaluating a decision to consolidate three regional loan application centers into a single centralized center. The data for the three regions and the proposed centralized center are shown in **Table 3.**

Table 3 Data for Three Regions of a Bank

	Region A	Region B	Region C	Entire System
Arrival rate (A) (customers per hour)	0.38	0.32	0.36	
Service rate (S) (customers per hour)	0.20	0.20	0.20	
Processors (m)	2.00	2.00	2.00	
Utilization (A/mS)	0.95	0.80	0.90	
CV_{IAT}	1.00	1.00	1.00	
CV_{ST}	1.00	1.00	1.00	
L_q (customers) (estimated)	17.60	2.90	7.70	28.2 (sum of three regions)
W_q (hours)	46.40	9.00	21.50	26.7 (weighted average of three regions)

For each region the average time in the queue was calculated using the L_q approximation. Little's Law was used to calculate the average waiting time. Note that the system averages 28.2 customers waiting, with an average time of 26.7 hours. What happens if we consolidate the regions? Consolidation yields an arrival rate of 1.06 per hour (the sum of the three individual regions). The utilization becomes $u = (A/mS) = 1.06/(6 \cdot 0.2) = 0.8833$. The L_q formula results in a line length of 5.4. $W_q = L_q/1.06 = 5.08$ hours. Note the tremendous effect of combining the regions. In the original system, there were 28.2 (17.6 + 2.9 + 7.7) people waiting at the three regions. Thus, combining the three regions results in an 81% reduction in the number of loans awaiting service ([28.2 – 5.4]/28.2 = 0.81)! This occurs for two reasons: (1) There is no longer a region with a high utilization (Region A) and (2) there is no longer the possibility that one region has a number of loans awaiting processing while another region is idle.

Call Center Staffing

A telephone call center for a mail-order catalog has the demand pattern shown in **Table 4** for weekdays between 8:00 a.m. and noon.

Table 4 Customer Demand at Call Center

Beginning Time	Average Number of Customers
8:00 a.m.	75
9:00 a.m.	110
10:00 a.m.	135
11:00 a.m.	185

The average time per customer call is 5 minutes, with a standard deviation of 6 minutes. The coefficient of variation for the inter-arrival times has historically been 1. The call center manager would like to know the number of staff required in each 1-hour block in order to have an average waiting time of under 1 minute.

We can use the L_q formula to answer this question. First, note that rather than use the average arrival rate for the entire four-hour period, we should determine a staffing level for each of the four-hour periods; that is, we will solve four separate problems. Aggregating the data would result in overestimating the staffing level for periods with below-average arrivals and underestimating the staffing level for periods with above-average arrivals.

We will use a trial-and-error approach with the L_q formula in order to determine the appropriate staffing levels for the target average waiting times. A will be given by the values, in customers per hour, in **Table 4**. Because the average service time is 5 minutes, the service rate per staff person is 60/5 = 12 customers per hour. $CV_{IAT} = 1$, as given, and $CV_{ST} = 6/5 = 1.2$. The results are shown in **Table 5.**

Table 5 Analysis of Call Center Data

Time Period	m	u	L_q (Customers)	W_q (Hours)	W_q (Minutes)
8:00 a.m.	9	0.69	0.78	0.010	0.625
9:00 a.m.	12	0.76	1.31	0.012	0.714
10:00 a.m.	14	0.80	1.87	0.014	0.833
11:00 a.m.	19	0.81	1.73	0.009	0.559

The m values were determined by increasing m until the average waiting times were under 1 minute. By looking at the utilization factors, we can see that bigger is indeed better because the larger systems can run at higher utilizations for the same target value of 1 minute average waiting time. The manager can perform sensitivity analyses of the number of servers (and utilization) versus the target average waiting time to see if the target is viable.

3 SUPPLEMENTAL READING

Perceived Versus Actual Waiting Time

One of the classic references on waiting is David Maister's 1985 "The Psychology of Waiting Lines," which made the intuitive but, at the time, novel argument that a customer's perception of service quality is influenced just as much by the subjective experience of waiting in a queue as it is by the objective measures of the waiting experience (such as the number of minutes spent in line). Maister formulated the First and Second Laws of Service—that is, companies can influence customer satisfaction in a waiting line by working on what the customer expects and what the customer perceives, especially in the early parts of the service encounter. Maister also identified eight psychological factors that increase a customer's negative perception of a wait, making it feel longer:[5]

- Unoccupied time feels longer than occupied time (distraction).[6]

- Pre-process waits feel longer than in-process waits (moment).

- Anxiety makes waits seem longer (anxiety).

- Uncertain waits are longer than certain waits (uncertainty).

- Unexplained waits are longer than explained waits (explanation).

- Unfair waits are longer than equitable waits (fairness).

- The more valuable the service, the longer people will wait (value).

- Solo waiting feels longer than waiting in a group (solo wait).

Maister proposed that companies remedy these factors by instituting the elements of customer service that most customers take for granted today, including updating them about their status in the queue, giving them a sense of control, providing value-added (or distracting) activities to occupy waiting time (such as perusing menus at a restaurant), promoting a sense of fairness (by, for instance, using a ticket system to determine order priority), and setting expectations. Especially if customers must go through a series of waits, companies should acknowledge that they have been "entered into the system" (through a registration procedure, such as one finds at a walk-in clinic), even though they may have another period of waiting before the service can be performed (the medical consultation). Companies also need to address even "irrational" sources of customer anxiety: If I switch lines, will the next one move faster? Are there enough seats on the plane for all ticketed passengers?

Maister's arguments were theoretical, but in 1984, the first empirical field study on waiting (in the retail industry) had verified a link between the conditions in which consumers wait and their subjective perception of the wait. For example, the study showed that people overestimated waiting time by 36% on average (a five-minute wait feels seven minutes long).[7] In the early 1990s, a survey of hotel and restaurant customers conducted by a group of United Kingdom Forte hotel managers independently confirmed many of Maister's proposed factors.[8] More than 70% of respondents were concerned about waiting times. However, the survey revealed a nuance: Although the customers believed that quality and value were worth waiting for, at a certain point a wait would become unacceptable and would lower their perception of quality. This was especially true for the hotel customers surveyed. In their operations, the hotel managers identified

six key points at which customers have to wait during an overnight stay (check in, luggage, telephone line, messages, room service, and checkout) and nine points during a restaurant meal. All of these provide opportunities to establish operational standards and to manage customers' perceptions.

A 1999 literature review of 18 empirical studies of wait management supported Maister's basic conceptual framework but also identified a hierarchy of factors influencing consumers' behavior.[9] The review's authors ranked their importance as temporal factors (real time/duration waited) first, individual factors (disposition of the customer) second, and situational factors (controlled by the company) third. The studies emphasized the importance of personal expectations of waits, although the authors of the 18 articles used different measures of the expectation concept, ranging from "probable duration" to "reasonable duration" to "maximum tolerable duration." Although there are many individual factors that companies can't control, they should try to identify customers whose preferences they can accommodate (such as by providing various checkout options at supermarkets) and those who worry more about waits. The review's only real revision to Maister's framework was to reclassify "anxiety" as a dependent variable, not a causal factor.

Different companies take different approaches to sharing information with customers. Some give actual expected waiting times: Disney does so for its amusement park rides, which helps families plan their day. Other companies "hide" visual information by, for instance, wrapping a line around the corner. Overestimating is also common; for example, a restaurant may tell customers they'll have a ten-minute wait but then seat them after eight minutes so that they feel gratified with the "fast service." Depending on the configuration of their operations, managers need to decide the best queue discipline. Will they process requests by arrival, according to priority (such as hospital triage), or on an appointment basis? Scheduling appointments involves a fine balance between leaving service providers unproductive (if appointments are scheduled too far apart) and not meeting consumers' high expectations (if appointments are scheduled so close together that the provider runs late).[10]

It's important not only to provide waiting customers with status information but also to show them work in progress as they wait for their service to be performed. In fact, according to recent research, even the "appearance of effort" improves customer satisfaction—customers who have to wait but receive visual cues (as the customers ahead of them are served) may in fact be happier than customers who experience no wait in the absence of visual cues.[11] Examples of this labor illusion, as the authors call it, are showing the names of airlines searched on the Kayak travel website or steaming the milk for each individual coffee order at Starbucks.

In their investigation of the impact of culture on queuing behavior, Graham Gillam, Kyle Simmons, and Elliott Weiss note that culture may affect how an individual perceives a queue and thus can affect his or her service experience.[12] Drawing on other research demonstrating that social justice, or a sense of fairness, often informs a customer's attitude toward waiting in a particular line,[13] they observe that this sense of justice varies with culture. In some countries, it is common for people of higher status to be ushered to the front of lines and be served immediately. This is viewed as "fair" in locations less concerned about equality among people. Malcolm Gladwell observes:[14]

> In cultures that aren't obsessed with punctuality or "wasted" time, chaotic lines for services are considered less of a problem. When Robert Levine, a psychologist at California State University at Fresno, studied the notoriously nonqueuing Brazilians, he found they had far fewer

clocks and watches per capita than similarly developed societies, and those they had were less likely to be accurate.

The British Broadcasting Corporation (BBC) also sheds light on some cultural attitudes toward queuing: [15]

- In India, although first come, first served lines are common in airports, those who jockey for position are often served first in railway and bus stations.

- Although Russians usually form orderly lines, exceptions may occur at doctor's offices when people ask for "a minute with a doctor, just to get his signature." One minute often turns into a half hour.

Gillam, Simmons, and Weiss conclude that companies trying to differentiate their brands in international markets must consider the challenges and opportunities presented by local queuing preferences. In particular, they should

- understand what is important to customers' satisfaction with queues

- determine the steps that optimize the experience for customers in a cost-effective manner

- research what industry peers are doing and how customers respond to their queues

- continually adapt the queue management system on the basis of past experiences and customers' evolving needs

4 KEY TERMS

Idle Time: The amount of time a server is inactive while waiting for customers to arrive.

Line Length: The number of customers or items waiting in line for a service to begin.

Little's Law: A formula that measures the relationship among the line length, the arrival rate, and the waiting time.

Queuing Approximation: A formula that estimates the average number of customers waiting in line. The formula is a function of the utilization, the number of servers, and the variability in arrivals and service as measured by the coefficient of variation of the customer times and the coefficient of variation of the service times.

Utilization Factor: The percentage of time that a server is busy with customers or items in service.

Waiting Time: The amount of time a customer or an item spends in line before a service begins.

5   ENDNOTES

[1] Frances Frei and Anne Morriss, *Uncommon Service: How to Win by Putting Customers at the Core of Your Business* (Boston: Harvard Business Review Press, 2012), pp. 1–12.

[2] John D. C. Little, "A Proof for the Queuing Formula: L = λ W," *Operations Research* 9, no. 3 (May–June 1961): 383–387.

[3] Hirotaka Sakasegawa. "An Approximation Formula for L_q Annals of the Institute of Statistical Mathematics," Volume 29, (1977): part A, pp. 67–75.

[4] Donald Gross et al., *Fundamentals of Queueing Theory*, (New York: Wiley-Interscience, 2008).

[5] David Maister, "Psychology of Waiting Lines," in *The Service Encounter: Managing Employee/Customer Interaction in Service Businesses*, ed. John A, Czepiel, Michael R. Solomon, and Carol F. Surprenant, pp. 113–115 (Lexington, MA: D.C. Heath/Lexington Books, 1985).

[6] Parenthetical terms are borrowed from Agnès Durrande-Moreau, "Waiting for Service: Ten Years of Empirical Research," *Journal of Service Management* 10, no. 2 (1999): 171–194.

[7] Jacob Hornik, "Subjective vs. Objective Time Measures: A Note on the Perception of Time in Consumer Behavior," *Journal of Consumer Research* 11, no. 1 (June 1984*)*: 615–618, as cited in Durrande-Moreau, "Waiting for Service: Ten Years of Empirical Research."

[8] Peter Jones and Michael Dent, "Improving Service: Managing Response Time in Hospitality Operations," *International Journal of Operations & Production Management* 14.5 (1994): 52.

[9] Agnes Durrande-Moreau, "Waiting for Service: Ten Years of Empirical Research," *Journal of Service Management* (1999): 171–194.

[10] David H. Maister, "Note on the Management of Queues," HBS No. 680-053 (Boston: Harvard Business School Publishing, 1979), http://hbsp.harvard.edu, accessed May 2013.

[11] Ryan W. Buell and Michael I. Norton, "Think Customers Hate Waiting? Not So Fast…," *Harvard Business Review,* May 2011: 2, http://hbsp.harvard.edu, accessed May 2013.

[12] Graham Gillam, Kyle Simmons, and Elliott Weiss, "Line, Line Everywhere a Line, the Impact of Culture on Waiting Line Management," (working paper, Darden School of Business, University of Virginia, 2013).

[13] Richard C. Larson, "Perspectives on Queues: Social Justics and the Psychology of Queueing," *Operations Research* 35, no. 6 (Nov/Dec 1987): 895–905, ABI/INFORM via ProQuest, accessed May 2013.

[14] Malcolm Gladwell, "You Are How You Wait—Queues Have Subtle Rules of Fairness and Justice," *The Seattle Times*, December 28, 1992, http://community.seattletimes.nwsource.com/archive/?date=19921228&slug=1532358

[15] Benjamin Walker, "Priority Queues: Paying to Get to the Front of the Line," *BBC News Magazine*, October 10, 2012, http://www.bbc.co.uk/news/magazine-19712847, accessed May 2013.

6 INDEX

Page numbers followed by *f* refer to figures. Page numbers followed by *i* refer to interactive illustrations. Page numbers followed by *t* refer to tables.