Chapter 7

Variability and Its Impact on Process Performance: Waiting Time Problems

For consumers, one of the most visible—and probably annoying—forms of supply–demand mismatches is waiting time. As consumers, we seem to spend a significant portion of our life waiting in line, be it in physical lines (supermarkets, check-in at airports) or in "virtual" lines (listening to music in a call center, waiting for a response e-mail).

It is important to distinguish between different types of waiting time:

- Waiting time predictably occurs when the expected demand rate exceeds the expected supply rate for some limited period of time. This happens especially in cases of constant capacity levels and demand that exhibits seasonality. This leads to implied utilization levels of over 100 percent for some time period. Queues forming at the gate of an airport after the flight is announced are an example of such queues.
- As we will see in the next section, in the presence of variability, queues also can arise if the implied utilization is below 100 percent. Such queues can thereby be fully attributed to the presence of variability, as there exists, on average, enough capacity to meet demand.

While the difference between these two types of waiting time probably does not matter much to the customer, it is of great importance from the perspective of operations management. The root cause for the first type of waiting time is a capacity problem; variability is only a secondary effect. Thus, when analyzing this type of a problem, we first should use the tools outlined in Chapters 3, 4, and 6 instead of focusing on variability.

The root cause of the second type of waiting time is variability. This makes waiting time unpredictable, both from the perspective of the customer as well as from the perspective of the operation. Sometimes, it is the customer (demand) waiting for service (supply) and, sometimes, it is the other way around. Demand just never seems to match supply in these settings.

Analyzing waiting times and linking these waiting times to variability require the introduction of new analytical tools, which we present in this chapter. We will discuss the tools for analyzing waiting times based on the example of An-ser Services, a call-center operation in

Wisconsin that specializes in providing answering services for financial services, insurance companies, and medical practices. Specifically, the objective of this chapter is to

- Predict waiting times and derive some performance metrics capturing the service quality provided to the customer.
- Recommend ways of reducing waiting time by choosing appropriate capacity levels, redesigning the service system, and outlining opportunities to reduce variability.

7.1 Motivating Example: A Somewhat Unrealistic Call Center

For illustrative purposes, consider a call center with just one employee from 7 a.m. to 8 a.m. Based on prior observations, the call-center management estimates that, on average, a call takes 4 minutes to complete (e.g., giving someone driving directions) and there are, on average, 12 calls arriving in a 60-minute period, that is, on average, one call every 5 minutes.

What will be the average waiting time for a customer before talking to a customer service representative? From a somewhat naïve perspective, there should be no waiting time at all. Since the call center has a capacity of serving 60/4 = 15 calls per hour and calls arrive at a rate of 12 calls per hour, supply of capacity clearly exceeds demand. If anything, there seems to be excess service capacity in the call center since its utilization, which we defined previously (Chapter 3) as the ratio between flow rate and capacity, can be computed as

$$\text{Utilization} = \frac{\text{Flow rate}}{\text{Capacity}} = \frac{\text{12 calls per hour}}{\text{15 calls per hour}} = 80\%$$

First, consider the arrivals and service times as depicted in Figure 7.1. A call arrives exactly every 5 minutes and then takes exactly 4 minutes to be served. This is probably the weirdest call center that you have ever seen! No need to worry, we will return to “real operations” momentarily, but the following thought experiment will help you grasp how variability can lead to waiting time.

Despite its almost robotlike service times and the apparently very disciplined customer service representative (“sorry, 4 minutes are over; thanks for your call”), this call center has one major advantage: no incoming call ever has to wait.

FIGURE 7.1
A Somewhat Odd Service Process

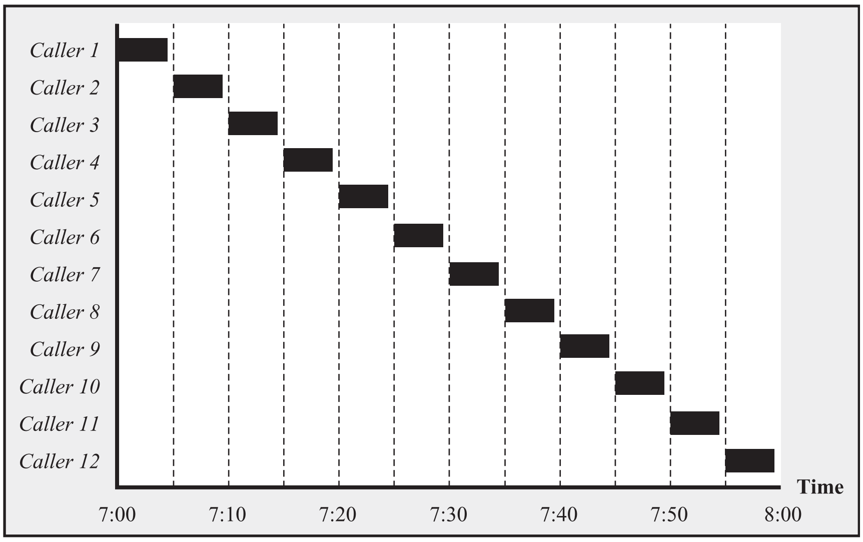
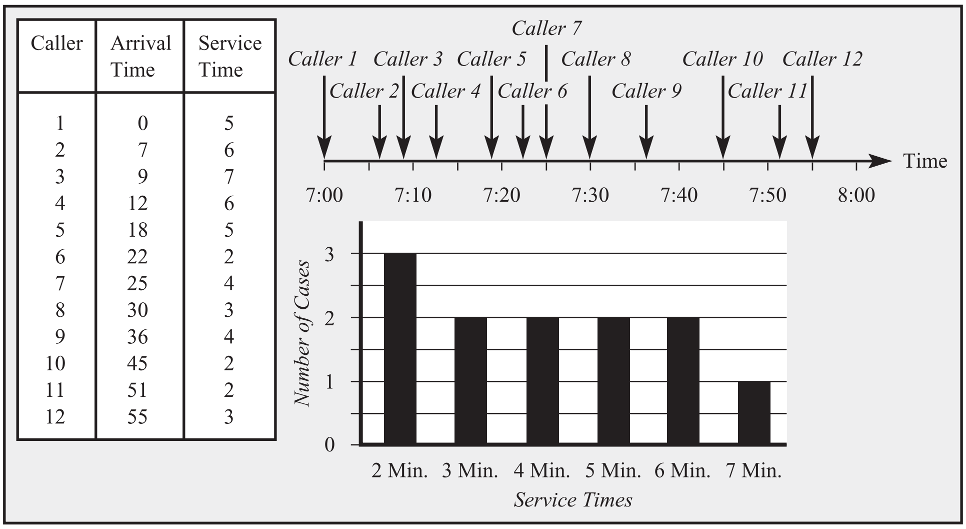


FIGURE 7.2
Data Gathered at a Call Center



Caller	Arrival Time	Service Time
1	0	5
2	7	6
3	9	7
4	12	6
5	18	5
6	22	2
7	25	4
8	30	3
9	36	4
10	45	2
11	51	2
12	55	3

Assuming that calls arrive like kick scooters at an assembly line and are then treated by customer service representatives that act like robots reflects a common mistake managers make when calculating process performance. These calculations look at the process at an aggregate level and consider how much capacity is available over the entire hour (day, month, quarter), yet ignore how the requests for service are spaced out within the hour.

If we look at the call center on a minute-by-minute basis, a different picture emerges. Specifically, we observe that calls do not arrive like kick scooters appear at the end of the assembly line, but instead follow a much less systematic pattern, which is illustrated by Figure 7.2.

Moreover, a minute-by-minute analysis also reveals that the actual service durations also vary across calls. As Figure 7.2 shows, while the average service time is 4 minutes, there exist large variations across calls, and the actual activity times range from 2 minutes to 7 minutes.

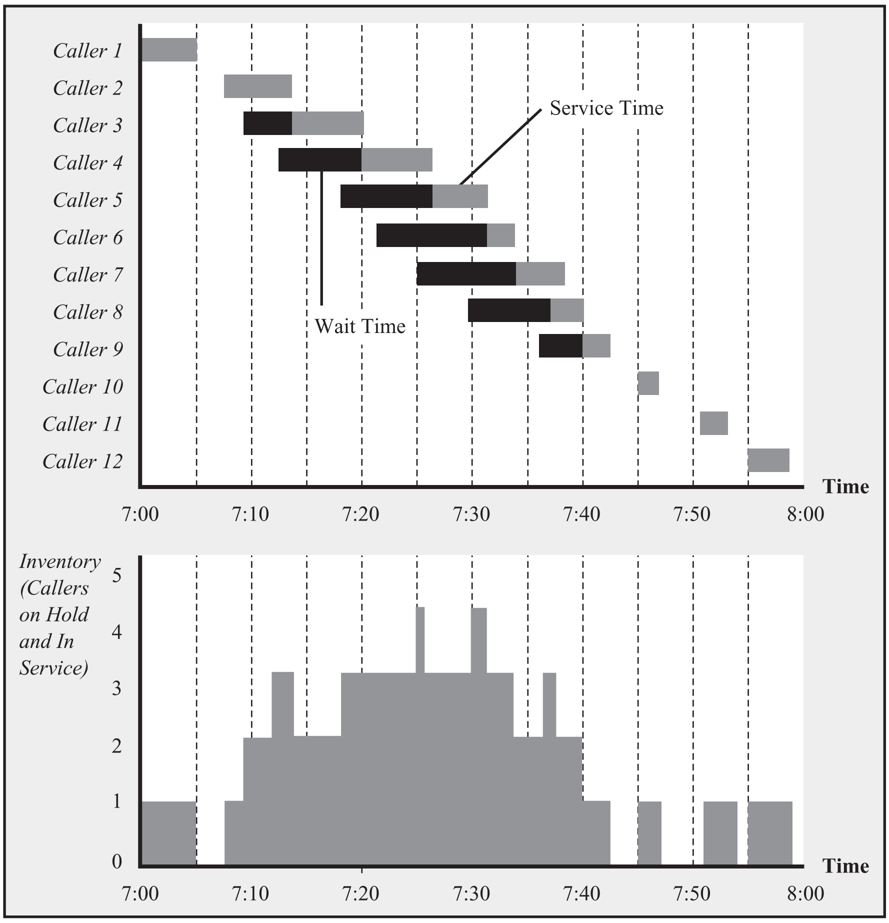
Now, consider how the hour from 7:00 a.m. to 8:00 a.m. unfolds. As can be seen in Figure 7.2, the first call comes in at 7:00 a.m. This call will be served without waiting time, and it takes the customer service representative 5 minutes to complete the call. The following 2 minutes are idle time from the perspective of the call center (7:05–7:07). At 7:07, the second call comes in, requiring a 6-minute service time. Again, the second caller does not have to wait and will leave the system at 7:13. However, while the second caller is being served, at 7:09 the third caller arrives and now needs to wait until 7:13 before beginning the service.

Figure 7.3 shows the waiting time and service time for each of the 12 customers calling between 7:00 a.m. and 8:00 a.m. Specifically, we observe that

- Most customers do have to wait a considerable amount of time (up to 10 minutes) before being served. This waiting occurs, although, on average, there is plenty of capacity in the call center.
- The call center is not able to provide a consistent service quality, as some customers are waiting, while others are not.
- Despite long waiting times and—because of Little's Law—long queues (see lower part of Figure 7.3), the customer service representative incurs idle time repeatedly over the time period from 7 a.m. to 8 a.m.

Why does variability not average out over time? The reason for this is as follows. In the call center example, the customer service representative can only serve a customer if there is capacity *and* demand at the same moment in time. Therefore, capacity can never "run

FIGURE 7.3
Detailed Analysis of Call Center



ahead" of demand. However, demand can "run ahead" of capacity, in which case the queue builds up. The idea that inventory can be used to decouple the supply process from demand, thereby restoring the flow rate to the level achievable in the absence of variability, is another version of the "buffer or suffer" principle that we already encountered in Chapter 6. Thus, if a service organization attempts to achieve the flow-rate levels feasible based on averages, long waiting times will result (unfortunately, in those cases, it is the customer who gets "buffered" and "suffers").

Taking the perspective of a manager attempting to match supply and demand, our objectives have not changed. We are still interested in calculating the three fundamental performance measures of an operation: inventory, flow rate, and flow time. Yet, as the above example illustrated, we realize that the process analysis tools we have discussed up to this point in the book need to be extended to appropriately deal with variability.

7.2 Variability: Where It Comes From and How It Can Be Measured

As a first step toward restoring our ability to understand a process's basic performance measures in the presence of variability, we take a more detailed look at the concept of variability itself. Specifically, we are interested in the sources of variability and how to measure variability.

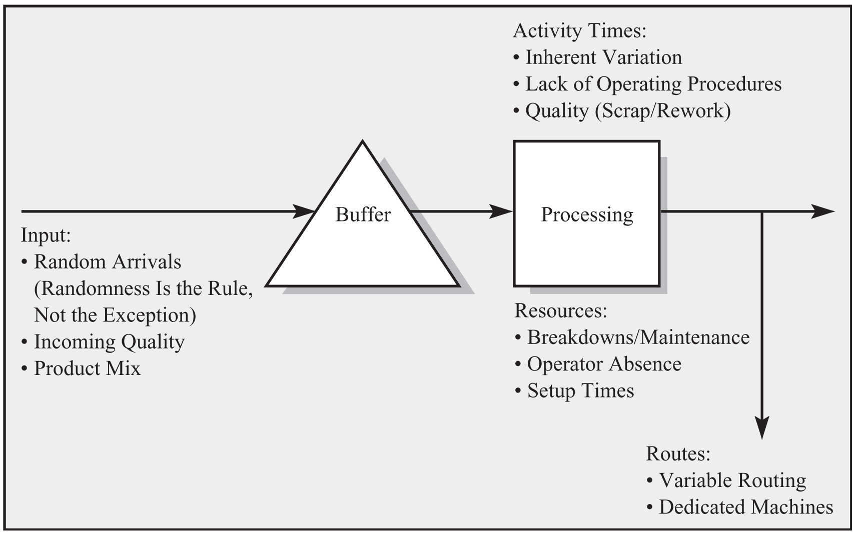


FIGURE 7.4
Variability and Where It Comes From

Why is there variability in a process to begin with? Drawing a simple (the most simple) process flow diagram suggests the following four sources of variability (these four sources are summarized in Figure 7.4):

- Variability from the inflow of flow units. The biggest source of variability in service organizations comes from the market itself. While some patterns of the customer-arrival process are predictable (e.g., in a hotel there are more guests checking out between 8 and 9 a.m. than between 2 and 3 p.m.), there always remains uncertainty about when the next customer will arrive.
- Variability in activity times. Whenever we are dealing with human operators at a resource, it is likely that there will be some variability in their behavior. Thus, if we would ask a worker at an assembly line to repeat a certain activity 100 times, we would probably find that some of these activities were carried out faster than others. Another source of variability in activity times that is specific to a service environment is that in most service operations, the customer him/herself is involved in many of the tasks constituting the activity time. At a hotel front desk, some guests might require extra time (e.g., the guest requires an explanation for items appearing on his or her bill), while others check out faster (e.g., simply use the credit card that they used for the reservation and only return their room key).
- Random availability of resources. If resources are subject to random breakdowns, for example, machine failures in manufacturing environments or operator absenteeism in service operations, variability is created.
- Random routing in case of multiple flow units in the process. If the path a flow unit takes through the process is itself random, the arrival process at each individual resource is subject to variability. Consider, for example, an emergency room in a hospital. Following the initial screening at the admissions step, incoming patients are routed to different resources. A nurse might handle easy cases, more complex cases might be handled by a general doctor, and severe cases are brought to specific units in the hospital (e.g., trauma center). Even if arrival times and service times are deterministic, this random routing alone is sufficient to introduce variability.

In general, any form of variability is measured based on the standard deviation. In our case of the call center, we could measure the variability of call durations based on collecting some data and then computing the corresponding standard deviation. The problem with this approach is that the standard deviation provides an *absolute* measure of variability. Does a standard deviation of 5 minutes indicate a high variability? A 5-minute standard deviation for call durations (activity times) in the context of a call center seems like a large number. In the context of a 2-hour surgery in a trauma center, a 5-minute standard deviation seems small.

For this reason, it is more appropriate to measure variability in *relative* terms. Specifically, we define the *coefficient of variation* of a random variable as

$$\text{Coefficient of variation} = \text{CV} = \frac{\text{Standard deviation}}{\text{Mean}}$$

As both the standard deviation and the mean have the same measurement units, the coefficient of variation is a unitless measure.

7.3 Analyzing an Arrival Process

Any process analysis we perform is only as good as the information we feed into our analysis. For this reason, Sections 7.3 and 7.4 focus on data collection and data analysis for the upcoming mathematical models. As a manager intending to apply some of the following tools, this data analysis is essential. However, as a student with only a couple of hours left to the final exam, you might be better off jumping straight to Section 7.5.

Of particular importance when dealing with variability problems is an accurate representation of the demand, which determines the timing of customer arrivals.

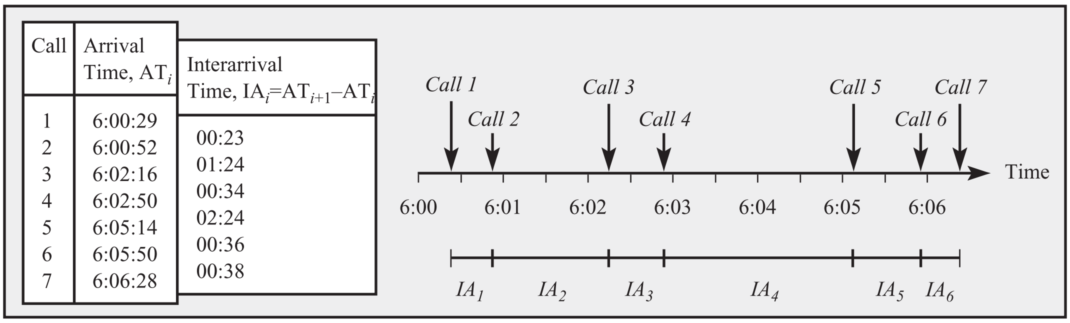
Assume we got up early and visited the call center of An-ser; say we arrived at their offices at 6:00 a.m. and we took detailed notes of what takes place over the coming hour. We would hardly have had the time to settle down when the first call comes in. One of the An-ser staff takes the call immediately. Twenty-three seconds later, the second call comes in; another 1:24 minutes later, the third call; and so on.

We define the time at which An-ser receives a call as the *arrival time.* Let AT_i denote the arrival time of the *i*th call. Moreover, we define the time between two consecutive arrivals as the *interarrival time,* IA. Thus, $\text{IA}_i = \text{AT}_{i+1} - \text{AT}_i$. Figure 7.5 illustrates these two definitions.

If we continue this data collection, we accumulate a fair number of arrival times. Such data are automatically recorded in call centers, so we could simply download a file that looks like Table 7.1.

Before we can move forward and introduce a mathematical model that predicts the effects of variability, we have to invest in some simple, yet important, data analysis. A major risk

FIGURE 7.5 The Concept of Interarrival Times



Call	Arrival Time, AT_i	Interarrival Time, $\text{IA}_i = \text{AT}_{i+1} - \text{AT}_i$
1	6:00:29	00:23
2	6:00:52	01:24
3	6:02:16	00:34
4	6:02:50	02:24
5	6:05:14	00:36
6	6:05:50	00:38
7	6:06:28	

TABLE 7.1 **Call Arrivals at An-ser on April 2, 2002, from 6:00 a.m. to 10:00 a.m.**

6:00:29	6:52:39	7:17:57	7:33:51	7:56:16	8:17:33	8:28:11	8:39:25	8:55:56	9:21:58
6:00:52	6:53:06	7:18:10	7:34:05	7:56:24	8:17:42	8:28:12	8:39:47	8:56:17	9:22:02
6:02:16	6:53:07	7:18:17	7:34:19	7:56:24	8:17:50	8:28:13	8:39:51	8:57:42	9:22:02
6:02:50	6:53:24	7:18:38	7:34:51	7:57:39	8:17:52	8:28:17	8:40:02	8:58:45	9:22:30
6:05:14	6:53:25	7:18:54	7:35:10	7:57:51	8:17:54	8:28:43	8:40:09	8:58:49	9:23:13
6:05:50	6:54:18	7:19:04	7:35:13	7:57:55	8:18:03	8:28:59	8:40:23	8:58:49	9:23:29
6:06:28	6:54:24	7:19:40	7:35:21	7:58:26	8:18:12	8:29:06	8:40:34	8:59:32	9:23:45
6:07:37	6:54:36	7:19:41	7:35:44	7:58:41	8:18:21	8:29:34	8:40:35	8:59:38	9:24:10
6:08:05	6:55:06	7:20:10	7:35:59	7:59:12	8:18:23	8:29:38	8:40:46	8:59:45	9:24:30
6:10:16	6:55:19	7:20:11	7:36:37	7:59:20	8:18:34	8:29:40	8:40:51	9:00:14	9:24:42
6:12:13	6:55:31	7:20:26	7:36:45	7:59:22	8:18:46	8:29:45	8:40:58	9:00:52	9:25:07
6:12:48	6:57:25	7:20:27	7:37:07	7:59:22	8:18:53	8:29:46	8:41:12	9:00:53	9:25:15
6:14:04	6:57:38	7:20:38	7:37:14	7:59:36	8:18:54	8:29:47	8:41:26	9:01:09	9:26:03
6:14:16	6:57:44	7:20:52	7:38:01	7:59:50	8:18:58	8:29:47	8:41:32	9:01:31	9:26:04
6:14:28	6:58:16	7:20:59	7:38:03	7:59:54	8:19:20	8:29:54	8:41:49	9:01:55	9:26:23
6:17:51	6:58:34	7:21:11	7:38:05	8:01:22	8:19:25	8:30:00	8:42:23	9:02:25	9:26:34
6:18:19	6:59:41	7:21:14	7:38:18	8:01:42	8:19:28	8:30:01	8:42:51	9:02:30	9:27:02
6:19:11	7:00:50	7:21:46	7:39:00	8:01:56	8:20:09	8:30:08	8:42:53	9:02:38	9:27:04
6:20:48	7:00:54	7:21:56	7:39:17	8:02:08	8:20:23	8:30:23	8:43:24	9:02:51	9:27:27
6:23:33	7:01:08	7:21:58	7:39:35	8:02:26	8:20:27	8:30:23	8:43:28	9:03:29	9:28:25
6:24:25	7:01:31	7:23:03	7:40:06	8:02:29	8:20:44	8:30:31	8:43:47	9:03:33	9:28:37
6:25:08	7:01:39	7:23:16	7:40:23	8:02:39	8:20:54	8:31:02	8:44:23	9:03:38	9:29:09
6:25:19	7:01:56	7:23:19	7:41:34	8:02:47	8:21:12	8:31:11	8:44:49	9:03:51	9:29:15
6:25:27	7:04:52	7:23:48	7:42:20	8:02:52	8:21:12	8:31:19	8:45:05	9:04:11	9:29:52
6:25:38	7:04:54	7:24:01	7:42:33	8:03:06	8:21:25	8:31:20	8:45:10	9:04:33	9:30:47
6:25:48	7:05:37	7:24:09	7:42:51	8:03:58	8:21:28	8:31:22	8:45:28	9:04:42	9:30:58
6:26:05	7:05:39	7:24:45	7:42:57	8:04:07	8:21:43	8:31:23	8:45:31	9:04:44	9:30:59
6:26:59	7:05:42	7:24:56	7:43:23	8:04:27	8:21:44	8:31:27	8:45:32	9:04:44	9:31:03
6:27:37	7:06:37	7:25:01	7:43:34	8:05:53	8:21:53	8:31:45	8:45:39	9:05:22	9:31:55
6:27:46	7:06:46	7:25:03	7:43:43	8:05:54	8:22:19	8:32:05	8:46:24	9:06:01	9:33:08
6:29:32	7:07:11	7:25:18	7:43:44	8:06:43	8:22:44	8:32:13	8:46:27	9:06:12	9:33:45
6:29:52	7:07:24	7:25:39	7:43:57	8:06:47	8:23:00	8:32:19	8:46:40	9:06:14	9:34:07
6:30:26	7:07:46	7:25:40	7:43:57	8:07:07	8:23:02	8:32:59	8:46:41	9:06:41	9:35:15
6:30:32	7:09:17	7:25:46	7:45:07	8:07:43	8:23:12	8:33:02	8:47:00	9:06:44	9:35:40
6:30:41	7:09:34	7:25:48	7:45:32	8:08:28	8:23:30	8:33:27	8:47:04	9:06:48	9:36:17
6:30:53	7:09:38	7:26:30	7:46:22	8:08:31	8:24:04	8:33:30	8:47:06	9:06:55	9:36:37
6:30:56	7:09:53	7:26:38	7:46:38	8:09:05	8:24:17	8:33:40	8:47:15	9:06:59	9:37:23
6:31:04	7:09:59	7:26:49	7:46:48	8:09:15	8:24:19	8:33:47	8:47:27	9:08:03	9:37:37
6:31:45	7:10:29	7:27:30	7:47:00	8:09:48	8:24:26	8:34:19	8:47:40	9:08:33	9:37:38
6:33:49	7:10:37	7:27:36	7:47:15	8:09:57	8:24:39	8:34:20	8:47:46	9:09:32	9:37:42
6:34:03	7:10:54	7:27:50	7:47:53	8:10:39	8:24:48	8:35:01	8:47:53	9:10:32	9:39:03
6:34:15	7:11:07	7:27:50	7:48:01	8:11:16	8:25:03	8:35:07	8:48:27	9:10:46	9:39:10
6:36:07	7:11:30	7:27:56	7:48:14	8:11:30	8:25:04	8:35:25	8:48:48	9:10:53	9:41:37
6:36:12	7:12:02	7:28:01	7:48:14	8:11:38	8:25:07	8:35:29	8:49:14	9:11:32	9:42:58
6:37:21	7:12:08	7:28:17	7:48:50	8:11:49	8:25:16	8:36:13	8:49:19	9:11:37	9:43:27
6:37:23	7:12:18	7:28:25	7:49:00	8:12:00	8:25:22	8:36:14	8:49:20	9:11:50	9:43:37
6:37:57	7:12:18	7:28:26	7:49:04	8:12:07	8:25:31	8:36:23	8:49:40	9:12:02	9:44:09
6:38:20	7:12:26	7:28:47	7:49:48	8:12:17	8:25:32	8:36:23	8:50:19	9:13:19	9:44:21
6:40:06	7:13:16	7:28:54	7:49:50	8:12:40	8:25:32	8:36:29	8:50:38	9:14:00	9:44:32
6:40:11	7:13:21	7:29:09	7:49:59	8:12:41	8:25:45	8:36:35	8:52:11	9:14:04	9:44:37
6:40:59	7:13:22	7:29:27	7:50:13	8:12:42	8:25:48	8:36:37	8:52:29	9:14:07	9:44:44
6:42:17	7:14:04	7:30:02	7:50:27	8:12:47	8:25:49	8:37:05	8:52:40	9:15:15	9:45:10
6:43:01	7:14:07	7:30:07	7:51:07	8:13:40	8:26:01	8:37:11	8:52:41	9:15:26	9:46:15
6:43:05	7:14:49	7:30:13	7:51:31	8:13:41	8:26:04	8:37:12	8:52:43	9:15:27	9:46:44
6:43:57	7:15:19	7:30:50	7:51:40	8:13:52	8:26:11	8:37:35	8:53:03	9:15:36	9:49:48
6:44:02	7:15:38	7:30:55	7:52:05	8:14:04	8:26:15	8:37:44	8:53:08	9:15:40	9:50:19
6:45:04	7:15:41	7:31:24	7:52:25	8:14:41	8:26:28	8:38:01	8:53:19	9:15:40	9:52:53
6:46:13	7:15:57	7:31:35	7:52:32	8:15:15	8:26:28	8:38:02	8:53:30	9:15:40	9:53:13
6:47:01	7:16:28	7:31:41	7:53:10	8:15:25	8:26:37	8:38:10	8:53:32	9:15:41	9:53:15
6:47:10	7:16:36	7:31:45	7:53:18	8:15:39	8:26:58	8:38:15	8:53:44	9:15:46	9:53:50
6:47:35	7:16:40	7:31:46	7:53:19	8:15:48	8:27:07	8:38:39	8:54:25	9:16:12	9:54:24
6:49:23	7:16:45	7:32:13	7:53:51	8:16:09	8:27:09	8:38:40	8:54:28	9:16:34	9:54:48
6:50:54	7:16:50	7:32:16	7:53:52	8:16:10	8:27:17	8:38:44	8:54:49	9:18:02	9:54:51
6:51:04	7:17:08	7:32:16	7:54:04	8:16:18	8:27:26	8:38:49	8:55:05	9:18:06	9:56:40
6:51:17	7:17:09	7:32:34	7:54:16	8:16:26	8:27:29	8:38:57	8:55:05	9:20:19	9:58:25
6:51:48	7:17:09	7:32:34	7:54:26	8:16:39	8:27:35	8:39:07	8:55:14	9:20:42	9:59:19
6:52:17	7:17:19	7:32:57	7:54:51	8:17:16	8:27:54	8:39:20	8:55:22	9:20:44	
6:52:17	7:17:22	7:33:13	7:55:13	8:17:24	8:27:57	8:39:20	8:55:25	9:20:54	
6:52:31	7:17:22	7:33:36	7:55:35	8:17:28	8:27:59	8:39:21	8:55:50	9:21:55	

related to any mathematical model or computer simulation is that these tools always provide us with a number (or a set of numbers), independent of the accuracy with which the inputs we enter into the equation reflect the real world.

Answering the following two questions before proceeding to any other computations improves the predictions of our models substantially.

- Is the arrival process *stationary;* that is, is the expected number of customers arriving in a certain time interval constant over the period we are interested in?
- Are the interarrival times *exponentially distributed,* and therefore form a so-called *Poisson* arrival process?

We now define the concepts of stationary arrivals and exponentially distributed interarrival times. We also describe how these two questions can be answered, both in general as well as in the specific setting of the call center described previously. We also discuss the importance of these two questions and their impact on the calculations in this and the next chapter.

Stationary Arrivals

Consider the call arrival pattern displayed in Table 7.1. How tempting it is to put these data into a spreadsheet, compute the mean and the standard deviation of the interarrival times over that time period, and end the analysis of the arrival pattern at this point, assuming that the mean and the standard deviation capture the entire behavior of the arrival process. Five minutes with Excel, and we could be done!

However, a simple graphical analysis (Figure 7.6) of the data reveals that there is more going on in the arrival process than two numbers can capture. As we can see graphically in Figure 7.6, the average number of customers calling within a certain time interval (e.g., 15 minutes) is not constant over the day.

To capture such changes in arrival processes, we introduce the following definitions:

- An arrival process is said to be *stationary* if, for any time interval (e.g., an hour), the expected number of arrivals in this time interval only depends on the length of the time interval, not on the starting time of the interval (i.e., we can move a time interval of a fixed length forth and back on a time line without changing the expected number of arrivals). In the context of Figure 7.6, we see that the arrival process is not stationary. For example, if we take a 3-hour interval, we see that there are many more customers arriving from 6 to 9 a.m. than there are from 1 to 4 a.m.
- An arrival process exhibits *seasonality* if it is not stationary.

FIGURE 7.6
Seasonality over the Course of a Day

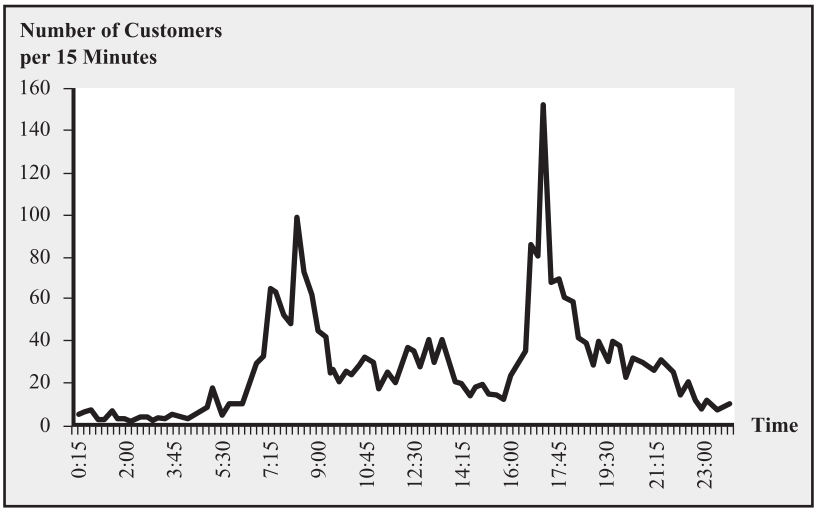
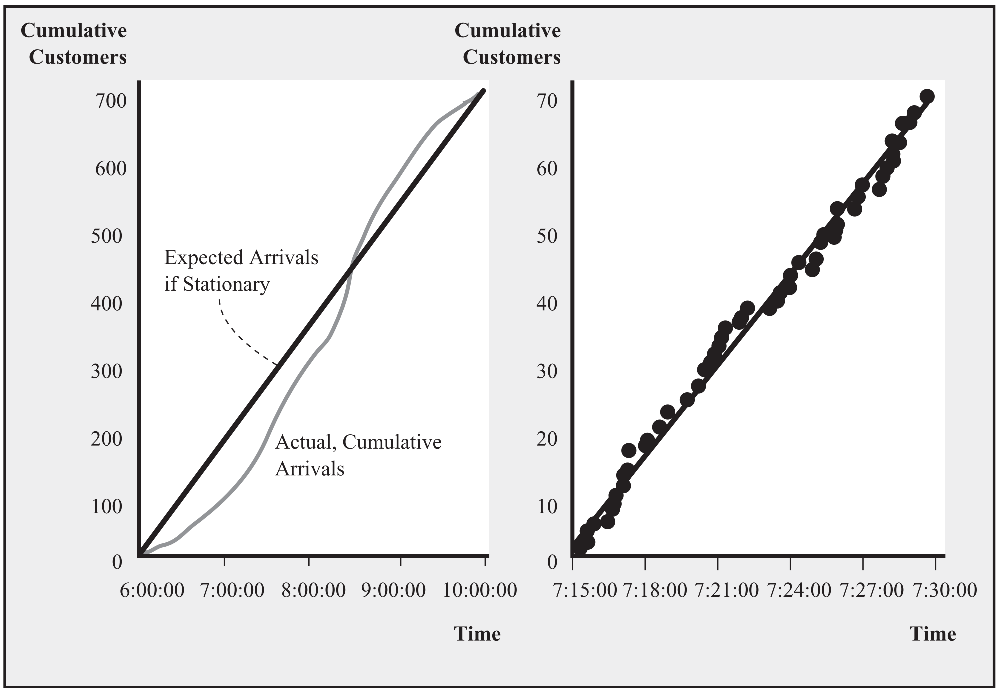


FIGURE 7.7 **Test for Stationary Arrivals**



When analyzing an arrival process, it is important that we distinguish between changes in demand (e.g., the number of calls in 15 minutes) that are a result of variability and changes in demand that are a result of seasonality. Both variability and seasonality are unpleasant from an operations perspecitive. However, the effect of seasonality alone can be perfectly predicted ex ante, while this is not possible for the case of variability (we might know the expected number of callers for a day, but the actual number is a realization of a random variable).

Based on the data at hand, we observe that the arrival process is not stationary over a period of several hours. In general, a simple analysis determines whether a process is stationary.

1. Sort all arrival times so that they are increasing in time (label them as $AT_1 \ldots AT_n$).
2. Plot a graph with (x AT_i; $y = i$) as illustrated by Figure 7.7.
3. Add a straight line from the lower left (first arrival) to the upper right (last arrival).

If the underlying arrival process is stationary, there will be no significant deviation between the graph you plotted and the straight line. In this case, however, in Figure 7.7 (left) we observe several deviations between the straight line and the arrival data. Specifically, we observe that for the first hour, fewer calls come in compared to the average arrival rate from 6 a.m. to 10 a.m. In contrast, around 8:30 a.m., the arrival rate becomes much higher than the average. Thus, our analysis indicates that the arrival process we face is not stationary.

When facing nonstationary arrival processes, the best way to proceed is to divide up the day (the week, the month) into smaller time intervals and have a separate arrival rate for each interval. If we then look at the arrival process within the smaller intervals—in our case, we use 15-minute intervals—we find that the seasonality within the interval is relatively low. In other words, within the interval, we come relatively close to a stationary arrival stream. The stationary behavior of the interarrivals within a 15-minute interval is illustrated by Figure 7.7 (right).

Figure 7.7 (left) is interesting to compare with Figure 7.7 (right): the arrival process behaves as stationary "at the micro-level" of a 15-minute interval, yet exhibits strong seasonality over the course of the entire day, as we observed in Figure 7.6. Note that the peaks in Figure 7.6 correspond to those time slots where the line of "actual, cumulative arrivals" in Figure 7.7 grows faster than the straight line "predicted arrivals."

In most cases in practice, the context explains this type of seasonality. For example, in the case of An-ser, the spike in arrivals corresponds to people beginning their day, expecting that the company they want to call (e.g., a doctor's office) is already "up and running." However, since many of these firms are not handling calls before 9 a.m., the resulting call stream is channeled to the answering service.

Exponential Interarrival Times

Interarrival times commonly are distributed following an *exponential distribution.* If IA is a random interarrival time and the interarrival process follows an exponential distribution, we have

$$\text{Probability } \{\text{IA} \leq t\} = 1 - e^{-\frac{t}{a}}$$

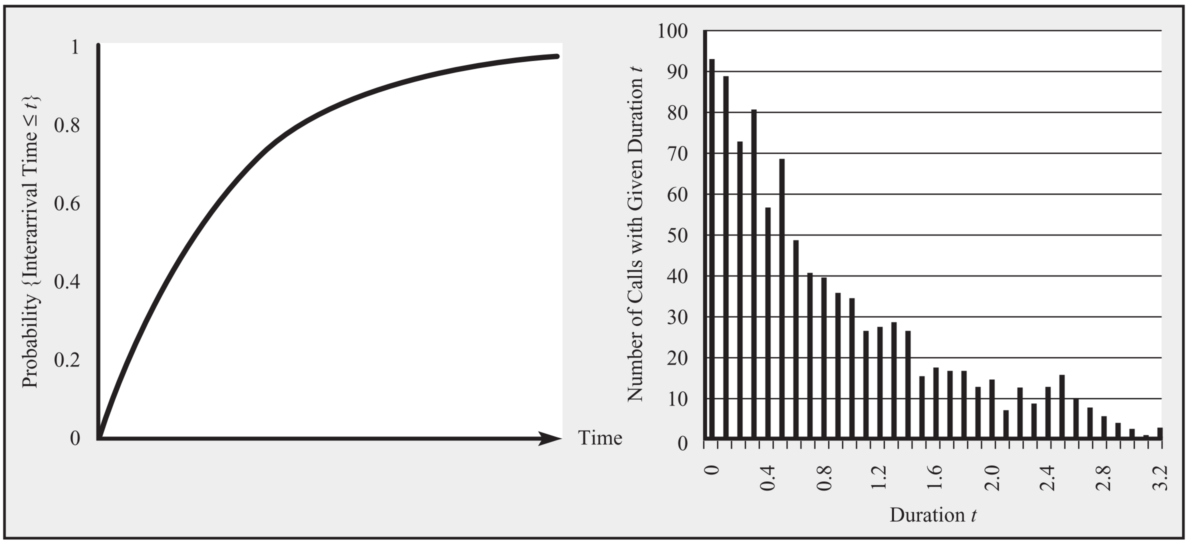
where a is the average interarrival time as defined above. Exponential functions are frequently used to model interarrival time in theory as well as practice, both because of their good fit with empirical data as well as their analytical convenience. If an arrival process has indeed exponential interarrival times, we refer to it as a *Poisson arrival process.*

It can be shown analytically that customers arriving independently from each other at the process (e.g., customers calling into a call center) form a demand pattern with exponential interarrival times. The shape of the cumulative distribution function for the exponential distribution is given in Figure 7.8. The average interarrival time is in minutes. An important property of the exponential distribution is that the standard deviation is also equal to the average, a.

Another important property of the exponential distribution is known as the *memoryless property.* The memoryless property simply states that the number of arrivals in the next time slot (e.g., 1 minute) is independent of when the last arrival has occurred.

To illustrate this property, consider the situation of an emergency room. Assume that, on average, a patient arrives every 10 minutes and no patients have arrived for the last

FIGURE 7.8 **Distribution Function of the Exponential Distribution (left) and an Example of a Histogram (right)**



20 minutes. Does the fact that no patients have arrived in the last 20 minutes increase or decrease the probability that a patient arrives in the next 10 minutes? For an arrival process with exponential interarrival times, the answer is *no.*

Intuitively, we feel that this is a reasonable assumption in many settings. Consider, again, an emergency room. Given that the population of potential patients for the ER is extremely large (including all healthy people outside the hospital), we can treat new patients as arriving independently from each other (the fact that Joan Wiley fell off her mountain bike has nothing to do with the fact that Joe Hoop broke his ankle when playing basketball).

Because it is very important to determine if our interarrival times are exponentially distributed, we now introduce the following four-step diagnostic procedure:

1. Compute the interarrival times $IA_1 \ldots IA_n$.
2. Sort the interarrival times in increasing order; let a_i denote the ith smallest interarrival time (a_1 is the smallest interarrival time; a_n is the largest).
3. Plot pairs ($x = a_i$, $y = i/n$). The resulting graph is called an empirical distribution function.
4. Compare the graph with an exponential distribution with "appropriately chosen parameter." To find the best value for the parameter, we set the parameter of the exponential distribution equal to the average interarrival time we obtain from our data. If a few observations from the sample are substantially remote from the resulting curve, we might adjust the parameter for the exponential distribution "manually" to improve fit.

Figure 7.9 illustrates the outcome of this process. If the underlying distribution is indeed exponential, the resulting graph will resemble the analytical distribution as in the case of Figure 7.9. Note that this procedure of assessing the goodness of fit works also for any other distribution function.

Nonexponential Interarrival Times

In some cases, we might find that the interarrival times are not exponentially distributed. For example, we might encounter a situation where arrivals are scheduled (e.g., every hour), which typically leads to a lower amount of variability in the arrival process.

FIGURE 7.9
Empirical versus Exponential Distribution for Interarrival Times

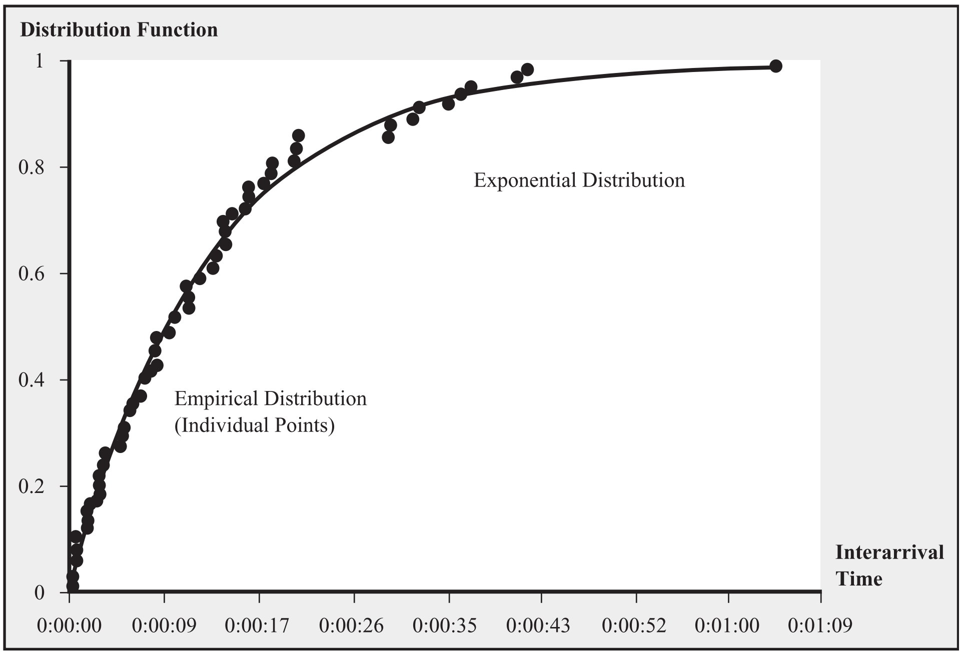
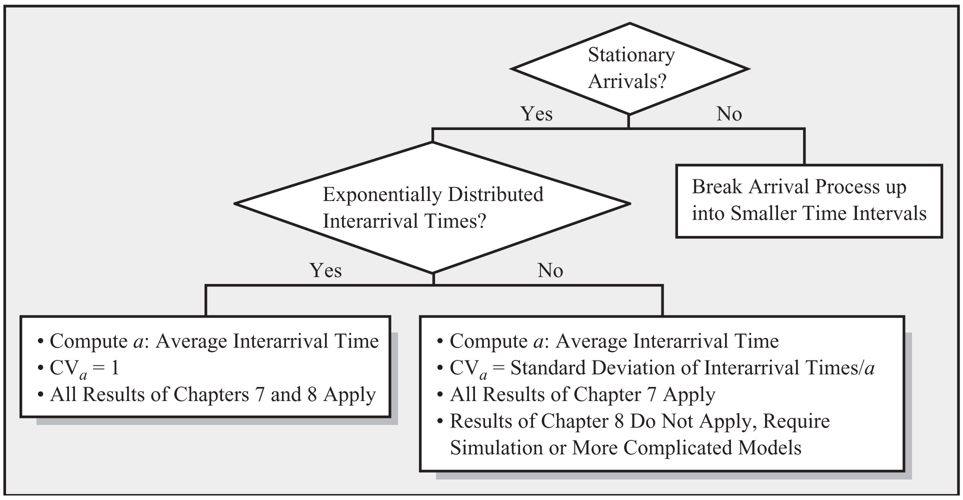


FIGURE 7.10
How to Analyze a Demand/Arrival Process


Stationary Arrivals?
Yes
No
Exponentially Distributed Interarrival Times?
Break Arrival Process up into Smaller Time Intervals
Yes
No
• Compute a: Average Interarrival Time
• $\mathrm{CV}_a = 1$
• All Results of Chapters 7 and 8 Apply
• Compute a: Average Interarrival Time
• CV_a = Standard Deviation of Interarrival Times/a
• All Results of Chapter 7 Apply
• Results of Chapter 8 Do Not Apply, Require Simulation or More Complicated Models

While in the case of the exponential distribution the mean interarrival time is equal to the standard deviation of interarrival times and, thus, one parameter is sufficient to characterize the entire arrival process, we need more parameters to describe the arrival process if interarrival times are not exponentially distributed.

Following our earlier definition of the coefficient of variation, we can measure the variability of an arrival (demand) process as

$$\mathrm{CV}_a = \frac{\text{Standard deviation of interarrival time}}{\text{Average interarrival time}}$$

Given that for the exponential distribution the mean is equal to the standard deviation, its coefficient of variation is equal to 1.

Summary: Analyzing an Arrival Process

Figure 7.10 provides a summary of the steps required to analyze an arrival process. It also shows what to do if any of the assumptions required for the following models (Chapters 7 and 8) are violated.

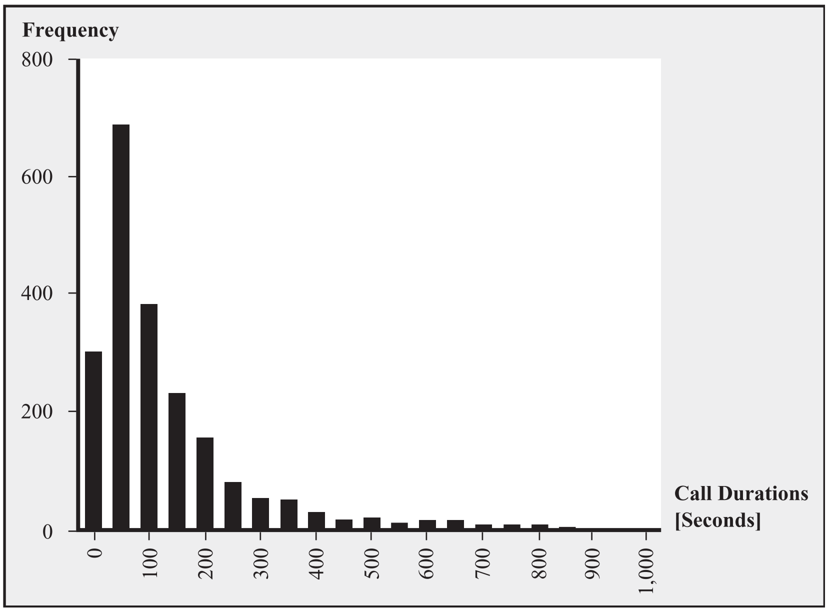
7.4 Service Time Variability

Just as exact arrival time of an individual call is difficult to predict, so is the actual duration of the call. Thus, service processes also have a considerable amount of variability from the supply side. Figure 7.11 provides a summary of call durations (service times from the perspective of the customer service representative) for the case of the An-ser call center.

We observe that the variability in service times is substantial. While some calls were completed in less than a minute, others took more than 10 minutes! Thus, in addition to the variability of demand, variability also is created within the process.

There have been reports of numerous different shapes of activity time distributions. For the purposes of this book, we focus entirely on their mean and standard deviation. In other words, when we collect data, we do not explicitly model the distribution of the service times, but assume that the mean and standard deviation capture all the relevant information. This information is sufficient for all computations in Chapters 7 and 8.

FIGURE 7.11
Service Times in Call Center

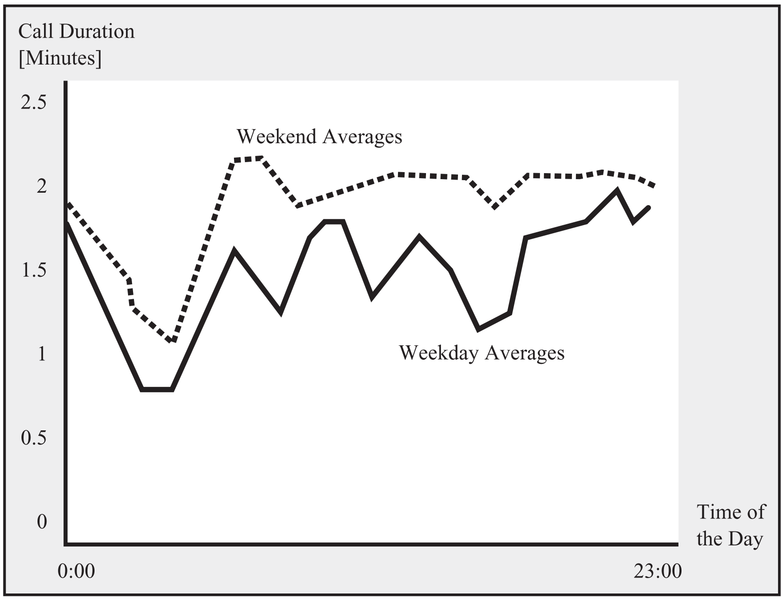


Based on the data summarized in Figure 7.11, we compute the mean call time as 120 seconds and the corresponding standard deviation as 150 seconds. As we have done with the interarrival times, we can now define the coefficient of variation, which we obtain by

$$\text{CV}_p = \frac{\text{Standard deviation of activity time}}{\text{Average activity time}}$$

As with the arrival process, we need to be careful not to confuse variability with seasonality. Seasonality in service times refers to known patterns of call durations as a function of the day of the week or the time of the day (as Figure 7.12 shows, calls take significantly longer on weekends than during the week). Call durations also differ depending on the time of the day.

FIGURE 7.12
Average Call Durations: Weekday versus Weekend



The models we introduce in Chapters 7 and 8 require a stationary service process (in the case of seasonality in the service process, just divide up the time line into smaller intervals, similar to what we did with the arrival process) but do not require any other properties (e.g., exponential distribution of service time). Thus, the standard deviation and mean of the service time are all we need to know.

7.5 Predicting the Average Waiting Time for the Case of One Resource

Based on our measures of variability, we now introduce a simple formula that restores our ability to predict the basic process performance measures: inventory, flow rate, and flow time.

In this chapter, we restrict ourselves to the most basic process diagram, consisting of one buffer with unlimited space and one single resource. This process layout corresponds to the call center example discussed above. Figure 7.13 shows the process flow diagram for this simple system.

Flow units arrive to the system following a demand pattern that exhibits variability. On average, a flow unit arrives every a time units. We labeled a as the average interarrival time. This average reflects the mean of interarrival times IA_1 to IA_n. After computing the standard deviation of the IA_1 to IA_n interarrival times, we can compute the coefficient of variation CV_a of the arrival process as discussed previously.

Assume that it takes on average p units of time to serve a flow unit. Similar to the arrival process, we can define p_1 to p_n as the empirically observed activity times and compute the coefficient of variation for the processing times, CV_p, accordingly. Given that there is only one single resource serving the arriving flow units, the capacity of the server can be written as $1/p$.

As discussed in the introduction to this chapter, we are considering cases in which the capacity exceeds the demand rate; thus, the resulting utilization is strictly less than 100 percent. If the utilization were above 100 percent, inventory would predictably build up and we would not need any sophisticated tools accounting for variability to predict that flow units will incur waiting times. However, the most important insight of this chapter is that flow units incur waiting time even if the server utilization is below 100 percent.

Given that capacity exceeds demand and assuming we never lose a customer (i.e., once a customer calls, he or she never hangs up), we are demand-constrained and, thus, the flow rate R is the demand rate. (Chapter 8 deals with the possibility of lost customers.) Specifically, since a customer arrives, on average, every a units of time, the flow rate $R = 1/a$. Recall that we can compute utilization as

$$\text{Utilization} = \frac{\text{Flow rate}}{\text{Capacity}} = \frac{1/a}{1/p} = p/a < 100\%$$

Note that, so far, we have not applied any concept that went beyond the deterministic process analysis we discussed in Chapters 3 to 6.

FIGURE 7.13
A Simple Process with One Queue and One Server

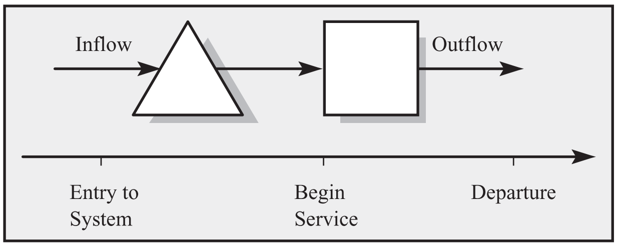
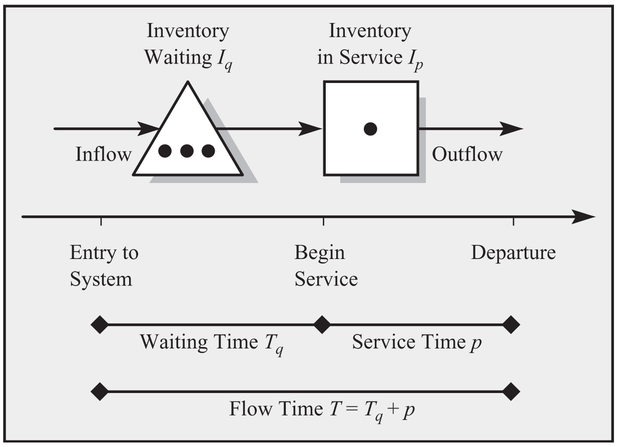


FIGURE 7.14
A Simple Process with One Queue and One Server



Now, take the perspective of a flow unit moving through the system (see Figure 7.14). A flow unit can spend time waiting in the queue (in a call center, this is the time when you listen to Music of the '70s). Let T_q denote the time the flow unit has to spend in the queue waiting for the service to begin. The subscript q denotes that this is only the time the flow unit waits in the queue. Thus, T_q does *not* include the actual service time, which we defined as p. Based on the waiting time in the queue T_q and the average service time p, we can compute the flow time (the time the flow unit will spend in the system) as

$$\text{Flow time} = \text{Time in queue} + \text{Activity time}$$
$$T = T_q + p$$

Instead of taking the perspective of the flow unit, we also can look at the system as a whole, wondering how many flow units will be in the queue and how many will be in service. Let I_q be defined as the inventory (number of flow units) that are in the queue and I_p be the number of flow units in process. Since the inventory in the queue I_q and the inventory in process I_p are the only places we can find inventory, we can compute the overall inventory in the system as $I = I_q + I_p$.

As long as there exists only one resource, I_p is a number between zero and one: sometimes there is a flow unit in service ($I_p = 1$); sometimes there is not ($I_p = 0$). The probability that at a random moment in time the server is actually busy, working on a flow unit, corresponds to the utilization. For example, if the utilization of the process is 30 percent, there exists a .3 probability that at a random moment in time the server is busy. Alternatively, we can say that over the 60 minutes in an hour, the server is busy for

$$.3 \times 60 \text{ [minutes/hour]} = 18 \text{ minutes}$$

While the inventory in service I_p and the activity time p are relatively easy to compute, this is unfortunately not the case for the inventory in the queue I_q or the waiting time in the queue T_q.

Based on the activity time p, the utilization, and the variability as measured by the coefficients of variation for the interarrival time CV_a and the processing time CV_p, we can compute the average waiting time in the queue using the following formula:

$$\text{Time in queue} = \text{Activity time} \times \left(\frac{\text{Utilization}}{1 - \text{Utilization}}\right) \times \left(\frac{\text{CV}_a^2 + \text{CV}_p^2}{2}\right)$$

The formula does not require that the service times or the interarrival times follow a specific distribution. Yet, for the case of nonexponential interarrival times, the formula only approximates the expected time in the queue, as opposed to being 100 percent exact. The formula should be used only for the case of a stationary process (see Section 7.3 for the definition of a stationary process as well as for what to do if the process is not stationary).

The above equation states that the waiting time in the queue is the product of three factors:

- The waiting time is expressed as multiples of the activity time. However, it is important to keep in mind that the activity time also directly influences the utilization (as Utilization = Activity time/Interarrival time). Thus, one should not think of the waiting time as increasing linearly with the activity time.
- The second factor captures the utilization effect. Note that the utilization has to be less than 100 percent. If the utilization is equal to or greater than 100 percent, the queue continues to grow. This is not driven by variability, but simply by not having the requested capacity. We observe that the utilization factor is nonlinear and becomes larger and larger as the utilization level is increased closer to 100 percent. For example, for Utilization = 0.8, the utilization factor is 0.8/(1 − 0.8) = 4; for Utilization = 0.9, it is 0.9/(1 − 0.9) = 9; and for Utilization = 0.95, it grows to 0.95/(1 − 0.95) = 19.
- The third factor captures the amount of variability in the system, measured by the average of the squared coefficient of variation of interarrival times CV_a and activity times CV_p. Since CV_a and CV_p affect neither the average activity time p nor the utilization u, we observe that the waiting time grows with the variability in the system.

The best way to familiarize ourselves with this newly introduced formula is to apply it and "see it in action." Toward that end, consider the case of the An-ser call center at 2:00 a.m. in the morning. An-ser is a relatively small call center and they receive very few calls at this time of the day (see Section 7.3 for detailed arrival information), so at 2:00 a.m., there is only one person handling incoming calls.

From the data we collected in the call center, we can quickly compute that the average activity time at An-ser at this time of the day is around 90 seconds. Given that we found in the previous section that the activity time does depend on the time of the day, it is important that we use the service time data representative for these early morning hours: Activity time $p = 90$ seconds.

Based on the empirical service times we collected in Section 7.4, we now compute the standard deviation of the service time to be 120 seconds. Hence, the coefficient of variation for the activity time is

$$\mathrm{CV}_p = 120 \text{ seconds}/90 \text{ seconds} = 1.3333$$

From the arrival data we collected (see Figure 7.6), we know that at 2:00 a.m. there are 3 calls arriving in a 15-minute interval. Thus, the interarrival time is $a = 5$ minutes = 300 seconds. Given the activity time and the interarrival time, we can now compute the utilization as

$$\begin{aligned}\text{Utilization} &= \text{Activity time/Interarrival time } (= p/a)\\ &= 90 \text{ seconds}/300 \text{ seconds} = 0.3\end{aligned}$$

Concerning the coefficient of variation of the interarrival time, we can take one of two approaches. First, we could take the observed interarrival times and compute the standard deviation empirically. Alternatively, we could view the arrival process during the time period as random. Given the good fit between the data we collected and the exponential distribution (see Figure 7.9), we assume that arrivals follow a Poisson process (interarrival times are exponentially distributed). This implies a coefficient of variation of

$$\mathrm{CV}_a = 1$$

Substituting these values into the waiting time formula yields

$$\begin{aligned}\text{Time in queue} &= \text{Activity time} \times \left(\frac{\text{Utilization}}{1-\text{Utilization}}\right) \times \left(\frac{\text{CV}_a^2 + \text{CV}_p^2}{2}\right)\\ &= 90 \times \frac{0.3}{1-0.3} \times \frac{1^2 + 1.3333^2}{2}\\ &= 53.57 \text{ seconds}\end{aligned}$$

Note that this result captures the average waiting time of a customer before getting served. To obtain the customer's total time spent for the call, including waiting time and service time, we need to add the activity time p for the actual service. Thus, the flow time can be computed as

$$T = T_q + p = 53.57 \text{ seconds} + 90 \text{ seconds} = 143.57 \text{ seconds}$$

It is important to point out that the value 53.57 seconds provides the average waiting time. The actual waiting times experienced by individual customers vary. Some customers get lucky and receive service immediately; others have to wait much longer than 53.57 seconds. This is discussed further below.

Waiting times computed based on the methodology outlined above need to be seen as long-run averages. This has the following two practical implications:

- If the system would start empty (e.g., in a hospital lab, where there are no patients before the opening of the waiting room), the first couple of patients are less likely to experience significant waiting time. This effect is transient: Once a sufficient number of patients have arrived, the system reaches a "steady-state." Note that given the 24-hour operation of An-ser, this is not an issue in this specific case.
- If we observe the system for a given time interval, it is unlikely that the average waiting time we observe within this interval is exactly the average we computed. However, the longer we observe the system, the more likely the expected waiting time T_q will indeed coincide with the empirical average. This resembles a casino, which cannot predict how much money a specific guest will win (or typically lose) in an evening, yet can well predict the economics of the entire guest population over the course of a year.

Now that we have accounted for the waiting time T_q (or the flow time T), we are able to compute the resulting inventory. With $1/a$ being our flow rate, we can use Little's Law to compute the average inventory I as

$$\begin{aligned}I = R \times T &= \frac{1}{a} \times (T_q + p)\\ &= 1/300 \times (53.57 + 90) = 0.479\end{aligned}$$

Thus, there is, on average, about half a customer in the system (it is 2:00 a.m. after all . . .). This inventory includes the two subsets we defined as inventory in the queue (I_q) and inventory in process (I_p):

- I_q can be obtained by applying Little's Law, but this time, rather than applying Little's Law to the entire system (the waiting line and the server), we apply it only to the waiting line in isolation. If we think of the waiting line as a mini process in itself (the corresponding process flow diagram consists only of one triangle), we obtain a flow time of T_q. Hence,

$$I_q = 1/a \times T_q = 1/300 \times 53.57 = 0.179$$

- At any given moment in time, we also can look at the number of customers that are currently talking to the customer service representative. Since we assumed there would only be one representative at this time of the day, there will never be more than one caller at this stage. However, there are moments in time when no caller is served, as the utilization of the employee is well below 100 percent. The average number of callers in service can thus be computed as

$$\begin{aligned} I_p &= \text{Probability}\{\text{0 callers talking to representative}\} \times 0 \\ &\quad + \text{Probability}\{\text{1 caller talking to representative}\} \times 1 \\ I_p &= (1 - u) \times 0 + u \times 1 = u \end{aligned}$$

In this case, we obtain $I_p = 0.3$.

7.6 Predicting the Average Waiting Time for the Case of Multiple Resources

After analyzing waiting time in the presence of variability for an extremely simple process, consisting of just one buffer and one resource, we now turn to more complicated operations. Specifically, we analyze a waiting time model of a process consisting of one waiting area (queue) and a process step performed by multiple, identical resources.

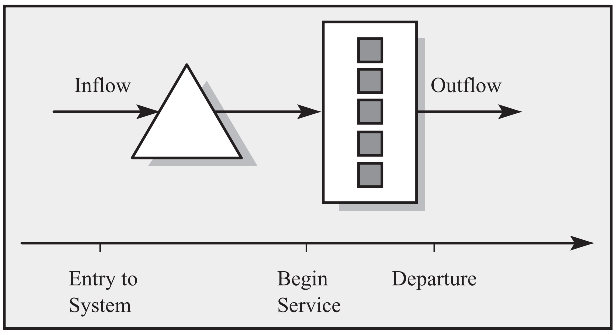
We continue our example of the call center. However, now we consider time slots at more busy times over the course of the day, when there are many more customer representatives on duty in the An-ser call center. The basic process layout is illustrated in Figure 7.15.

Let m be the number of parallel servers we have available. Given that we have m servers working in parallel, we now face a situation where the average service time is likely to be much longer than the average interarrival time. Taken together, the m resources have a capacity of m/p, while the demand rate continues to be given by $1/a$. We can compute the utilization u of the service process as

$$\text{Utilization} = \frac{\text{Flow rate}}{\text{Capacity}} = \frac{1/\text{Interarrival time}}{(\text{Number of resources}/\text{Activity time})} = \frac{1/a}{m/p} = \frac{p}{a \times m}$$

Similar to the case with one single resource, we are only interested in the cases of utilization levels below 100 percent.

FIGURE 7.15
A Process with One Queue and Multiple, Parallel Servers ($m = 5$)



The flow unit will initially spend T_q units of time waiting for service. It then moves to the next available resource, where it spends p units of time for service. As before, the total flow time is the sum of waiting time and service time:

$$\text{Flow time} = \text{Waiting time in queue} + \text{Activity time}$$
$$T = T_q + p$$

Based on the activity time p, the utilization u, the coefficients of variation for both service (CV_p) and arrival process (CV_a) as well as the number of resources in the system (m), we can compute the average waiting time T_q using the following formula:[1]

$$\text{Time in queue} = \left(\frac{\text{Activity time}}{m}\right) \times \left(\frac{\text{Utilization}^{\sqrt{2(m+1)}-1}}{1 - \text{Utilization}}\right) \times \left(\frac{\text{CV}_a^2 + \text{CV}_p^2}{2}\right)$$

As in the case of one single resource, the waiting time is expressed as the product of the activity time, a utilization factor, and a variability factor. We also observe that for the special case of $m = 1$, the above formula is exactly the same as the waiting time formula for a single resource. Note that all other performance measures, including the flow time (T), the inventory in the system (I), and the inventory in the queue (I_q), can be computed as discussed before.

While the above expression does not necessarily seem an inviting equation to use, it can be programmed without much effort into a spreadsheet. Furthermore, it provides the average waiting time for a system that otherwise could only be analyzed with much more sophisticated software packages.

Unlike the waiting time formula for the single resource case, which provides an exact quantification of waiting times as long as the interarrival times follow an exponential distribution, the waiting time formula for multiple resources is an approximation. The formula works well for most settings we encounter, specifically if the ratio of utilization u to the number of servers m is large (u/m is high).

Now that we have computed waiting time, we can again use Little's Law to compute the average number of flow units in the waiting area I_q, the average number of flow units in service I_p, and the average number of flow units in the entire system $I = I_p + I_q$. Figure 7.16 summarizes the key performance measures.

[1] Hopp and Spearman (1996); the formula initially had been proposed by Sakasegawa (1977) and used successfully by Whitt (1983). For $m = 1$, the formula is exactly the same as in the previous section. The formula is an approximation for $m > 1$. An exact expression for this case does not exist.

FIGURE 7.16
Summary of Key Performance Measures

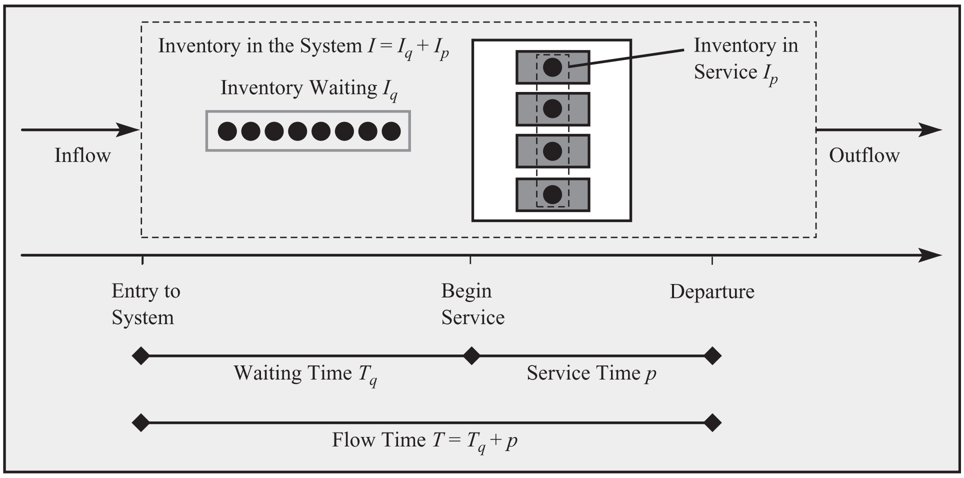


Exhibit 7.1

SUMMARY OF WAITING TIME CALCULATIONS

1. Collect the following data:
   - Number of servers, m
   - Activity time, p
   - Interarrival time, a
   - Coefficient of variation for interarrival (CV_a) and processing time (CV_p)
2. Compute utilization: $u = \frac{p}{a \times m}$
3. Compute expected waiting time:

$$T_q = \left(\frac{\text{Activity time}}{m}\right) \times \left(\frac{\text{Utilization}^{\sqrt{2(m+1)}-1}}{1 - \text{Utilization}}\right) \times \left(\frac{\mathrm{CV}_a^2 + \mathrm{CV}_a^2}{2}\right)$$

4. Based on T_q, we can compute the remaining performance measures as

$$\begin{aligned} \text{Flow time } T &= T_q + p \\ \text{Inventory in service } I_p &= m \times u \\ \text{Inventory in the queue } I_q &= T_q/a \\ \text{Inventory in the system } I &= I_p + I_q \end{aligned}$$

Note that in the presence of multiple resources serving flow units, there can be more than one flow unit in service simultaneously. If u is the utilization of the process, it is also the utilization of each of the m resources, as they process demand at the same rate. We can compute the expected number of flow units at any of the m resources *in isolation* as

$$u \times 1 + (1 - u) \times 0 = u$$

Adding up across the m resources then yields

$$\begin{aligned} \text{Inventory in process} &= \text{Number of resources} \times \text{Utilization} \\ I_p &= m \times u \end{aligned}$$

We illustrate the methodology using the case of An-ser services. Assuming we would work with a staff of 10 customer service representatives (CSRs) for the 8:00 a.m. to 8:15 a.m. time slot, we can compute the utilization as follows:

$$\text{Utilization } u = \frac{p}{a \times m} = \frac{90 \text{ [seconds/call]}}{11.39 \times 10 \text{ [seconds/call]}} = 0.79$$

where we obtained the interarrival time of 11.39 seconds between calls by dividing the length of the time interval (15 minutes = 900 seconds) by the number of calls received over the interval (79 calls). This now allows us to compute the average waiting time as

$$\begin{aligned} T_q &= \left(\frac{p}{m}\right) \times \left(\frac{u^{\sqrt{2(m+1)}-1}}{1-u}\right) \times \left(\frac{CV_a^2 + \mathrm{CV}_p^2}{2}\right) \\ &= \left(\frac{90}{10}\right) \times \left(\frac{0.79^{\sqrt{2(10+1)}-1}}{1 - 0.79}\right) \times \left(\frac{1 + 1.3333^2}{2}\right) = 24.98 \text{ seconds} \end{aligned}$$

The most important calculations related to waiting times caused by variability are summarized in Exhibit 7.1.

7.7 Service Levels in Waiting Time Problems

So far, we have focused our attention on the average waiting time in the process. However, a customer requesting service from our process is not interested in the average time he or she waits in queue or the average total time to complete his or her request (waiting time T_q and flow time T respectively), but in the wait times that he or she experiences personally.

Consider, for example, a caller who has just waited for 15 minutes listening to music while on hold. This caller is likely to be unsatisfied about the long wait time. Moreover, the response from the customer service representative of the type "we are sorry for your delay, but our average waiting time is only 4 minutes" is unlikely to reduce this dissatisfaction.

Thus, from a managerial perspective, we not only need to analyze the average wait time, but also the likelihood that the wait time exceeds a certain *target wait time* (*TWT*). More formally, we can define the *service level* for a given target wait time as the percentage of customers that will begin service in TWT or less units of waiting time:

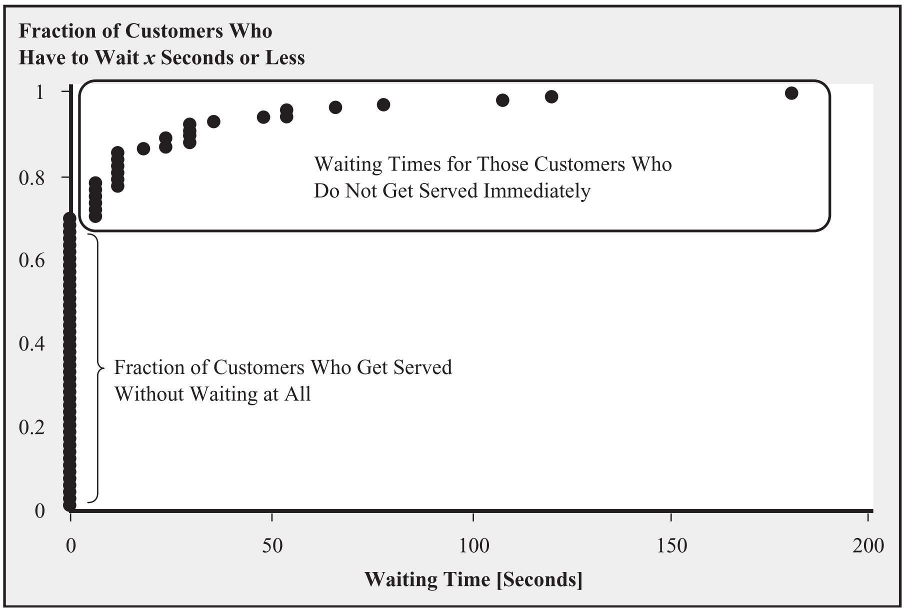
$$\text{Service level} = \text{Probability}\{\text{Waiting time} \leq \text{TWT}\}$$

This service level provides us with a way to measure to what extent the service is able to respond to demand within a consistent waiting time. A service level of 95 percent for a target waiting time of TWT = 2 minutes means that 95 percent of the customers are served in less than 2 minutes of waiting time.

Figure 7.17 shows the empirical distribution function (see Section 7.3 on how to create this graph) for waiting times at the An-ser call center for a selected time slot. Based on the graph, we can distinguish between two groups of customers. About 65 percent of the customers did not have to wait at all and received immediate service. The remaining 35 percent of the customers experienced a waiting time that strongly resembles an exponential distribution.

We observe that the average waiting time for the entire calling population (not just the ones who had to wait) was, for this specific sample, about 10 seconds. For a target wait time TWT = 30 seconds, we find a service level of 90 percent; that is, 90 percent of the callers had to wait 30 seconds or less.

FIGURE 7.17
Empirical Distribution of Waiting Times at An-ser



Service levels as defined above are a common performance measure for service operations in practice. They are used internally by the firm in charge of delivering a certain service. They also are used frequently by firms that want to outsource a service, such as a call center, as a way to contract (and track) the responsiveness of their service provider.

There is no universal rule of what service level is right for a given service operation. For example, responding to large public pressure, the German railway system (Deutsche Bundesbahn) has recently introduced a policy that 80 percent of the calls to their customer complaint number should be handled within 20 seconds. Previously, only 30 percent of the calls were handled within 20 seconds. How fast you respond to calls depends on your market position and the importance of the incoming calls for your business. A service level that worked for the German railway system in 2003 (30 percent within 20 seconds) is likely to be unacceptable in other, more competitive environments.

7.8 Economic Implications: Generating a Staffing Plan

So far, we have focused purely on analyzing the call center for a given number of customer service representatives (CSRs) on duty and predicted the resulting waiting times. This raises the managerial question of how many CSRs An-ser should have at work at any given moment in time over the day. The more CSRs we schedule, the shorter the waiting time, but the more we need to pay in terms of wages.

When making this trade-off, we need to balance the following two costs:

- Cost of waiting, reflecting increased line charges for 1-800 numbers and customer dissatisfaction (line charges are incurred for the actual talk time as well as for the time the customer is on hold).
- Cost of service, resulting from the number of CSRs available.

Additional costs that could be factored into the analysis are

- Costs related to customers calling into the call center but who are not able to gain access even to the waiting line, that is, they receive a busy signal (blocked customers; this will be discussed further in Chapter 8).
- Costs related to customers who hang up while waiting for service.

In the case of An-ser, the average salary of a CSR is $10 per hour. Note that CSRs are paid independent of being idle or busy. Variable costs for a 1-800 number are about $0.05 per minute. A summary of various costs involved in managing a call center—or service operations in general—is given by Figure 7.18.

FIGURE 7.18
Economic Consequences of Waiting

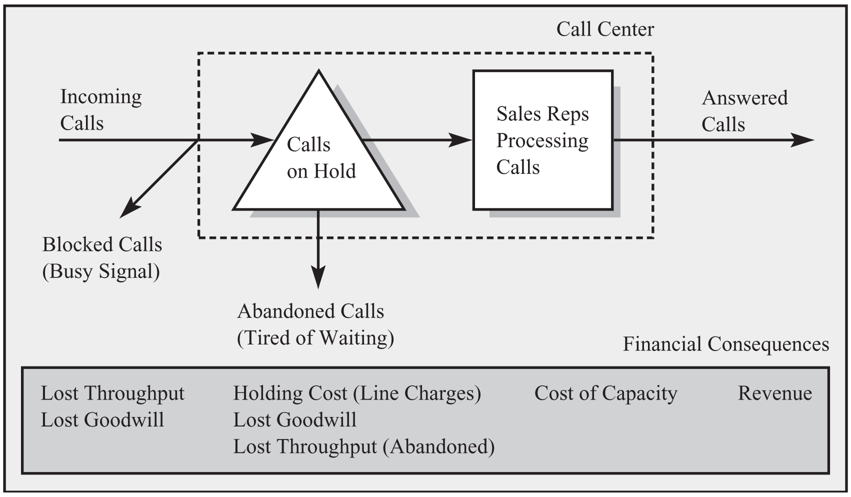

Call Center
Incoming Calls
Calls on Hold
Sales Reps Processing Calls
Answered Calls
Blocked Calls (Busy Signal)
Abandoned Calls (Tired of Waiting)
Financial Consequences
Lost Throughput
Lost Goodwill
Holding Cost (Line Charges)
Lost Goodwill
Lost Throughput (Abandoned)
Cost of Capacity
Revenue

TABLE 7.2
Determining the Number of CSRs to Support Target Wait Time

Number of CSRs, m	Utilization $u = p/(a \times m)$	Expected Wait Time T_q [seconds] Based on Waiting Time Formula
8	0.99	1221.23
9	0.88	72.43
10	0.79	24.98
11	0.72	11.11
12	0.66	5.50
13	0.61	2.89
14	0.56	1.58

When deciding how many CSRs to schedule for a given time slot, we first need to decide on how responsive we want to be to our customers. For the purpose of our analysis, we assume that the management of An-ser wants to achieve an average wait time of 10 seconds. Alternatively, we also could set a service level and then staff according to a TWT constraint, for example, 95 percent of customers to be served in 20 seconds or less.

Now, for a given arrival rate, we need to determine the number of CSRs that will correspond to an average wait time of 10 seconds. Again, consider the time interval from 8:00 to 8:15 a.m.Table 7.2 shows the utilization level as well as the expected wait time for different numbers of customer service representatives. Note that using fewer than 8 servers would lead to a utilization above one, which would mean that queues would build up independent of variability, which is surely not acceptable.

Table 7.2 indicates that adding CSRs leads to a reduction in waiting time. For example, while a staff of 8 CSRs would correspond to an average waiting time of about 20 minutes, the average waiting time falls below 10 seconds once a twelfth CSR has been added. Thus, working with 12 CSRs allows An-ser to meet its target of an average wait time of 10 seconds. In this case, the actual service would be even better and we expect the average wait time for this specific time slot to be 5.50 seconds.

Providing a good service level does come at the cost of increased labor. The more CSRs are scheduled to serve, the lower is their utilization. In Chapter 4 we defined the cost of direct labor as

$$\text{Cost of direct labor} = \frac{\text{Total wages per unit of time}}{\text{Flow rate per unit of time}}$$

where the total wages per unit of time are determined by the number of CSRs m times their wage rate (in our case, \$10 per hour or 16.66 cents per minute) and the flow rate is determined by the arrival rate. Therefore,

$$\text{Cost of direct labor} = \frac{m \times 16.66 \text{ cents/minute}}{1/a} = a \times m \times 16.66 \text{ cents/minute}$$

An alternative way of writing the cost of labor uses the definition of utilization ($u = p/(a \times m)$). Thus, in the above equation, we can substitute p/u for $a \times m$ and obtain

$$\text{Cost of direct labor} = \frac{p \times 16.66 \text{ cents/minute}}{u}$$

This way of writing the cost of direct labor has a very intuitive interpretation: The actual activity time p is inflated by a factor of 1/Utilization to appropriately account for idle time. For example,

TABLE 7.3
Economic Implications of Various Staffing Levels

Number of Servers	Utilization	Cost of Labor per Call	Cost of Line Charges per Call	Total Cost per Call
8	0.988	0.2531	1.0927	1.3458
9	0.878	0.2848	0.1354	0.4201
10	0.790	0.3164	0.0958	0.4122
11	0.718	0.3480	0.0843	0.4323
12	0.658	0.3797	0.0796	0.4593
13	0.608	0.4113	0.0774	0.4887
14	0.564	0.4429	0.0763	0.5193
15	0.527	0.4746	0.0757	0.5503

if utilization were 50 percent, we are charged a $1 of idle time penalty for every $1 we spend on labor productively. In our case, the utilization is 66 percent; thus, the cost of direct labor is

$$\text{Cost of direct labor} = \frac{1.5\ \text{minutes/call} \times 16.66\ \text{cents/minute}}{0.66} = 38\ \text{cents/call}$$

This computation allows us to extend Table 7.2 to include the cost implications of the various staffing scenarios (our calculations do not consider any cost of lost goodwill). Specifically, we are interested in the impact of staffing on the cost of direct labor per call as well as in the cost of line charges.

Not surprisingly, we can see in Table 7.3 that moving from a very high level of utilization of close to 99 percent (using 8 CSRs) to a more responsive service level, for example, as provided by 12 CSRs, leads to a significant increase in labor cost.

At the same time, though, line charges drop from over $1 per call to almost $0.075 per call. Note that $0.075 per call is the minimum charge that can be achieved based on staffing changes, as it corresponds to the pure talk time.

Adding line charges and the cost of direct labor allows us to obtain total costs. In Table 7.3, we observe that total costs are minimized when we have 10 CSRs in service.

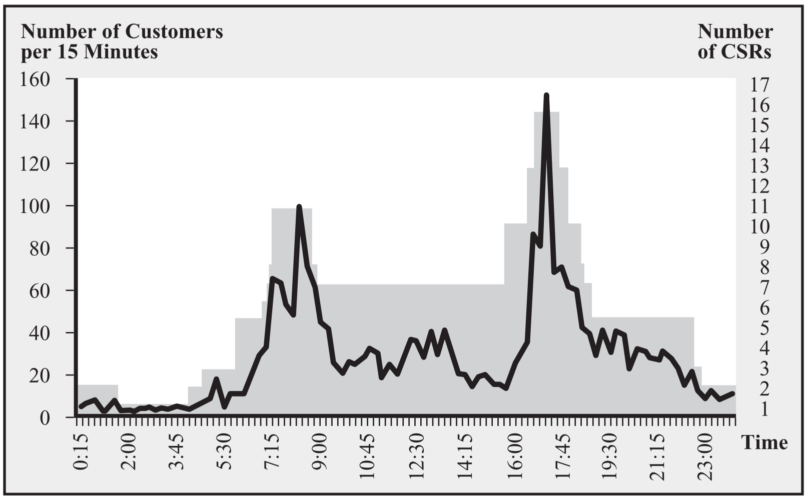
However, we need to be careful in labeling this point as the optimal staffing level, as the total cost number is a purely internal measure and does not take into account any information about the customer's cost of waiting. For this reason, when deciding on an appropriate staffing level, it is important to set acceptable service levels for waiting times as done in Table 7.2 and then staffing up to meet these service levels (opposed to minimizing internal costs).

If we repeat the analysis that we have conducted for the 8:00 to 8:15 a.m. time slot over the 24 hours of the day, we obtain a staffing plan. The staffing plan accounts for both the seasonality observed throughout the day as well as the variability and the resulting need for extra capacity. This is illustrated by Figure 7.19.

When we face a nonstationary arrival process as in this case, a common problem is to decide into how many intervals one should break up the time line to have close to a stationary arrival process within a time interval (in this case, 15 minutes). While we cannot go into the theory behind this topic, the basic intuition is this: It is important that the time intervals are large enough so that

- We have enough data to come up with reliable estimates for the arrival rate of the interval (e.g., if we had worked with 30-second intervals, our estimates for the number of calls arriving within a 30-second time interval would have been less reliable).
- Over the course of an interval, the queue needs sufficient time to reach a "steady state"; this is achieved if we have a relatively large number of arrivals and service completions within the duration of a time interval (more than 10).

FIGURE 7.19
Staffing and Incoming Calls over the Course of a Day



In practice, finding a staffing plan can be somewhat more complicated, as it needs to account for

- Breaks for the operators.
- Length of work period. It is typically not possible to request an operator to show up for work for only a one-hour time slot. Either one has to provide longer periods of time or one would have to temporarily route calls to other members of the organization (supervisor, back-office employees).

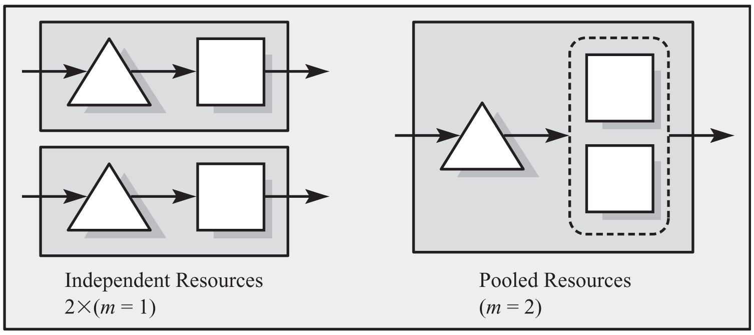
Despite these additional complications, the analysis outlined above captures the most important elements typical for making supply-related decisions in service environments.

7.9 Impact of Pooling: Economies of Scale

Consider a process that currently corresponds to two (m) demand arrival processes that are processed by two (m) identical servers. If demand cannot be processed immediately, the flow unit waits in front of the server where it initially arrived. An example of such a system is provided in Figure 7.20 (left).

Here is an interesting question: Does combining the two systems into a single system with one waiting area and two (m) identical servers lead to lower average waiting times? We refer to such a combination of multiple resources into one "mega-resource" as *pooling*.

FIGURE 7.20
The Concept of Pooling



Consider, for example, two small food services at an airport. For simplicity, assume that both of them have a customer arrival stream with an average interarrival time a of 4 minutes and a coefficient of variation equal to one. The activity time p is three minutes per customer and the coefficient of variation for the service process also is equal to one. Consequently, both food services face a utilization of $p/a = 0.75$.

Using our waiting time formula, we compute the average waiting time as

$$\begin{aligned} T_q &= \text{Activity time} \times \left(\frac{\text{Utilization}}{1 - \text{Utilization}}\right) \times \left(\frac{\text{CV}_a^2 + \text{CV}_p^2}{2}\right) \\ &= 3 \times \left(\frac{0.75}{1 - 0.75}\right) \times \left(\frac{1 + 1}{2}\right) \\ &= 3 \times (0.75/0.25) = 9 \text{ minutes} \end{aligned}$$

Now compare this with the case in which we combine the capacity of both food services to serve the demand of both services. The capacity of the pooled process has increased by a factor of two and now is $\frac{2}{3}$ unit per minute. However, the demand rate also has doubled: If there was one customer every four minutes arriving for service 1 and one customer every four minutes arriving for service 2, the pooled service experiences an arrival rate of one customer every $a = 2$ minutes (i.e., two customers every four minutes is the same as one customer every two minutes).

We can compute the utilization of the pooled process as

$$\begin{aligned} u &= \frac{p}{a \times m} \\ &= 3/(2 \times 2) = 0.75 \end{aligned}$$

Observe that the utilization has not changed compared to having two independent services. Combining two processes with a utilization of 75 percent leads to a pooled system with a 75 percent utilization. However, a different picture emerges when we look at the waiting time of the pooled system. Using the waiting time formula for multiple resources, we can write

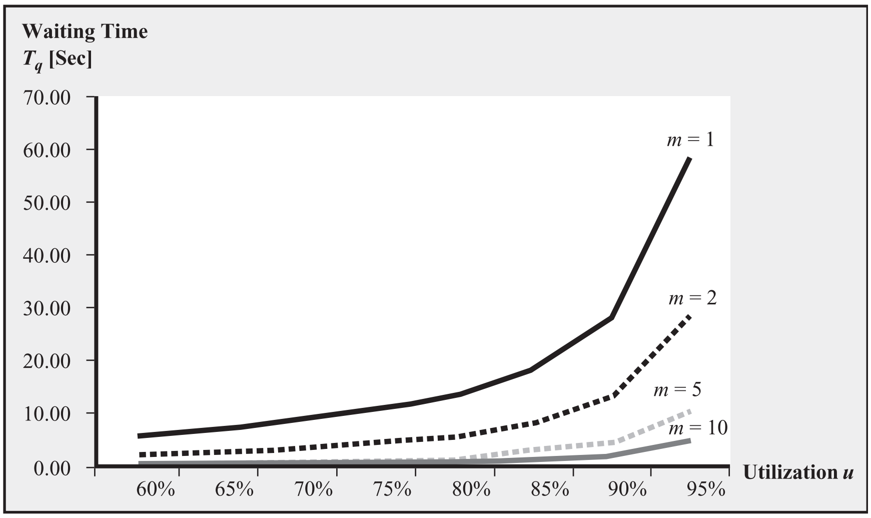
$$\begin{aligned} T_q &= \left(\frac{\text{Activity time}}{m}\right) \times \left(\frac{\text{Utilization}^{\sqrt{2(m+1)}-1}}{1 - \text{Utilization}}\right) \times \left(\frac{\text{CV}_a^2 + \text{CV}_p^2}{2}\right) \\ &= \left(\frac{3}{2}\right) \times \left(\frac{0.75^{\sqrt{2(2+1)}-1}}{1 - 0.75}\right) = 3.95 \text{ minutes} \end{aligned}$$

In other words, the pooled process on the right of Figure 7.20 can serve the same number of customers using the same service time (and thereby having the same utilization), but in only *half* the waiting time!

While short of being a formal proof, the intuition for this result is as follows. The pooled process uses the available capacity more effectively, as it prevents the case that one resource is idle while the other faces a backlog of work (waiting flow units). Thus, pooling identical resources balances the load for the servers, leading to shorter waiting times. This behavior is illustrated in Figure 7.21.

Figure 7.21 illustrates that for a given level of utilization, the waiting time decreases with the number of servers in the resource pool. This is especially important for higher levels of utilization. While for a system with one single server waiting times tend to "go through the

FIGURE 7.21
How Pooling Can Reduce Waiting Time



roof ” once the utilization exceeds 85 percent, a process consisting of 10 identical servers can still provide reasonable service even at utilizations approaching 95 percent.

Given that a pooled system provides better service than individual processes, a service organization can benefit from pooling identical branches or work groups in one of two forms:

- The operation can use pooling to reduce customer waiting time without having to staff extra workers.
- The operation can reduce the number of workers while maintaining the same responsiveness.

These economic benefits of pooling can be illustrated nicely within the context of the An-ser case discussed above. In our analysis leading to Table 7.2, we assumed that there would be 79 calls arriving per 15-minute time interval and found that we would need 12 CSRs to serve customers with an average wait time of 10 seconds or less.

Assume we could pool An-ser’s call center with a call center of comparable size; that is, we would move all CSRs to one location and merge both call centers’ customer populations. Note that this would not necessarily require the two call centers to “move in” with each other; they could be physically separate as long as the calls are routed through one joint network.

Without any consolidation, merging the two call centers would lead to double the number of CSRs and double the demand, meaning 158 calls per 15-minute interval. What would be the average waiting time in the pooled call center? Or, alternatively, if we maintained an average waiting time of 10 seconds or less, how much could we reduce our staffing level? Table 7.4 provides the answers to these questions.

First, consider the row of 24 CSRs, corresponding to pooling the entire staff of the two call centers. Note specifically that the utilization of the pooled call center is not any different from what it was in Table 7.2. We have doubled the number of CSRs, but we also have doubled the number of calls (and thus cut the interarrival time by half). With 24 CSRs, we expect an average waiting time of 1.2 seconds (compared to almost 6 seconds before).

Alternatively, we could take the increased efficiency benefits resulting from pooling by reducing our labor cost. We also observe from Table 7.4 that a staff of 20 CSRs would be able to answer calls with an average wait time of 10 seconds. Thus, we could increase

TABLE 7.4
Pooling Two Call Centers

Number of CSRs	Utilization	Expected Wait Time [seconds]	Labor Cost per Call	Line Cost per Call	Total Cost
16	0.988	588.15	0.2532	0.5651	0.8183
17	0.929	72.24	0.2690	0.1352	0.4042
18	0.878	28.98	0.2848	0.0992	0.3840
19	0.832	14.63	0.3006	0.0872	0.3878
20	0.790	8.18	0.3165	0.0818	0.3983
21	0.752	4.84	0.3323	0.0790	0.4113
22	0.718	2.97	0.3481	0.0775	0.4256
23	0.687	1.87	0.3639	0.0766	0.4405
24	0.658	1.20	0.3797	0.0760	0.4558
25	0.632	0.79	0.3956	0.0757	0.4712
26	0.608	0.52	0.4114	0.0754	0.4868
27	0.585	0.35	0.4272	0.0753	0.5025
28	0.564	0.23	0.4430	0.0752	0.5182
29	0.545	0.16	0.4589	0.0751	0.5340
30	0.527	0.11	0.4747	0.0751	0.5498

utilization to almost 80 percent, which would lower our cost of direct labor from \$0.3797 to \$0.3165. Given an annual call volume of about 700,000 calls, such a saving would be of significant impact for the bottom line.

Despite the nice property of pooled systems outlined above, pooling should not be seen as a silver bullet. Specifically, pooling benefits are much lower than expected (and potentially negative) in the following situations:

- Pooling benefits are significantly lower when the systems that are pooled are not truly independent. Consider, for example, the idea of pooling waiting lines before cash registers in supermarkets, similar to what is done at airport check-ins. In this case, the individual queues are unlikely to be independent, as customers in the current, nonpooled layout will intelligently route themselves to the queue with the shortest waiting line. Pooling in this case will have little, if any, effect on waiting times.
- Similar to the concept of line balancing we introduced earlier in this book, pooling typically requires the service workforce to have a broader range of skills (potentially leading to higher wage rates). For example, an operator sufficiently skilled that she can take orders for hiking and running shoes, as well as provide answering services for a local hospital, will likely demand a higher wage rate than someone who is just trained to do one of these tasks.
- In many service environments, customers value being treated consistently by the same person. Pooling several lawyers in a law firm might be desirable from a waiting-time perspective but ignores the customer desire to deal with one point of contact in the law firm.
- Similarly, pooling can introduce additional setups. In the law-firm example, a lawyer unfamiliar with the situation of a certain client might need a longer time to provide some quick advice on the case and this extra setup time mitigates the operational benefits from pooling.
- Pooling can backfire if pooling combines different customer classes because this might actually increase the variability of the service process. Consider two clerks working in a retail bank, one of them currently in charge of simple transactions (e.g., activity time of 2 minutes per customer), while the other one is in charge of more complex cases (e.g., activity time of 10 minutes). Pooling these two clerks makes the service process more variable and might actually increase waiting time.

7.10 Priority Rules in Waiting Lines

Choosing an appropriate level of capacity helps to prevent waiting lines from building up in a process. However, in a process with variability, it is impossible to eliminate waiting lines entirely. Given, therefore, that at some point in time some customers will have to wait before receiving service, we need to decide on the order in which we permit them access to the server. This order is determined by a *priority rule,* sometimes also referred to as a queuing discipline.

Customers are assigned priorities by adding a (small) step at the point in the process where customers arrive. This process step is called the *triage step.* At triage, we collect information about some of the characteristics of the arriving customer, which we use as input for the priority rule. Below we discuss priority rules based on the following characteristics:

- The service time or the expected service time of the customer (service-time-dependent priority rules).
- Service-time-independent priority rules, including priority rules based on customer arrival time and priority rules based on customer importance or urgency.

Service-Time-Dependent Priority Rules

If it is possible to observe the customer's service time or his or her expected service time prior to initiating the service process, this information should be incorporated when assigning a priority to the customer. The most commonly used service-time-dependent priority rule is the shortest processing time (SPT) rule.

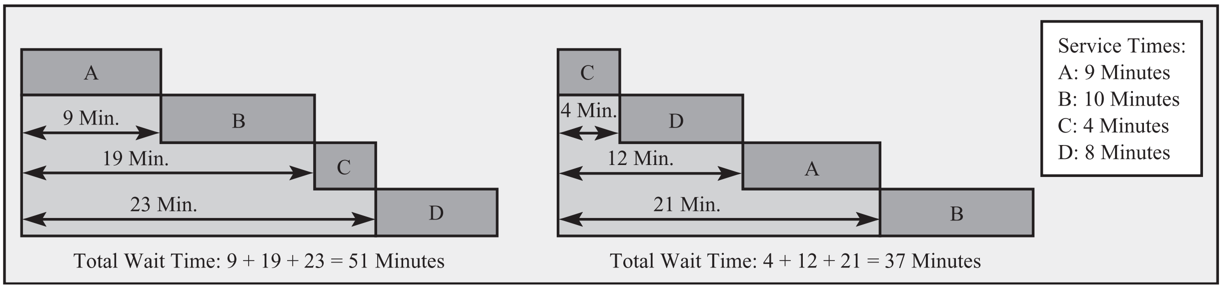
Under the SPT rule, the next available server is allocated to the customer with the shortest (expected) processing time of all customers currently in the waiting line. The SPT rule is extremely effective and performs well, with respect to the expected waiting time as well as to the variance of the waiting time. If the service times are not dependent on the sequence with which customers are processed, the SPT rule can be shown to lead to the shortest average flow time. Its basic intuition is summarized by Figure 7.22.

Service-Time-Independent Priority Rules

In many cases, it is difficult or impossible to assess the service time or even the expected service time prior to initiating the service process. Moreover, if customers are able to misrepresent their service time, then they have an incentive to suggest that their service time is less than it really is when the SPT rule is applied (e.g., "Can I just ask a quick question? . . ."). In contrast, the customer arrival time is easy to observe and difficult for the customer to manipulate.

For example, a call center receiving calls for airline reservations knows the sequence with which callers arrive but does not know which customer has already gathered all relevant information and is ready to order and which customer still requires explanation and discussion.

FIGURE 7.22 The Shortest Processing Time (SPT) Rule (used in the right case)



The most commonly used priority rule based on arrival times is the first-come, first-served (FCFS) rule. With the FCFS rule, the next available server is allocated to the customer in the waiting line with the earliest arrival time.

In addition to using arrival time information, many situations in practice require that characteristics such as the urgency or the importance of the case are considered in the priority rule. Consider the following two examples:

- In an emergency room, a triage nurse assesses the urgency of each case and then assigns a priority to the patient. Severely injured patients are given priority, independent of their arrival times.
- Customers calling in for investor services are likely to experience different priorities, depending on the value of their invested assets. Customers with an investment of greater than $5 million are unlikely to wait, while customers investing only several thousand dollars might wait for 20 minutes or more.

Such urgency-based priority rules are also independent of the service time. In general, when choosing a service-time-independent priority rule, the following property should be kept in mind: Whether we serve customers in the order of their arrival, in the reverse order of their arrival (last-come, first-served), or even in alphabetical order, the expected waiting time does not change. Thus, higher priority service (shorter waiting time) for one customer always requires lower priority (longer waiting time) for other customers.

From an implementation perspective, one last point is worth noting. Using priority rules other than FCFS might be perceived as unfair by the customers who arrived early and are already waiting the longest. Thus, while the average waiting time does not change, serving latecomers first increases the variance of the waiting time. Since variability in waiting time is not desirable from a service-quality perspective, the following property of the FCFS rule is worth remembering: Among service-time-independent priority rules, the FCFS rule minimizes the variance of waiting time and flow time.

7.11 Reducing Variability

In this chapter, we have provided some new methods to evaluate the key performance measures of flow rate, flow time, and inventory in the presence of variability. We also have seen that variability is the enemy of all operations (none of the performance measures improves as variability increases). Thus, in addition to just taking variability as given and adjusting our models to deal with variability, we should always think about ways to reduce variability.

Ways to Reduce Arrival Variability

One—somewhat obvious—way of achieving a match between supply and demand is by “massaging” demand such that it corresponds exactly to the supply process. This is basically the idea of *appointment systems* (also referred to as reservation systems in some industries).

Appointment systems have the potential to reduce the variability in the arrival process as they encourage customers to arrive at the rate of service. However, one should not overlook the problems associated with appointment systems, which include

- Appointment systems do not eliminate arrival variability. Customers do not perfectly arrive at the scheduled time (and some might not arrive at all, “no-shows”). Consequently, any good appointment system needs ways to handle these cases (e.g., extra charge or extra waiting time for customers arriving late). However, such actions are typically very difficult to implement, due to what is perceived to be “fair” and/or “acceptable,” or because variability in service times prevents service providers from always keeping on schedule (and if the doctor has the right to be late, why not the patient?).

- What portion of the available capacity should be reserved in advance. Unfortunately, the customers arriving at the last minute are frequently the most important ones: emergency operations in a hospital do not come through an appointment system and business travelers paying 5 to 10 times the fare of low-price tickets are not willing to book in advance (this topic is further explored in the revenue management chapter, Chapter 15).

The most important limitation, however, is that appointment systems might reduce the variability of the arrival process as seen by the operation, but they do not reduce the variability of the true underlying demand. Consider, for example, the appointment system of a dental office. While the system (hopefully) reduces the time the patient has to wait before seeing the dentist on the day of the appointment, this wait time is not the only performance measure that counts, as the patient might already have waited for three months between requesting to see the dentist and the day of the appointment. Thus, appointment systems potentially hide a much larger supply–demand mismatch and, consequently, any good implementation of an appointment system includes a continuous measurement of both of the following:

- The inventory of customers who have an appointment and are now waiting for the day they are scheduled to go to the dentist.
- The inventory of customers who wait for an appointment in the waiting room of the dentist.

In addition to the concept of appointment systems, we can attempt to influence the customer arrival process (though, for reasons similar to the ones discussed, not the true underlying demand pattern) by providing incentives for customers to avoid peak hours. Frequently observed methods to achieve this include

- Early-bird specials at restaurants or bars.
- Price discounts for hotels during off-peak days (or seasons).
- Price discounts in transportation (air travel, highway tolls) depending on the time of service.
- Pricing of air travel depending on the capacity that is already reserved.

It is important to point out that, strictly speaking, the first three items do not reduce variability; they level expected demand and thereby reduce seasonality (remember that the difference between the two is that seasonality is a pattern known already ex ante). The fourth item refers to the concept of revenue management, which is discussed in Chapter 15.

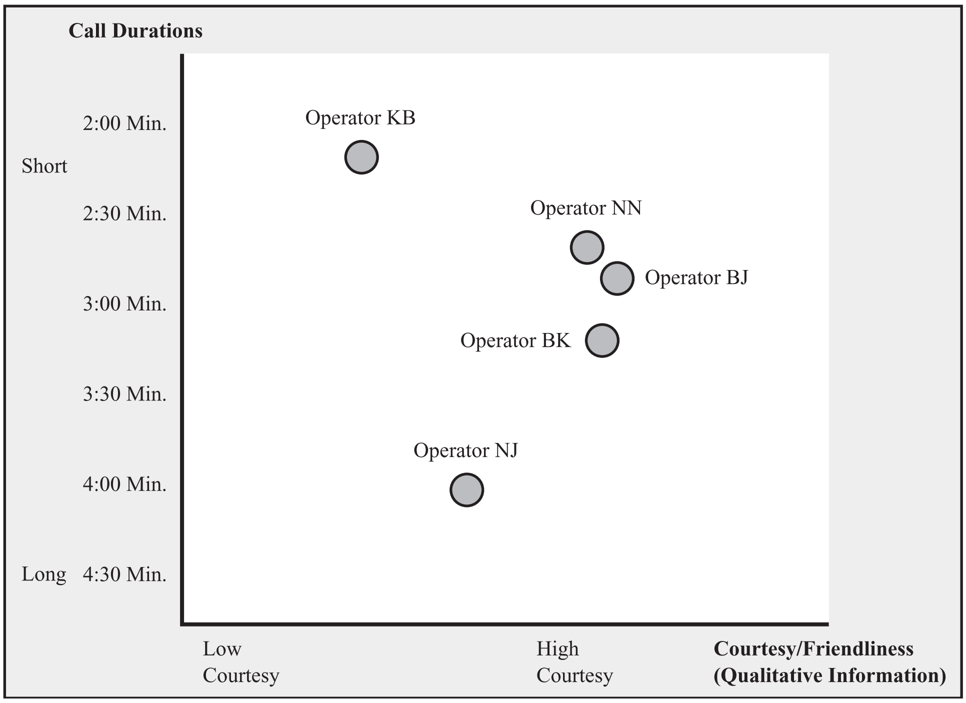
Ways to Reduce Service Time Variability

In addition to reducing variability by changing the behavior of our customers, we also should consider how to reduce internal variability. However, when attempting to standardize activities (reducing the coefficient of variation of the service times) or shorten activity times, we need to find a balance between operational efficiency (call durations) and the quality of service experienced by the customer (perceived courtesy).

Figure 7.23 compares five of An-ser's operators for a specific call service along these two dimensions. We observe that operators NN, BK, and BJ are achieving relatively short call durations while being perceived as friendly by the customers (based on recorded calls). Operator KB has shorter call durations, yet also scores lower on courtesy. Finally, operator NJ has the longest call durations and is rated medium concerning courtesy.

Based on Figure 7.23, we can make several interesting observations. First, observe that there seems to exist a frontier capturing the inherent trade-off between call duration and courtesy. Once call durations for this service go below 2.5 minutes, courtesy seems hard to maintain. Second, observe that operator NJ is away from this frontier, as he is neither

FIGURE 7.23 **Operator Performance Concerning Call Duration and Courtesy**



overly friendly nor fast. Remarkably, this operator also has the highest variability in call durations, which suggests that he is not properly following the operating procedures in place (this is not visible in the graph).

To reduce the inefficiencies of operators away from the frontier (such as NJ), call centers invest heavily in training and technology. For example, technology allows operators to receive real-time instruction of certain text blocks that they can use in their interaction with the customer (scripting). Similarly, some call centers have instituted training programs in which operators listen to tapes of other operators or have operators call other operators with specific service requests. Such steps reduce both the variability of service times as well as their means and, therefore, represent substantial improvements in operational performance.

There are other improvement opportunities geared primarily toward reducing the variability of the service times:

- Although in a service environment (or in a make-to-order production setting) the operator needs to acknowledge the idiosyncrasy of each customer, the operator still can follow a consistent process. For example, a travel agent in a call center might use predefined text blocks (scripts) for his or her interaction with the customer (welcome statement, first question, potential up-sell at the end of the conversation). This approach allowed operators NN, BK, and BJ in Figure 7.23 to be fast and friendly. Thus, being knowledgeable about the process (when to say what) is equally important as being knowledgeable about the product (what to say).
- Activity times in a service environment—unlike activity times in a manufacturing context—are not under the complete control of the resource. The customer him/herself plays a crucial part in the activity at the resource, which automatically introduces a certain amount of variability (e.g., having the customer provide his or her credit card number, having the customer bag the groceries, etc.) What is the consequence of this? At least from a variability perspective, the answer is clear: Reduce the involvement of the customer during

the service at a scarce resource wherever possible (note that if the customer involvement does not occur at a scarce resource, having the customer be involved and thereby do part of the work might be very desirable, e.g., in a self-service setting).

- Variability in service times frequently reflects quality problems. In manufacturing environments, this could include reworking a unit that initially did not meet specifications. However, rework also occurs in service organizations (e.g., a patient who is released from the intensive care unit but later on readmitted to intensive care can be thought of as rework).

Many of these concepts are discussed further in Chapter 9.

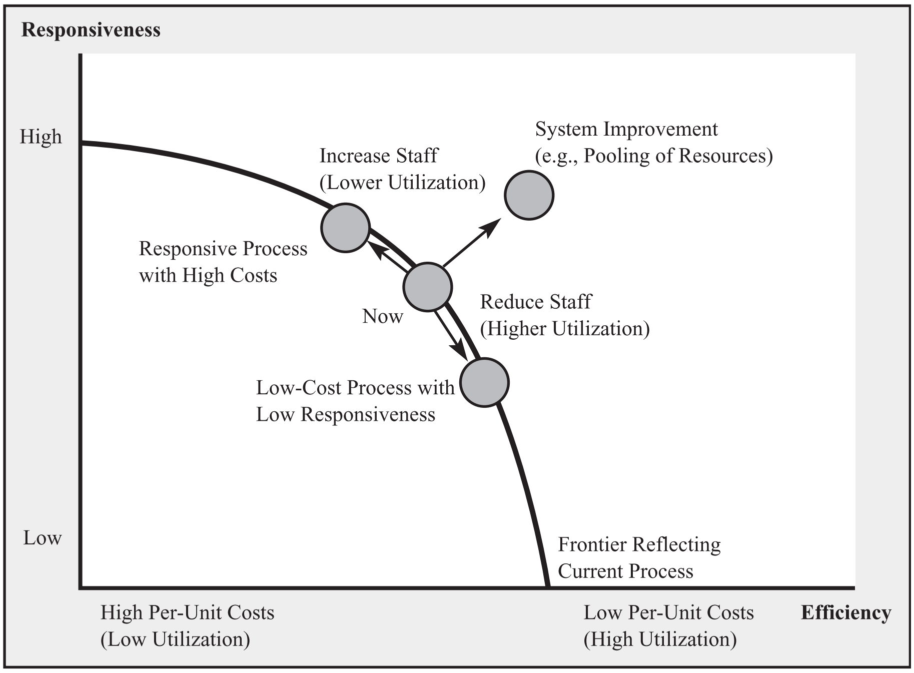
7.12 Summary

In this chapter, we have analyzed the impact of variability on waiting times. As we expected from our more qualitative discussion of variability in the beginning of this chapter, variability causes waiting times, even if the underlying process operates at a utilization level of less than 100 percent. In this chapter, we have outlined a set of tools that allows us to quantify this waiting time, with respect to both the average waiting time (and flow time) as well as the service level experienced by the customer.

There exists an inherent tension between resource utilization (and thereby cost of labor) and responsiveness: Adding service capacity leads to shorter waiting times but higher costs of labor (see Figure 7.24). Waiting times grow steeply with utilization levels. Thus, any responsive process requires excess capacity. Given that capacity is costly, it is important that only as much capacity is installed as is needed to meet the service objective in place for the process. In this chapter, we have outlined a method that allows a service operation to find the point on the frontier that best supports their business objectives (service levels).

However, our results should be seen not only as a way to predict/quantify the waiting time problem. They also outline opportunities for improving the process. Improvement opportunities can be broken up into capacity-related opportunities and system-design-related opportunities, as summarized below.

FIGURE 7.24
Balancing Efficiency with Responsiveness


Responsiveness
High
Low
Increase Staff (Lower Utilization)
System Improvement (e.g., Pooling of Resources)
Responsive Process with High Costs
Now
Reduce Staff (Higher Utilization)
Low-Cost Process with Low Responsiveness
Frontier Reflecting Current Process
High Per-Unit Costs (Low Utilization)
Low Per-Unit Costs (High Utilization)
Efficiency

Capacity-Related Improvements

Operations benefit from flexibility in capacity, as this allows management to adjust staffing levels to predicted demand. For example, the extent to which a hospital is able to have more doctors on duty at peak flu season is crucial in conducting the staffing calculations outlined in this chapter. A different form of flexibility is given by the operation's ability to increase capacity in the case of unpredicted demand. For example, the extent to which a bank can use supervisors and front-desk personnel to help with unexpected spikes in inbound calls can make a big difference in call center waiting times. This leads to the following two improvement opportunities:

- Demand (and sometimes supply) can exhibit seasonality over the course of the day. In such cases, the waiting time analysis should be done for individual time intervals over which the process behaves relatively stationary. System performance can be increased to the extent the organization is able to provide time-varying capacity levels that mirror the seasonality of demand (e.g., Figure 7.19).
- In the presence of variability, a responsive process cannot avoid excess capacity, and thereby will automatically face a significant amount of idle time. In many operations, this idle time can be used productively for tasks that are not (or at least are less) time critical. Such work is referred to as background work. For example, operators in a call center can engage in outbound calls during times of underutilization.

System-Design-Related Improvements

Whenever we face a trade-off between two conflicting performance measures, in this case between responsiveness and efficiency, finding the right balance between the measures is important. However, at least equally important is the attempt to improve the underlying process, shifting the frontier and allowing for higher responsiveness and lower cost (see Figure 7.24). In the context of services suffering from variability-induced waiting times, the following improvement opportunities should be considered:

- By combining similar resources into one joint resource pool (pooling resources), we are able to either reduce wait times for the same amount of capacity or reduce capacity for the same service level. Processes that face variability thereby exhibit very strong scale economies.
- Variability is not exogenous and we should remember to reduce variability wherever possible.
- By introducing a triage step before the actual service process that sequences incoming flow units according to a priority rule (service-time-dependent or service-time-independent), we can reduce the average wait time, assign priority to the most important flow units, or create a waiting system that is perceived as fair by customers waiting in line.

7.13 Further Reading

Gans, Koole, and Mandelbaum (2003) is a recent overview on call-center management from a queuing theory perspective. Further quantitative tools on queueing can be found in Hillier and Lieberman (2002).

Hall (1997) is a very comprehensive and real-world-focused book that provides numerous tools related to variability and its consequences in services and manufacturing.

7.14 Practice Problems

Q7.1* **(Online Retailer)** Customers send e-mails to a help desk of an online retailer every 2 minutes, on average, and the standard deviation of the interarrival time is also 2 minutes. The online retailer has three employees answering e-mails. It takes on average 4 minutes to write a response e-mail. The standard deviation of the service times is 2 minutes.

(* indicates that the solution is at the end of the book)

a. Estimate the average customer wait before being served.

b. How many e-mails would there be, on average, that have been submitted to the online retailer but not yet answered?

Q7.2 **(My-law.com)** My-law.com is a recent start-up trying to cater to customers in search of legal services who are intimidated by the idea of talking to a lawyer or simply too lazy to enter a law office. Unlike traditional law firms, My-law.com allows for extensive interaction between lawyers and their customers via telephone and the Internet. This process is used in the upfront part of the customer interaction, largely consisting of answering some basic customer questions prior to entering a formal relationship.

In order to allow customers to interact with the firm's lawyers, customers are encouraged to send e-mails to my-lawyer@My-law.com. From there, the incoming e-mails are distributed to the lawyer who is currently "on call." Given the broad skills of the lawyers, each lawyer can respond to each incoming request.

E-mails arrive from 8 a.m. to 6 p.m. at a rate of 10 e-mails per hour (coefficient of variation for the arrivals is 1). At each moment in time, there is exactly one lawyer "on call," that is, sitting at his or her desk waiting for incoming e-mails. It takes the lawyer, on average, 5 minutes to write the response e-mail. The standard deviation of this is 4 minutes.

a. What is the average time a customer has to wait for the response to his/her e-mail, ignoring any transmission times? *Note:* This includes the time it takes the lawyer to start writing the e-mail *and* the actual writing time.

b. How many e-mails will a lawyer have received at the end of a 10-hour day?

c. When not responding to e-mails, the lawyer on call is encouraged to actively pursue cases that potentially could lead to large settlements. How much time on a 10-hour day can a My-law.com lawyer dedicate to this activity (assume the lawyer can instantly switch between e-mails and work on a settlement)?

To increase the responsiveness of the firm, the board of My-law.com proposes a new operating policy. Under the new policy, the response would be highly standardized, reducing the standard deviation for writing the response e-mail to 0.5 minute. The average writing time would remain unchanged.

d. How would the amount of time a lawyer can dedicate to the search for large settlement cases change with this new operating policy?

e. How would the average time a customer has to wait for the response to his/her e-mail change? *Note:* This includes the time until the lawyer starts writing the e-mail *and* the actual writing time.

Q7.3 **(Car Rental Company)** The airport branch of a car rental company maintains a fleet of 50 SUVs. The interarrival time between requests for an SUV is 2.4 hours, on average, with a standard deviation of 2.4 hours. There is no indication of a systematic arrival pattern over the course of a day. Assume that, if all SUVs are rented, customers are willing to wait until there is an SUV available. An SUV is rented, on average, for 3 days, with a standard deviation of 1 day.

a. What is the average number of SUVs parked in the company's lot?

b. Through a marketing survey, the company has discovered that if it reduces its daily rental price of $80 by $25, the average demand would increase to 12 rental requests per day and the average rental duration will become 4 days. Is this price decrease warranted? Provide an analysis!

c. What is the average time a customer has to wait to rent an SUV? Please use the initial parameters rather than the information in (b).

d. How would the waiting time change if the company decides to limit all SUV rentals to *exactly* 4 days? Assume that if such a restriction is imposed, the average interarrival time will increase to 3 hours, with the standard deviation changing to 3 hours.

Q7.4 **(Tom Opim)** The following situation refers to Tom Opim, a first-year MBA student. In order to pay the rent, Tom decides to take a job in the computer department of a local department store. His only responsibility is to answer telephone calls to the department, most of which are inquiries about store hours and product availability. As Tom is the only person answering calls, the manager of the store is concerned about queuing problems.

Currently, the computer department receives an average of one call every 3 minutes, with a standard deviation in this interarrival time of 3 minutes.

Tom requires an average of 2 minutes to handle a call. The standard deviation in this activity time is 1 minute.

The telephone company charges $5.00 per hour for the telephone lines whenever they are in use (either while a customer is in conversation with Tom or while waiting to be helped).

Assume that there are no limits on the number of customers that can be on hold and that customers do not hang up even if forced to wait a long time.

a. For one of his courses, Tom has to read a book (*The Pole,* by E. Silvermouse). He can read 1 page per minute. Tom's boss has agreed that Tom could use his idle time for studying, as long as he drops the book as soon as a call comes in. How many pages can Tom read during an 8-hour shift?

b. How long does a customer have to wait, on average, before talking to Tom?

c. What is the average total cost of telephone lines over an 8-hour shift? Note that the department store is billed whenever a line is in use, including when a line is used to put customers on hold.

Q7.5 **(Atlantic Video)** Atlantic Video, a small video rental store in Philadelphia, is open 24 hours a day, and—due to its proximity to a major business school—experiences customers arriving around the clock. A recent analysis done by the store manager indicates that there are 30 customers arriving every hour, with a standard deviation of interarrival times of 2 minutes. This arrival pattern is consistent and is independent of the time of day. The checkout is currently operated by one employee, who needs on average 1.7 minutes to check out a customer. The standard deviation of this check-out time is 3 minutes, primarily as a result of customers taking home different numbers of videos.

a. If you assume that every customer rents at least one video (i.e., has to go to the checkout), what is the average time a customer has to wait in line before getting served by the checkout employee, not including the actual checkout time (within 1 minute)?

b. If there are no customers requiring checkout, the employee is sorting returned videos, of which there are always plenty waiting to be sorted. How many videos can the employee sort over an 8-hour shift (assume no breaks) if it takes exactly 1.5 minutes to sort a single video?

c. What is the average number of customers who are at the checkout desk, either waiting or currently being served (within 1 customer)?

d. Now assume *for this question only* that 10 percent of the customers do not rent a video at all and therefore do not have to go through checkout. What is the average time a customer has to wait in line before getting served by the checkout employee, not including the actual checkout time (within 1 minute)? Assume that the coefficient of variation for the arrival process remains the same as before.

e. As a special service, the store offers free popcorn and sodas for customers waiting in line at the checkout desk. (*Note:* The person who is currently being served is too busy with paying to eat or drink.) The store owner estimates that every minute of customer waiting time costs the store 75 cents because of the consumed food. What is the optimal number of employees at checkout? Assume an hourly wage rate of $10 per hour.

Q7.6 **(RentAPhone)** RentAPhone is a new service company that provides European mobile phones to American visitors to Europe. The company currently has 80 phones available at Charles de Gaulle Airport in Paris. There are, on average, 25 customers per day requesting a phone. These requests arrive uniformly throughout the 24 hours the store is open. (*Note:* This means customers arrive at a faster rate than 1 customer per hour.) The corresponding coefficient of variation is 1.

Customers keep their phones on average 72 hours. The standard deviation of this time is 100 hours.

Given that RentAPhone currently does not have a competitor in France providing equally good service, customers are willing to wait for the telephones. Yet, during the waiting period, customers are provided a free calling card. Based on prior experience, RentA-Phone found that the company incurred a cost of $1 per hour per waiting customer, independent of day or night.

a. What is the average number of telephones the company has in its store?

b. How long does a customer, on average, have to wait for the phone?

c. What are the total monthly (30 days) expenses for telephone cards?

d. Assume RentAPhone could buy additional phones at $1,000 per unit. Is it worth it to buy one additional phone? Why?

e. How would waiting time change if the company decides to limit all rentals to *exactly* 72 hours? Assume that if such a restriction is imposed, the number of customers requesting a phone would be reduced to 20 customers per day.

Q7.7 **(Webflux, Inc.)** Webflux is an Internet-based DVD rental business specializing in hard-to-find, obscure films. Its operating model is as follows. When a customer finds a film on the Webflux Web site and decides to watch it, she puts it in the virtual shopping cart. If a DVD is available, it is shipped immediately (assume it can be shipped during weekends and holidays, too). If not available, the film remains in the customer's shopping cart until a rented DVD is returned to Webflux, at which point it is shipped to the customer if she is next in line to receive it. Webflux maintains an internal queue for each film and a returned DVD is shipped to the first customer in the queue (first-in, first-out).

Webflux has one copy of the 1990 film *Sundown, the Vampire in Retreat,* starring David Carradine and Bruce Campbell. The average time between requests for the DVD is 10 days, with a coefficient of variation of 1. On average, a customer keeps the DVD for 5 days before returning it. It also takes 1 day to ship the DVD to the customer and 1 day to ship it from the customer back to Webflux. The standard deviation of the time between shipping the DVD out from Webflux and receiving it back is 7 days (i.e., it takes on average 7 days to (a) ship it, (b) have it with the customer, and (c) ship it back); hence, the coefficient of variation of this time is 1.

a. What is the average time that a customer has to wait to receive *Sundown, the Vampire in Retreat* DVD after the request? Recall it takes 1 day for a shipped DVD to arrive at a customer address (i.e., in your answer, you have to include the 1-day shipping time).

b. On average, how many customers are in Webflux's internal queue for *Sundown?* Assume customers do not cancel their items in their shopping carts.

Thanks to David Carradine's renewed fame after the recent success of *Kill Bill Vol. I* and *II* which he starred in, the demand for *Sundown* has spiked. Now the average interarrival time for the DVD requests at Webflux is 3 days. Other numbers (coefficient of variation, time in a customer's possession, shipping time) remain unchanged. *For the following question only,* assume sales are lost for customers who encounter stockouts; that is those who cannot find a DVD on the Webflux Web site simply navigate away without putting it in the shopping cart.

c. To satisfy the increased demand, Webflux is considering acquiring a second copy of the *Sundown* DVD. If Webflux owns a total of two copies of *Sundown* DVDs (whether in Webflux's internal stock, in customer's possession, or in transit), what percentage of the customers are turned away because of a stockout? (Note: to answer this question, you will need material from chapter 8.)

Q7.8 **(Security Walking Escorts)** A university offers a walking escort service to increase security around campus. The system consists of specially trained uniformed professional security officers that accompany students from one campus location to another. The service is operated 24 hours a day, seven days a week. Students request a walking escort by phone. Requests for escorts are received, on average, every 5 minutes with a coefficient of variation of 1. After receiving a request, the dispatcher contacts an available escort (via a mobile phone), who immediately proceeds to pick up the student and walk her/him to her/his destination. If there are no escorts available (that is, they are all either walking a student to her/his destination or walking to pick up a student), the dispatcher puts the request in a queue until an escort becomes available. An escort takes, on average, 25 minutes for picking up a student and taking her/him to her/his desired location (the coefficient of variation of this time is also 1). Currently, the university has 8 security officers who work as walking escorts.

a. How many security officers are, on average, available to satisfy a new request?

b. How much time does it take—on average—from the moment a student calls for an escort to the moment the student arrives at her/his destination?

For the next two questions, consider the following scenario. During the period of final exams, the number of requests for escort services increases to 19.2 per hour (one request every 3.125 minutes). The coefficient of variation of the time between successive requests equals 1. However, if a student requesting an escort finds out from the dispatcher that her/his request would have to be put in the queue (i.e., all security officers are busy walking other students), the student cancels the request and proceeds to walk on her/his own.

c. How many students per hour who called to request an escort end up canceling their request and go walking on their own? (Note: to answer this question, you will need material from chapter 8.)

d. University security regulations require that at least 80 percent of the students' calls to request walking escorts have to be satisfied. What is the minimum number of security officers that are needed in order to comply with this regulation?

Q7.9 **(Mango Electronics Inc.)** Mango Electronics Inc. is a *Fortune* 500 company that develops and markets innovative consumer electronics products. The development process proceeds as follows.

Mango researches new technologies to address unmet market needs. Patents are filed for products that have the requisite market potential. Patents are granted for a period of 20 years starting from the date of issue. After receiving a patent, the patented technologies are then developed into marketable products at five independent development centers. Each product is only developed at one center. Each center has all the requisite skills to bring any of the products to market (a center works on one product at a time). On average, Mango files a patent every 7 months (with standard deviation of 7 months). The average development process lasts 28 months (with standard deviation of 56 months).

a. What is the utilization of Mango's development facilities?

b. How long does it take an average technology to go from filing a patent to being launched in the market as a commercial product?

c. How many years of patent life are left for an average product launched by Mango Electronics?

Q7.10 **(UPS Shipping)** A UPS employee, Davis, packs and labels three types of packages: basic packages, business packages, and oversized packages. Business packages take priority over basic packages and oversized packages because those customers paid a premium to have guaranteed two-day delivery. During his nine-hour shift, he has, on average, one container of packages containing a variety of basic, business, and oversized packages to process every 3 hours. As soon as Davis processes a package, he passes it to the next employee, who loads it onto a truck. The times it takes him to process the three different types of packages and the average number of packages per container are shown in the table below.

	Basic	Business	Oversized
Average number of minutes to label and package each unit	5 minutes	4 minutes	6 minutes
Average number of units per container	10	10	5

Davis currently processes packages from each container as follows. First, he processes all business packages in the container. Then he randomly selects either basic packages or oversized packages for processing until the container is empty. However, his manager suggested to Davis that, for each container, he should process all the business packages first, second the basic packages, and last the oversized packages.

a. If Davis follows his supervisor's advice, what will happen to Davis's utilization?

b. What will happen to the average time that a package spends in the container?