2. W. Carley, "A Ride on the Eastern Shuttle Is Rush Hour in Crowded Skies," *The Wall Street Journal*, September 24, 1985, p. 31. Reprinted by permission of *The Wall Street Journal*, © Dow Jones and Company, Inc., 1985. All rights reserved.

3. R. Kotulak, "Transplant Gift in Short Supply," *Chicago Tribune*, July 16, 1986, p. 1. © Copyrighted, Chicago Tribune Company. All rights reserved. Used with permission.

4. D. E. Pitt, "Drug Cases Clog New York Courts," *The New York Times*, April 4, 1989. © Copyright 1989 by The New York Times Company. Reprinted by permission.

5. P. B. Crosby, *Quality Is Free* (New York: McGraw-Hill, 1979), p. 1.

6. From *First Prize: Fifteen Years!* by C. Banc and A. Dundes. Fairleigh Dickinson University Press, Rutherford, N.J., 1986, p. 58. Reprinted by permission of Associated University Presses.

PROBLEMS

1. Based on your understanding of each of the following queueing systems, describe the server, customer, queue organization, queue discipline, and reneging and jockeying processes.
   - **a.** Toll booths on a highway
   - **b.** Traffic intersection controlled by a traffic light
   - **c.** Taxi service at a hotel
   - **d.** Elevators in a large building
   - **e.** Copying service at a copy shop
   - **f.** Mail sorting by the postal service
   - **g.** Medical service at a health center
   - **h.** Airplanes landing at an airport
   - **i.** Emergency telephone lines
   - **j.** A sit-down restaurant
   - **k.** Time-shared computer system
2. Describe a situation (or situations) in which waiting or delay is harmful to productivity. Discuss how the situation might be improved.

EXERCISE: QUEUEING TOUR

Select five queueing systems to visit and observe (your instructor may provide suggestions). Answer the following questions for each site. Be succinct.

1. Describe the server(s) and service process. How many servers are there?
2. Describe the customer(s).
3. Diagram the server configuration and the queue organization.
4. Record the service time for five successive customers. Describe the factors that influence the service time.
5. Count and record the number of customers that arrive over five successive 1-minute intervals. Describe the factors that influence the arrival process.
6. If you observe any reneging or jockeying, describe the process.
7. Describe the queue discipline.
8. Record the time of day and the date.

chapter 2

Observation and Measurement

A doctor examining a patient looks for symptoms of whether the patient is healthy or ill. Blood pressure, body temperature, and pulse are just a few of the things to check. An analyst examining a queueing system also looks for symptoms of health or illness. These symptoms are known as measures of performance (MOP).

Observing the queueing system is the key to the formulation stage of systems analysis. It directly answers the questions: What are the data? and What are the conditions? Understanding how the system currently operates should also be the basis for choosing a new form of operation. Observation should provide insights to help answer the question: What is the unknown? That is, it should help identify, define, and isolate the problem.

Chapter 1 describes how to assess the characteristics of a queueing system. This chapter describes how to measure the performance of the system. It details which queueing symptoms to examine and how they should be measured. It begins by listing some of the key MOPs that apply to almost any queueing system. Next, a major topic that will be used throughout the book is introduced: cumulative arrival and departure diagrams. These diagrams are used to display and analyze the most important MOPs. The diagrams are followed by a discussion of how to calculate averages and standard deviations of the performance measures and how to create an empirical probability distribution. Little's formula is introduced. The chapter concludes with a review of techniques for

measuring and observing queues. Chapter 2 does not address the question of how long to observe the system, nor does it consider how to use observations to predict future queue performance, subjects that are covered later in the book.

2.1 MEASURES OF PERFORMANCE

One should consider two perspectives in measuring the performance of a queueing system. From the customer's perspective, the major concern is the quality of the service provided. From the server's perspective (or the server's employer's perspective), the major concerns are the cost of providing the service and the impact of service quality on business (that is, revenue). Measures of performance allow these concerns to be quantified.

The following sections divide the MOPs along the lines of customer concerns and server concerns. But before discussing the measures of performance, a few basic terms must be defined:

Definitions 2.1

Arrival time The time that the customer arrives at the queue

Departure time The time that the customer completes service and leaves the queueing system

Departure time from queue The time that the customer leaves the queue to enter service

Time in queue Departure time *from queue* minus the arrival time

Service time Departure time minus departure time *from queue*

Time in system Departure time minus the arrival time = time in queue plus service time

As an example, suppose that a customer joined a bank queue at 2:00, stepped up to the teller window at 2:03, and finished the transaction at 2:05. Thus, 2:00 would be the arrival time, 2:03 the departure time from queue, and 2:05 the departure time from the system. The time in queue would be 3 minutes, service time 2 minutes, and the time in system 5 minutes. Therefore, *a customer is "in the system" when it is either in the queue or in service.*

2.1.1. Customer Measures of Performance

In most situations, the major concerns of the customer are the length of time spent in the queue and in service and the cost associated with this time. These are reflected in the following measures of performance:

Time in queue In essence, a shorter wait is better than a longer wait.

Service time The time in service might be treated separately from the time in queue if customers find their time waiting in queue more costly than their time being served (or vice versa). If they are perceived more or less the same, *time in system* might substitute for service time and time in queue.

Waiting cost Waiting time might be more costly for some customers than others, in which case some weighting scheme might be used to combine the various waiting times. The combined waiting time could represent cost.

Proportion of work completed on time Customers may have deadlines to meet. If they are not served by a deadline, a penalty is incurred.

Tardiness If a job does not meet the deadline, it is better to be late by a small amount than a long amount. The tardiness equals zero if the job is completed on time and equals the departure time minus the deadline time if the job is late.

Example 1

A job is received at 3:00, begins processing at 4:00, and is finished at 4:30. The deadline for the job is 4:45. The time in queue is 60 minutes, the service time is 30 minutes, and the tardiness is zero (job completed on time).

Example 2

Another job is received at 3:15, begins processing at 4:30, and is finished at 5:00. The deadline for the job is also 4:45. The time in queue is 75 minutes, the service time is 30 minutes, and the tardiness is 15 minutes (job completed late).

Because many MOPs are random variables, they should be defined more precisely. It is not enough simply to say that the MOP is time in queue. The MOP should be specified as *average* time in queue, *standard deviation* of the time in queue, or *maximum* and *minimum* time in queue. The MOPs might also vary in a predictable (nonrandom) fashion with time of day, day of week, or the like. For example, the waiting time at a toll plaza might be very long every day from 7:00 A.M. to 8:00 A.M. and short the remainder of the day. If this is the case, the MOPs should be calculated for distinct time periods.

In addition to the quantitative measures, such qualitative factors, as the waiting environment should be examined. Is the customer idle while waiting? Do customers receive adequate information about how long they will have to wait? Can customers sit? Is the room crowded? Is there adequate ventilation? All of these factors are important to queue performance.

2.1.2 Server Measures of Performance

The server is concerned with minimizing the cost of providing service and attracting and keeping customers. The cost of providing the service is reflected in the service time and server utilization; attracting and keeping customers is reflected in arrivals and reneging and balking. The MOPs are listed below:

Service time A shorter service time is indicative of a more efficient operation.

Proportional utilization This is the proportion of time that servers are busy serving customers. The proportional utilization is a number between zero and one, and the closer the number is to one the more efficient the operation is.

Throughput The rate at which customers are served. In some systems (such as toll plazas), the throughput increases as queues become longer because customers are better prepared before they reach the server; in others, service deteriorates when queues become long due to customer inattention or server fatigue.

Arrival rate Arrival rate is a measure of the amount of business. A large arrival rate is desirable because it means more revenue.

Reneging proportion Each customer who reneges may translate into lost business. Reneging might be measured as the proportion of customers who renege after joining the queue, or as the proportion who decide not to join the queue after seeing its length. Reneging is the most difficult MOP to measure because the customer who reneges may never be recorded.

Queue length A long queue is more costly to accommodate than a short queue because of the space required.

Qualitative aspects include the following: Do servers have something to do when they are not serving customers? Are servers fully productive? Are there problem servers? Are servers courteous? Are servers ever idle when customers are waiting in the queue?

2.1.3 Relation of Measures of Performance to Organization Goals

It is important to be aware of how the measures of performance fit within the organization's overall goals. If the organization is a private business, that goal is almost certainly to maximize long-term profitability. In government, the goal is more nebulous, but probably has something to do with maximizing the welfare of its constituents. Ideally, when evaluating queueing alternatives, one would like to measure directly the impact of the alternative on the overall goal. Rarely is this possible. Instead, one looks at the measures of performance as indicators of the potential impact on profits and the potential impact on welfare. As illustrated in Fig. 2.1, both time in system and queue length affect the arrival rate and reneging, which in turn affects revenue. The utilization, queue length, and service time affect cost. Combined, cost and revenue determine profitability. The measures of performance are proxies for the overall goal, which, in practice, cannot be directly measured.

2.2 CUMULATIVE ARRIVAL AND DEPARTURE DIAGRAMS

The word *cumulative* means "resulting from accumulation." Thus, a *cumulative diagram* indicates how many customers have accumulated up to some point in time from an initial starting time. A *cumulative arrival diagram* indicates how many customers have arrived

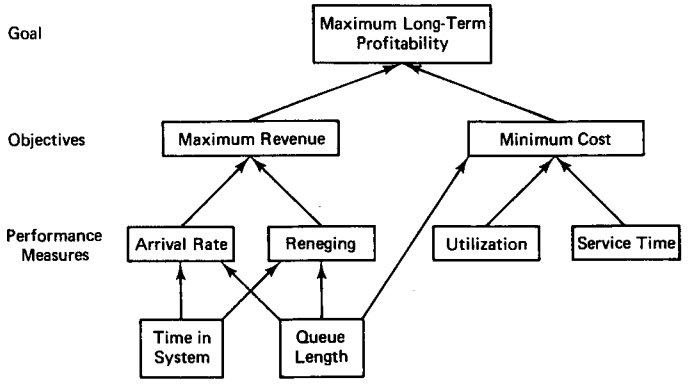


Figure 2.1 Performance measures indicate whether system objectives and goals are achieved.

up to a point in time, and a *cumulative departure diagram* indicates how many customers have departed up to a point in time.

Cumulative diagrams are important because they provide many of the measures of performance in one simple picture. Creating a cumulative diagram is the easiest and simplest way to assess the health of a queueing system. (Remember, though, cumulative diagrams are not the only way to evaluate a queueing system and should not be the sole tool used.)

Definitions 2.2

$A(t)$ = *cumulative arrivals* from time 0 to time t

$D_s(t)$ = *cumulative departures from the system* from time 0 to time t

$D_q(t)$ = *cumulative departures from the queue* from time 0 to time t

The starting time, time 0, can be set at any time that is convenient to the analysis. For example, if a store opens at 9:30 A.M., time 0 would be 9:30 and $A(t)$ would be the number of customers who arrived between 9:30 and time t.

Consider the following data:

Customer	Arrival time	Departure from queue	Departure from system
1	9:36	9:36	9:40
2	9:37	9:40	9:44
3	9:38	9:44	9:48
4	9:40	9:48	9:52
5	9:45	9:52	9:56

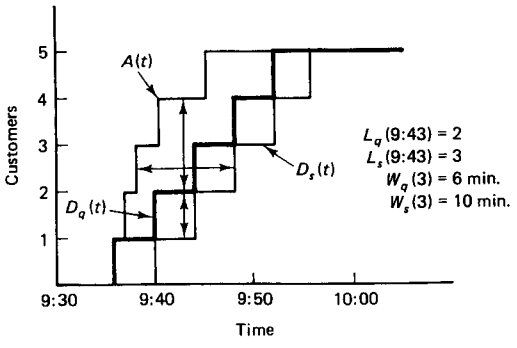


Figure 2.2 Cumulative arrival and departure diagram. Queue lengths are determined by vertical separation between curves. Waiting times are determined by horizontal separation.

The first customer arrives at 9:36, finds no other customer in the system, and goes directly into service. The second customer arrives at 9:37, before the first customer completes service, and joins the queue. The fact that this customer did not enter service immediately indicates that the system has a single server. As soon as customer 1 completes service (9:40), customer 2 departs from the queue and begins service. Each subsequent customer enters service at the time that the previous customer completes service, indicating that the queue operates on a first-come, first-served basis.

The data presented above are plotted in Fig. 2.2 in the form of cumulative arrival and departure diagrams, with time 0 set at 9:30. Each step has height of one and represents one customer. The placement of the step on the horizontal axis corresponds to an "event" (arrival, departure from queue, or departure from the system). The steps in the arrival diagram correspond to the times that customers arrived; the steps in the departure diagram correspond to the times that customers departed.

At any point in time, the height of each curve represents the number of customers that have passed a certain point in the system. At time 9:43, four customers have arrived, two have departed from the queue, and one has departed from the system. One can imagine that these values represent counts made by three observers, stationed at the entrance to the queue, the exit from the queue, and the exit from the system. The vertical distance between the arrival curve and the departure curves indicates how many customers are in the queue and in the system:

Definitions 2.3

$$L_q(t) = \text{number of customers in the queue at time } t$$
$$= A(t) - D_q(t)$$
$$L_s(t) = \text{number of customers in the system at time } t$$
$$= A(t) - D_s(t)$$

At time 9:43, $A(t) = 4$ and $D_q(t) = 2$. Therefore, two customers remain in the queue ($L_q(t) = 2$). At the same time, only one customer so far has left the system ($D_s(t) =$

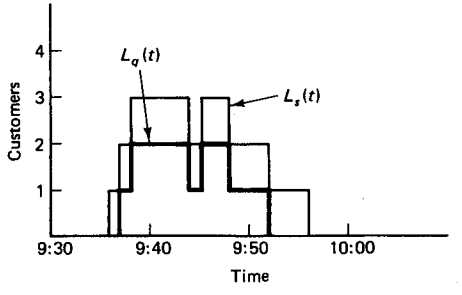


Figure 2.3 Customers in the system and in the queue versus time.

1), so the number of customers in the system is three ($L_s(t) = 3$). Figure 2.2 shows that the queue eventually vanishes at 9:52 and the last customer leaves the system at 9:56. For comparison, Fig. 2.3 plots $L_q(t)$ and $L_s(t)$. Note that Fig. 2.3 provides no information that is not already shown in Fig. 2.2. And, as will be discussed later in the book, the cumulative diagram provides information that cannot be shown on the graphs of $L_q(t)$ and $L_s(t)$.

A customer cannot leave the queue before it arrives, and it cannot leave the system before it leaves the queue. Therefore

$$A(t) \geq D_q(t) \geq D_s(t) \qquad t \geq 0$$

which guarantees that both $L_q(t)$ and $L_s(t)$ are greater than or equal to zero.

2.2.1 Waiting Time

When Fig. 2.2 is read vertically, the queue size and number of customers in the system are identified. Reading Fig. 2.2 horizontally reveals the time in queue and the time in system.

Definitions 2.4

$$A^{-1}(n) = \text{time of the } n\text{th arrival}$$
$$D_q^{-1}(n) = \text{time of the } n\text{th departure from queue}$$
$$D_s^{-1}(n) = \text{time of the } n\text{th departure from system}$$

Whereas $A(t)$, $D_q(t)$, and $D_s(t)$ convert a time into a customer number, $A^{-1}(n)$, $D_q^{-1}(n)$ and $D_s^{-1}(n)$ take a customer number and convert it to a time. They correspond exactly to the data provided before:

n	$A^{-1}(n)$	$D_q^{-1}(n)$	$D_s^{-1}(n)$
1	9:36	9:36	9:40
2	9:37	9:40	9:44
3	9:38	9:44	9:48
4	9:40	9:48	9:52
5	9:45	9:52	9:56

Definitions 2.5

$W_q(n)$ = time in queue, for nth customer to arrive

$W_s(n)$ = time in system, for nth customer to arrive

When the discipline is FCFS, the waiting times, $W_q(n)$ and $W_s(n)$, are found by computing the horizontal distance between the steps in Fig. 2.2:

FCFS Waiting Time

$$W_q(n) = D_q^{-1}(n) - A^{-1}(n) \tag{2.1}$$

$$W_s(n) = D_s^{-1}(n) - A^{-1}(n) \tag{2.2}$$

For example, the steps for customer 3 occur at 9:38, 9:44, and 9:48. Therefore, $W_q(3) =$ 6 minutes and $W_s(3) = 10$ minutes. Equations (2.1) and (2.2) do not apply to non-FCFS disciplines because the customers do not necessarily depart in the same order that they arrive. An alternative representation for non-FCFS systems is presented later in the chapter.

2.2.2 Other Measures of Performance

Server utilization (the proportion of time that the server is busy) is also found on cumulative diagrams. The number of servers that are busy at any time equals $D_q(t) - D_s(t)$, the number of customers that have entered service but have not yet completed service. The ***absolute utilization*** is the average number of busy servers (over time), and the ***proportional utilization*** is the absolute utilization divided by the total number of available servers. In Fig. 2.2, the single server is busy for 20 of the 30 minutes, so both utilizations are 2/3.

The cumulative diagram can also be examined to see whether servers are ever idle when customers are waiting in the queue. When this occurs, $D_q(t) - D_s(t)$ is less than the number of servers, and $A(t) - D_q(t) = L_q(t)$ is greater than zero. This is an obvious sign of inefficiency, for all servers ought to be busy when customers are waiting. Although this situation is not illustrated in Fig. 2.2, it frequently occurs in practice.

Reneging can be presented on a cumulative diagram, but it should be treated separately from ordinary service, as in Fig. 2.4. A major difficulty in creating a reneging diagram is identifying exactly which customers reneged. Reneges do not occur in a strict FCFS order; one customer might remain in the queue 5 seconds before reneging, another 5 minutes. So it is difficult to correlate the time a customer reneged with the time the customer arrived. In Fig. 2.4, four customers are shown to renege between 9:43 and 9:46.

If both an arrival/departure diagram and a reneging diagram are created, the queue size equals the combined number of customers in both diagrams. The waiting time should be based on the arrival/departure diagram alone (this is the waiting time for the customers

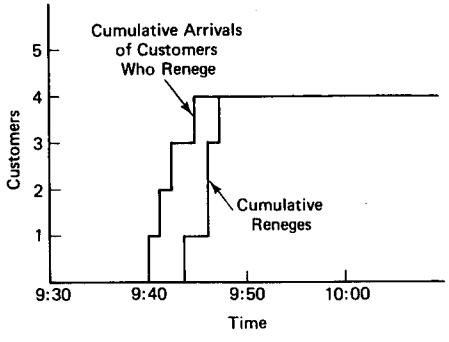


Figure 2.4 Cumulative diagram of reneges.

served). However, in presenting the data, the percentage of customers that reneged, along with their average time in queue, should be noted.

If for some reason reneging was not measured, then the cumulative arrival/departure diagram will contain a discrepancy between the number of customers that arrived and the number that were served, which will lead to overestimation of queue lengths and waiting times. If the discrepancy is small, this problem might be resolved by scaling the departure curve upward by a slight percentage. However, there is no accurate way to resolve a large discrepancy other than measuring the data again by a different method.

2.2.3 Average and Standard Deviation of Waiting Time

The average waiting time per customer equals the total waiting time divided by the total number of customers served. To be most meaningful, it should be computed over a period of time that begins and ends with no customers in the system. If this is not possible, only those customers that arrived before the "oldest" customer in the queue (that is, the customer that has been in the queue the longest time) should be included in the average. Keep in mind that the average waiting time may be different for the customers remaining in the queue, particularly if the discipline is not FCFS.

Definitions 2.6

W_q = average waiting time in queue

$$= \frac{\sum_{n=1}^{N} W_q(n)}{N} \tag{2.3}$$

W_s = average waiting time in system

$$= \frac{\sum_{n=1}^{N} W_s(n)}{N} \tag{2.4}$$

where N is the number of data points in the sample. For our sample data, $W_q = (0 + 3 + 6 + 8 + 7)/5 = 4.8$ minutes, and $W_s = (4 + 7 + 10 + 12 + 11)/5 = 8.8$ minutes. The difference between W_q and W_s is the average service time, 4 minutes.

The standard deviation is an indicator of the variation in data. If the standard deviation is small, then the time in queue and time in system did not vary much between customers; if the standard deviation is large, the variation was large. Because a predictable wait is generally preferable to an unpredictable wait, the standard deviation is an important measure of performance.

Definitions 2.7

$$\sigma_{W_q} = \text{standard deviation of time in queue} = \sqrt{\frac{\sum_{n=1}^{N} [W_q(n) - W_q]^2}{N}} \tag{2.5}$$

$$\sigma_{W_s} = \text{standard deviation of time in system} = \sqrt{\frac{\sum_{n=1}^{N} [W_s(n) - W_s]^2}{N}} \tag{2.6}$$

For our sample problem, $\sigma_{W_q} = \sigma_{W_s} = \sqrt{(4.8^2 + 1.8^2 + 1.2^2 + 3.2^2 + 2.2^2)/5} = 2.9$ minutes. The standard deviations are the same for the problem because the service time is the same for all customers.

It is important to recognize that the average and standard deviation of the waiting time depend on when the data are recorded and on how much data are recorded. Waiting times tend to be highly correlated between successive customers. That is, if a customer has a long wait, the next customer is likely to have a long wait as well. The average and standard deviation are foremost descriptors of what happened in a given time interval. One should not necessarily infer that the values are indicative of what should always occur.

2.2.4 Average and Standard Deviation of Queue Length

One can think of the average waiting time as an average that is computed vertically across Fig. 2.2. The number in the denominator of Eqs. (2.3) and (2.4) is the number of customers that arrived. The average queue length is computed horizontally, and the number in the denominator is an elapsed time. Hence, the average queue length is interpreted as the average number of customers in the queue over some interval of time. Suppose that a = start time of interval and b = end time of interval. Then

Definition 2.8

$$L_q = \text{average queue length (customers)} = \frac{\int_a^b L_q(t)dt}{b - a} \tag{2.7}$$

The numerator is defined by an integral rather than a summation. The reason for this is simple. Whereas customers are discrete entities, time is continuous. An integral is the continuous analog of a summation.

The integral of $L_q(t)$ equals the area between $A(t)$ and $D_q(t)$ over the interval from a to b. The integral does not have to be calculated in the customary form of an equation. It can be calculated geometrically by determining the length of time that one customer is in the queue, the length of time that two customers are in the queue, and so on. (The appendix at the end of this chapter expands on this concept.) For our sample data:

Time interval	$L_q(t)$
9:30 – 9:37	0
9:37 – 9:38	1
9:38 – 9:44	2
9:44 – 9:45	1
9:45 – 9:48	2
9:48 – 9:52	1
9:52 –10:00	0

$L_q(t)$ equals zero for 15 minutes, one for 6 minutes, and two for 9 minutes. Therefore, the integral of $L_q(t)$ equals $1 \times 6 + 2 \times 9 = 24$ *customer-minutes*. The average queue length over the half-hour interval is 24 customer-minutes/30 minutes = .8 customer. Whereas waiting time is measured in terms of time (minutes, hours), queue length is measured in terms of customers.

The standard deviation of the queue length, like the average queue length, is defined over a time interval:

Definition 2.9

$$\sigma_{L_q} = \text{standard deviation of queue length} = \sqrt{\frac{\int_a^b [L_q(t) - L_q]^2 dt}{b - a}} \tag{2.8}$$

For our sample data, the numerator of Eq. (2.8) is $(.8^2 \times 15 + .2^2 \times 6 + 1.2^2 \times 9) = 22.8$ customer^2-minutes. Dividing through by 30 minutes and taking the square root, the standard deviation of $L_q(t)$ is .87 customer.

The average and standard deviation of the number of customers in the system are calculated in similar fashion:

Definitions 2.10

$$L_s = \text{average number of customers in the system} = \frac{\int_a^b L_s(t)dt}{b - a} \tag{2.9}$$

σ_{L_s} = standard deviation of customers in the system

$$= \sqrt{\frac{\int_a^b [L_s(t) - L_s]^2 dt}{b - a}} \tag{2.10}$$

For our sample problem, the integral in Eq. (2.9) equals 44 customer-minutes, so the average number of customers in the system is 44/30 = 1.47 customers. The standard deviation is 1.23 customers. Even though the service time is the same for all customers, the standard deviation of $L_s(t)$ does not equal the standard deviation of $L_q(t)$.

2.3 LITTLE'S FORMULA

In the calculations for average queue length, a coincidence occurred. The numerator of the equation for L_q equaled the numerator of the equation for W_q, and the numerator of the equation for L_s equaled the numerator of the equation for W_s. Actually, this is no coincidence at all. It is a simple implication of one of the most important results of queueing theory: Little's formula (Little 1961).

So why does the following occur?

$$\sum_{n=1}^{N} W_q(n) = \int_a^b L_q(t)dt \tag{2.11}$$

Both the left and right sides of Eq. (2.11) equal the area between the curves $A(t)$ and $D_q(t)$, *whether or not* the discipline is FCFS (an exception is mentioned below). The left side calculates the area with a vertical summation and the right side calculates the area with a horizontal integral (Fig. 2.5). The two methods measure the exact same quantity in two different ways and must, therefore, arrive at the same result. This area can be interpreted as the total waiting time for all customers:

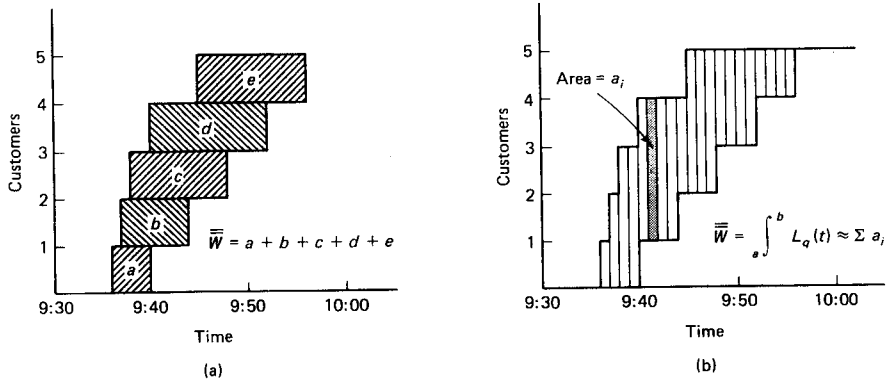


Figure 2.5 Total waiting time can be calculated by (a) summing individual waiting times or (b) summing queue lengths over short time intervals.

Definition 2.11

$\bar{\bar{W}}_q$ = total time in queue for all customers

$\bar{\bar{W}}_q$ has the dimension customers × time. For example, $\bar{\bar{W}}_q$ might equal 100 customer-minutes, meaning that 100 customers might have waited 1 minute each, or 50 customers might have waited 2 minutes each, or any combination whose product is 100 customer-minutes.

One caveat: The areas are only equal if the system begins and ends the time interval with no customers in the system. Otherwise, a few customers will be left dangling and there will be a slight discrepancy in the calculations.

Little's formula, which is quite general and applies to any queue discipline, specifies how to convert the average waiting time into the average time in queue, and vice versa. Based on the definition of $\bar{\bar{W}}_q$, the average time in queue and average customers in queue can be defined as follows:

$$W_q = \frac{\bar{\bar{W}}_q}{N} \qquad L_q = \frac{\bar{\bar{W}}_q}{b - a} \tag{2.12}$$

Combining these two equations provides the following:

$$NW_q = (b - a)L_q \tag{2.13}$$

Definition 2.12

λ = average arrival rate over an interval $[a,b]$

$= N/(b - a)$

The arrival rate, λ, is interpreted as the average number of customers that arrive per unit time. Substitution of λ in Eq. (2.13) yields

Little's Formula I

$$L_q = \lambda W_q \tag{2.14}$$

The same result applies to time in system:

Little's Formula II

$$L_s = \lambda W_s \tag{2.15}$$

The only restriction on Little's formula is that the time interval begins and ends with no customers in the system. However, even if there are customers in the system, Little's formula is approximately correct. It is correct even if there is reneging. However, the waiting time must account for the waiting time of customers who left the system without receiving service.

Returning to our sample problem, recall that W_q = 4.8 minutes and that λ = 5/30 = .167 customer-minute. Thus, L_q must be .167 × 4.8 = .80 customer, the value

calculated earlier. Also recall that $W_s = 8.8$ minutes, so L_s must be $.167 \times 8.8 = 1.47$ customers, also the same value calculated previously.

Little's formula is particularly significant with respect to data measurement. Because average waiting time can be calculated from average queue length, and vice versa, one does not have to measure both queue lengths and waiting times to estimate the averages of both; only one of the two must be measured.

Example

A large machine shop has records on the number of jobs in the work queue, by hourly interval, for the month of March. It also has records on the number of jobs submitted each month. A queueing analyst calculated the average queue length to be 16 jobs. The number of jobs submitted during the month was 253. From Little's formula, average waiting time in queue is approximately 16/253 = .063 month (1.9 days).

No simple relation holds between the standard deviation of waiting time and the standard deviation of queue length. Depending on the queue discipline, it is possible to have a very small deviation in queue length yet at the same time have very large deviations in waiting time. This is especially true if the discipline is not first-come, first-served.

2.4 PARALLEL SERVERS

Cumulative diagrams can be used when servers work in parallel. Consider the following data:

n	$A^{-1}(n)$	$D_q^{-1}(n)$	$D_s^{-1}(n)$
1	9:36	9:36	9:40
2	9:37	9:37	9:41
3	9:38	9:40	9:44
4	9:40	9:41	9:45
5	9:45	9:45	9:49

The arrival data are the same as before, but the departure data are different. There are now two servers working in parallel. Customer 2 is immediately served by the second server, and customer 3 begins service as soon as customer 1 (not customer 2) completes service. Figure 2.6 is the cumulative diagram. The separation between $D_q(t)$ and $D_s(t)$ now equals two (the number of servers) from 9:37 to 9:44. Between 9:36 and 9:37 and between 9:44 and 9:49 only one of the two servers is busy and the separation equals one. In general, cumulative diagrams can be used for any number of parallel servers and, as will be shown in later chapters, any number of serial servers.

Time in queue, queue length, customers in the system, and utilization are calculated in the exact same manner as for a single server. Time in system is also calculated in the same way, with an additional caveat. It is possible for a parallel server queue to follow a first-come, first-*served* discipline without following a first-come, first-*out* (FCFO) discipline. For example, the first customer might begin service at 2:00, ahead of the second customer, which begins at 2:02. But because the service time is longer for customer 1 than customer 2, the first does not complete service until 2:10 while the second finishes at 2:07. Thus, the discipline is FCFS but not FCFO. Time in system only equal $D_s^{-1}(n) - A^{-1}(n)$ if the discipline is FCFO.

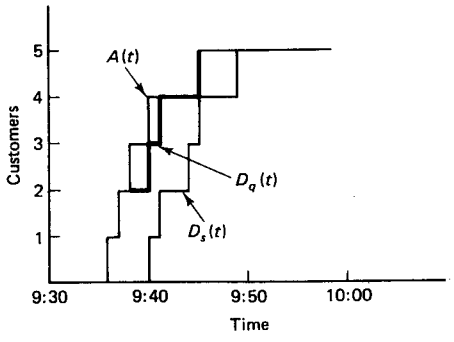


Figure 2.6 Cumulative arrival and departure diagram with two parallel servers.

2.5 REPRESENTATION OF NON-FCFS SYSTEMS

As already noted, caution must be exercised in interpreting the cumulative graphs when the discipline is not FCFS (or not FCFO for parallel servers), because the nth customer to arrive is not necessarily the nth customer to depart. Non-FCFS disciplines present no problem with respect to measuring queue lengths. They also present no problem with respect to measuring average waiting times. But they do present problems with respect to measuring waiting times of individual customers and calculating the waiting time standard deviation.

Consider the following data:

n	$A^{-1}(n)$	Departure from Queue	Departure from System
1	9:36	9:36	9:40
2	9:37	9:52	9:56
3	9:38	9:44	9:48
4	9:40	9:40	9:44
5	9:45	9:48	9:52

The arrival data are the same as before, but customers are served according to the last-come, first-served discipline. When the first customer arrives, no one is in the system, and the customer begins service immediately. When customer 1 completes service, customers 2, 3, and 4 are in the queue, and customer 4 is served first because it arrived last.

Figure 2.7 is a ***Gantt chart*** representation of the data. One bar represents each

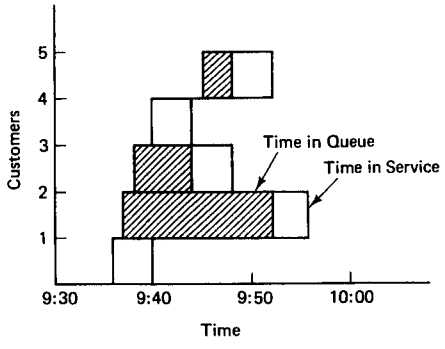


Figure 2.7 Gantt chart showing last-come, first-served (LCFS) discipline.

customer, and the bar is divided into two segments, representing time in queue and time in service. The entire length of the bar is time in system. By comparing with Fig. 2.2, we see that the LCFS discipline allows customers 4 to 5 to leave earlier, forces customer 2 to leave much later, and has no impact on customers 1 and 3. The average time in queue (and time in system) for LCFS is no different from FCFS (4.8 minutes and 8.8 minutes). However, the standard deviation is now much larger (5.9 minutes). Thus, the major impact of the LCFS discipline is an increase in the *variation* of the waiting times, not an increase in the average waiting time.

Gantt charts are also useful when customers have deadlines, whether or not the discipline is FCFS. Vertical lines indicating customer deadlines can be compared against the times service was actually completed. The separation between the deadline and the departure time from the system reveals the lateness.

Note that the Gantt chart is not a good way to show queue lengths. If this is the purpose, then the cumulative arrival and departure curves should be used (independent of the discipline).

2.6 EMPIRICAL PROBABILITY DISTRIBUTIONS

We have already seen that it is important to calculate the average and standard deviation of the queue length and waiting time. These statistics reduce a large set of data into a few measures of performance. Sometimes probabilistic data should be presented in greater detail, in the form of a diagram.

A probability distribution function represents the probability that a random variable occurs at or below a set value. For example, suppose that a coin is flipped five times. The total number of heads recorded has a binomial distribution with parameters $n = 5$ (number of flips) and $p = .5$ (probability of heads). The binomial distribution function indicates the probability of observing no heads, the probability of having one or fewer heads, the probability of having two or fewer heads, and so on, in the sample of five flips.

Unlike flipping coins, there may be no reason to believe that the data from a queueing system conform to any of the familiar theoretical distributions. This does not

mean that there is no explanation for the underlying process generating the data, only that the process is something with which we are not familiar.

Whether or not the process conforms to a theoretical distribution, the observations can be displayed as an ***empirical distribution***. The empirical (*empirical* means "relying on observation") distribution indicates the proportion of data points *in a sample* that falls at or below a set value. The empirical probability distribution is drawn in much the same manner that a cumulative arrival and departure diagram is drawn (even though the cumulative diagrams are not directly concerned with probability).

Suppose that the following data are the service times (in seconds) recorded for 20 consecutive customers:

69	55	49	75	36	50	57	77	99	11
94	86	30	60	33	14	75	79	23	3

The first step in creating the empirical distribution is to sort the data in ascending order:

3	11	14	23	30	33	36	49	50	55
57	60	69	75	75	77	79	86	94	99

The second step is to plot the data in ascending order as a step curve. Each step has a height of one and is placed on the horizontal axis at a point corresponding to one of the service times. The graph is completed when the vertical axis is rescaled from zero to one, as in Fig. 2.8. The empirical distribution, $F(x)$, indicates the proportion of data points that have values at or below the value x. For example, 60 percent of the data points are less than or equal to 65.

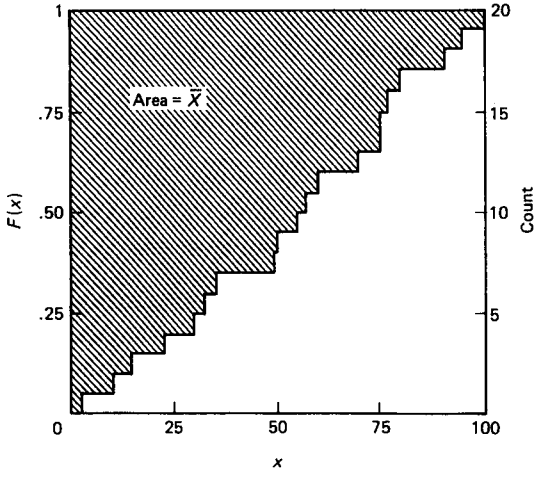


Figure 2.8 Empirical probability distribution showing calculation of mean.

If all of the data points are greater than or equal to zero, the cumulative diagram can also be used to find the average of the data points. This average is defined by

$$\bar{X} = \int_0^\infty [1 - F(x)]dx \qquad F(0) = 0 \tag{2.16}$$

That is, the average equals the area between $F(x)$ and the horizontal line 1. (Equation (2.16) also applies to theoretical distribution functions.)

2.7 MEASUREMENT TECHNIQUES

A fundamental principle in performing any type of experiment is that the measurement technique should not interfere with the process observed. Unfortunately, this goal cannot always be achieved. From the Heisenberg uncertainty principle, we know that it is impossible to measure both the position and motion of a wave with unlimited accuracy. Though no hard-and-fast rules apply to queueing systems, there are times when something cannot be measured without disturbing the system. Sometimes the mere presence of an observer will change customer behavior. The only exceptions seem to be automated systems, such as time-shared computers and telephone information lines. These systems can continuously monitor and record queueing information and produce reports, as shown in Fig. 2.9.

The task of determining cumulative arrivals and departures can be accomplished in any of several ways:

1. Record the time that each customer arrives, departs from queue, and departs from system.

2. Record the time that each customer departs from the system. Periodically count the number of customers in the queue and the number of customers in service.

3. Record the time that each customer departs from the system. Periodically time how long a customer spends in queue and spends in service.

2.7.1 Arrival and Departure Data

Departures from the system are the easiest to measure. The server can record departures with a counting device (Fig. 2.10) or perhaps directly from a cash register. Arrivals are usually much more difficult to measure.

2.7.1.1 Arrivals. Generally, measuring arrivals requires the added cost of stationing an observer (either human or an automated counting device) at a point beyond the end of the queue. Difficulties arise if the "free-flow" time for the customer to travel from the measurement point to the server is long. Vehicle queues at toll plazas sometimes stretch a mile or more, meaning that the vehicle arrives at the end of the queue at least a minute before it would have arrived at the server if there had been no queue. If a vehicle spends 5 minutes in the queue, eliminating the delay can at most reduce waiting time by 4 minutes—the 1-minute travel time cannot be eliminated. This 1-minute difference should

ANSWERED-CALL PROFILE
TRUNK GROUP WATS
DECEMBER 14, 1985

TIME OF DAY	NO. OF CALLS ANSW'D	% OF CALLS ANSW'D WITHIN X SECONDS 90	180	270	360	450	540	AVG. TIME-BEFORE-ANSWER (SECONDS)	0 90 180 270 360 450 540
07:30-08:00	18	72	83	100	100	100	100	75	+xx
08:00-08:30	74	14	31	62	66	88	100	258	+xxxxxxxxxxxx
08:30-09:00	86	0	1	44	57	80	98	326	+xxxxxxxxxxxxxxxx
09:00-09:30	116	0	23	89	100	100	100	214	+xxxxxxxxxxx
09:30-10:00	103	0	32	68	94	100	100	228	+xxxxxxxxxxxx
10:00-10:30	62	0	23	31	47	65	87	360	+xxxxxxxxxxxxxxxxx
10:30-11:00	136	4	65	71	91	93	97	200	+xxxxxxxxxx
11:00-11:30	110	13	68	96	100	100	100	150	+xxxxxxx
11:30-12:00	125	18	66	82	95	100	100	171	+xxxxxxxx
12:00-12:30	112	16	94	100	100	100	100	123	+xxxxxx
12:30-13:00	102	14	40	72	100	100	100	198	+xxxxxxxxx
13:00-13:30	71	0	0	4	37	99	100	367	+xxxxxxxxxxxxxxxxx
13:30-14:00	75	0	0	24	53	64	79	387	+xxxxxxxxxxxxxxxxx
14:00-14:30	116	0	34	67	88	100	100	232	+xxxxxxxxxx
14:30-15:00	92	0	14	53	98	100	100	261	+xxxxxxxxxxx
15:00-15:30	82	0	0	23	71	89	100	328	+xxxxxxxxxxxxx
15:30-16:00	135	19	70	92	99	100	100	150	+xxxxxxx
16:00-16:30	95	5	25	58	100	100	100	238	+xxxxxxxxx
16:30-17:00	81	0	0	21	81	98	100	311	+xxxxxxxxxxxx
17:00-17:30	12	50	67	75	100	100	100	120	+xxxxx
TOTAL	1803	7	38	64	86	95	98	233	

(a)

Figure 2.9 Output generated by call sequencing system.
Source: Reprinted by permission of IBM Corporation (Rolm Systems).

SEQUENCE # 53782

ABANDONED-CALL PROFILE
TRUNK GROUP WATS
DECEMBER 13, 1985

TIME OF DAY	NO. OF CALLS ABAN'D	% OF CALLS ABAN'D WITHIN X SECONDS 10	20	30	40	50	60	AVERAGE TIME-BEFORE-ABANDONMENT (SECONDS) 0 10 20 30 40 50 60
07:30-08:00	4	25	50	75	75	100	100	21 +xxxxxxxxxx
08:00-08:30	14	43	57	57	57	64	79	29 +xxxxxxxxxxxxx
08:30-09:00	10	20	70	90	90	100	100	18 +xxxxxxxx
09:00-09:30	18	56	67	83	89	100	100	15 +xxxxxxx
09:30-10:00	24	50	67	75	83	92	100	17 +xxxxxxxx
10:00-10:30	24	13	29	50	67	83	96	30 +xxxxxxxxxxxxxx
10:30-11:00	12	67	83	83	83	83	92	16 +xxxxxxxx
11:00-11:30	22	41	91	100	100	100	100	14 +xxxxxxx
11:30-12:00	19	53	84	95	95	95	95	15 +xxxxxxxx
TOTAL	147	41	67	78	83	90	96	19

(b)

Figure 2.9 (*continued*)

be compensated for when drawing the cumulative arrival diagram, as in Fig. 2.11. (The large arrival rate has smoothed out the steps in the arrival and departure curves.)

As another example, a patient visiting a medical clinic might have to check in at a receptionist before waiting to see a doctor. Depending on the line for the receptionist, the patient might have several minutes of delay before his or her arrival is recorded. These minutes may never be accounted for in measuring how long patients wait to see doctors.

Another problem when the discipline is not FCFS is correlating a specific arrival with a specific departure. A counting device does not indicate which customer passed a given point, only that some customer passed the point. To measure variation in waiting time, one must be able to identify exactly which customer passed each point. When the arrival rate is large, this task is tedious at best. If vehicles are involved, the license plate can be recorded. Time-lapse photography can be used when people are being counted.

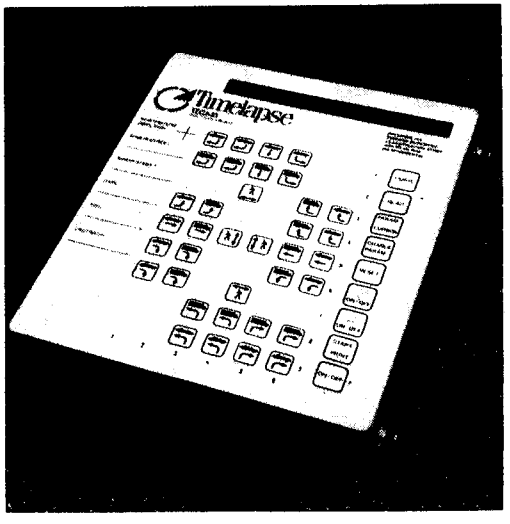


Figure 2.10 Device for counting customers.
Source: Courtesy of Timelapse, Inc., St. Petersburg, Florida.

Measuring arrivals is not as difficult when the customers are work orders or things rather than people. Computer systems routinely record the times that jobs are submitted, begin processing, and complete processing. For packages and objects, bar codes and automated reading devices can greatly simplify the task of recording arrival and departure data. Even handwritten work slips might contain the necessary information.

2.7.1.2 Departures from Queue. Departures from queue can be derived from the data on departures from the system and arrivals. When all servers are busy, one departure from queue coincides with one departure from system. When servers are idle, a

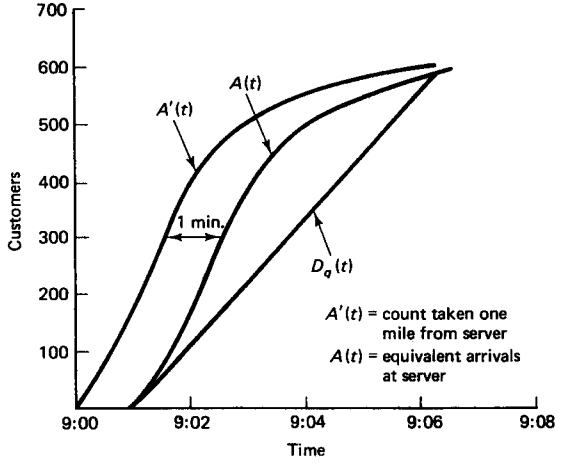


Figure 2.11 When customers are counted upstream from the server, the cumulative count should be transformed to obtain the arrival curve.

departure from queue coincides with an arrival. This at least is how a queue performs in theory. In reality, there is often a lag between the time a customer completes service and the time the next customer begins service. This lag may represent two things: a reaction time and a movement time.

Example

George Waites arrived at his bank at 9:00, at which time he immediately joined a teller queue. Because the line was long, George began reading his newspaper. At 9:03, George reached the front of the queue, and at 9:04 a customer completed service. By this time, George was enraptured in the comics and did not notice that a teller was free. Five seconds later, seeing that George did not move, the teller announced, "Next in line." But George did not hear the teller. Another 5 seconds later, the customer behind George poked him in the back, and George realized a teller was free. It then took George another 4 seconds to walk up to the teller. George completed service 30 seconds later.

For the purposes of analysis, George departed from the queue at 9:04, even though he did not reach the teller until 14 seconds later. George departed from the system at 9:04.44 and the service time is 44 seconds, the total time devoted by the server to George. Of the 44 seconds, reaction took 10 seconds and movement took 4 seconds.

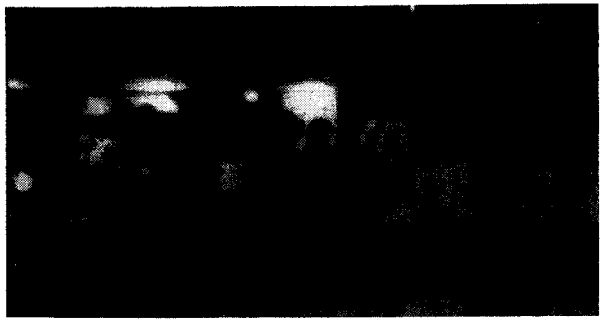
When a queue exists, a departure from queue occurs when a server becomes available, not when the customer realizes the server is available. Depending on the purpose of the analysis, one may wish to record the reaction and movement times in addition to the total service time. Such information may be important if changes in the queue's design are considered (see Chap. 11). Keep in mind that the reaction time and movement time may be different when a queue exists than when a queue does not exist. This may lead to slightly different service times.

2.7.1.3 Busy Systems. For systems such as highways, telephone lines, and large cafeterias, the number of customers served can be so large that recording each customer's arrival and departure time is unwieldy. Interval counts should be used instead. An interval count would tell how many customers arrived and departed in each time interval (each minute, for example). This simplification is especially important when arrivals and departures are recorded manually, because it is impossible to write down the time of each event. In making these counts, one may wish to use a metronome to emit a sound at the end of each time interval.

2.7.2 Queue Lengths

Provided that there is no reneging, the cumulative departures from the queue equal the cumulative departures from the system plus the number of customers in service; the cumulative arrivals equal the cumulative departures from the system plus the number of customers in the system. Therefore, measurement of departures from the system, customers in the system, and customers in service provides the necessary information.

Queue lengths must be measured periodically, at time intervals that are sufficiently short to detect any major changes. In most cases, some type of recording device is needed to count customers. For example, the queue might be photographed periodically and



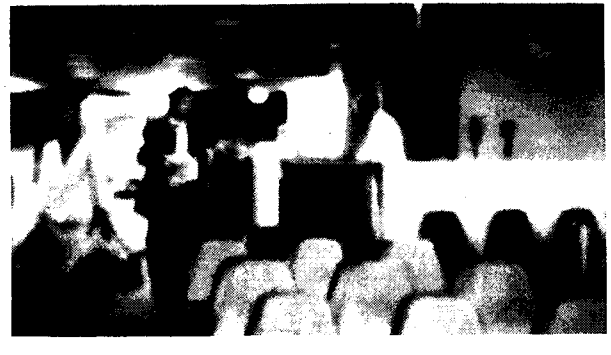
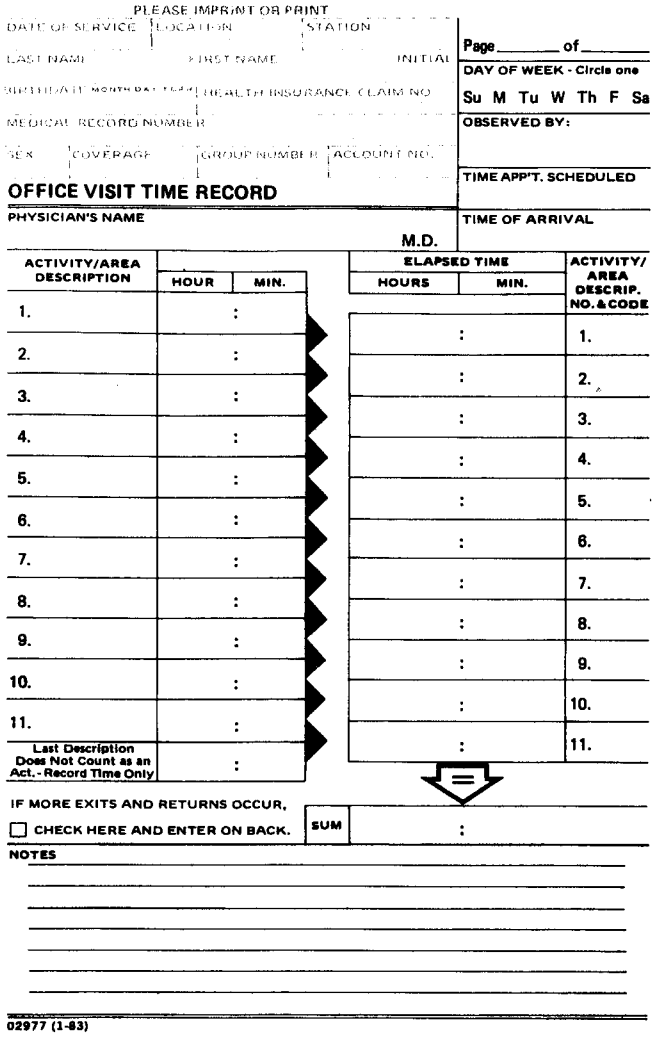


Figure 2.12 Time-lapse photography records the formation of a queue.

counted afterward (Fig. 2.12). Or markers might be positioned according to different queue lengths so that an observer would only have to note which marker best approximates the queue length.

2.7.3 Waiting Times

The cumulative arrival diagram and the cumulative departure from queue diagram can also be developed by periodically measuring customer waiting times. The time in service



PLEASE IMPRINT OR PRINT

DATE OF SERVICE | LOCATION | STATION

LAST NAME | FIRST NAME | INITIAL

BIRTH DATE: MONTH DAY YEAR | HEALTH INSURANCE CLAIM NO.

MEDICAL RECORD NUMBER

SEX | COVERAGE | GROUP NUMBER | ACCOUNT NO.

Page _____ of _____

DAY OF WEEK - Circle one

Su M Tu W Th F Sa

OBSERVED BY:

TIME APP'T. SCHEDULED

OFFICE VISIT TIME RECORD

PHYSICIAN'S NAME

M.D.

TIME OF ARRIVAL

ACTIVITY/AREA DESCRIPTION	HOUR	MIN.	ELAPSED TIME HOURS	ELAPSED TIME MIN.	ACTIVITY/ AREA DESCRIP. NO. & CODE
1.	:		:		1.
2.	:		:		2.
3.	:		:		3.
4.	:		:		4.
5.	:		:		5.
6.	:		:		6.
7.	:		:		7.
8.	:		:		8.
9.	:		:		9.
10.	:		:		10.
11.	:		:		11.
Last Description Does Not Count as an Act. - Record Time Only	:				

IF MORE EXITS AND RETURNS OCCUR,

☐ CHECK HERE AND ENTER ON BACK.

SUM :

NOTES

02977 (1-83)

Figure 2.13 Sample sheet for recording customer data.
Source: Courtesy of Kaiser Permanente, Oakland, California.

is subtracted from departures from the system to obtain departures from the queue; time in system is subtracted from departures from the system to obtain arrivals.

Waiting times are measured by sending test customers through the system. This technique is used in communication systems with test data and sometimes on road networks with test vehicles. In addition to being potentially expensive, one drawback of

this approach is that the test interferes with the queue. The added customers increase overall delay.

Another approach is to observe real customers and record their departure and service times. Figure 2.13 is a sample record sheet that can be used when customers are processed at several different servers.

2.7.4 Reneging

Unless the queueing system is totally automated, as in an airline telephone reservation line, reneging is bound to be difficult to measure.

Example

Diners place their names on a waiting list as soon as they arrive at a busy restaurant. As tables becomes available, the maître d'hôtel calls off names in the order that customers arrived. Some of the diners have reneged before their names are called.

The waiting list allows the restaurant to determine when customers arrive and when they depart from queue. It also allows them to determine which customers reneged. It does not tell the restaurant when the customers reneged.

Example

An observer counts customers as they join a ticket queue for a popular movie, and the cashier records ticket sales. Before they reach the front of the line, some of the customers decide that the wait is too long and leave to attend another movie.

The system here provides arrival and departure data. But because customers renege from the middle of the queue (not the end or the beginning), they may not be detected until the queue finally vanishes.

Example

A company uses work tickets to keep track of customer orders. The times the order is received, begun, completed, and delivered are recorded. Work tickets are discarded if an order is canceled.

Again, the system provides data on arrivals and departures but not on reneges.

Unfortunately, measuring reneging often requires interfering with the queue's operation. For example, customers might be required to sign out before departing. Or barriers might be erected that prevent customers from leaving the middle of the line. For the work orders, a system must be instituted to collect data that the company would otherwise prefer to discard.

2.7.5 Detailed Service Data

Often large queues are the product of inefficient servers, and the best way to eliminate the queue is not to add more servers but to encourage the existing servers to work more productively. The only way to tackle this problem is first to observe the servers in action,

but this in itself raises a host of problems, especially if the server is a human. It is a fair assumption that workers do not like to be watched, timed, or measured in any way, nor are they likely to behave in the same manner when they are observed as when they are not observed. Needless to say, the task is difficult. Entire books are written on the subjects of time-motion study and work methods, and it would not be appropriate to cover this topic in any detail here. Refer to the books listed at the end of this chapter.

2.8 CHAPTER SUMMARY

Queueing systems should be examined for symptoms of health and illness. These symptoms are the measures of performance (MOP), the most important of which are:

1. Time in system and time in queue
2. Customers in queue and in service
3. Service time
4. Utilization
5. Arrival rate
6. Reneging

A poor score on any one of these MOPs indicates that something is wrong, and further examination is needed.

The measures of performance should be presented through cumulative diagrams and Gantt charts, which show how the queues evolve over time. Average waiting time and average queue length should also be calculated. From Little's formula, average waiting time can be converted into average queue length with a simple calculation. The standard deviation of the waiting time and queue length provide an indication of how much these two factors vary between customers and over time, respectively. Finally, empirical probability distributions for the service time and interarrival time provide a more detailed picture of how the factors vary.

Data measurement may involve recording arrivals, departures, queue lengths, and/or waiting times. Counters, photographs, and test customers are some of the tools used in data measurement. Particular caution should be exercised when measuring arrivals and reneging. No matter what tool is used, the measurement technique should be as unobtrusive as possible.

Though data are important in themselves, they are no substitute for observing the system with your own eyes. A simple record of what happened does not necessarily reveal *why* it happened. Explanations for unusually long service times (server attitude, mechanical breakdown, problem customer) may never appear in the data. Although quantitative data are needed in evaluating the merits of an alternative, qualitative information is essential to defining and isolating the problem. It is also essential in creating models, discussion of which begins in the following chapter.

FURTHER READING

BARNES, R. M. 1980. *Motion and Time Study, Design and Measurement of Work*. New York: John Wiley.

BROWNLEE, K. A. 1965. *Statistical Theory and Methodology in Science and Engineering*. New York: John Wiley.

FREEDMAN, D., R. PISANI, and R. PURVES. 1978. *Statistics*. New York: W. W. Norton.

LITTLE, J. D. C. 1961. "A Proof for the Queueing Formula: $L = \lambda W$," *Operations Research*, 9, 383–387.

MAXWELL, E. A. 1983. *Introduction to Statistical Thinking*. Englewood Cliffs, N.J.: Prentice Hall.

MUNDEL, M. E. 1985. *Motion and Time Study, Improving Productivity*. Englewood Cliffs, N.J.: Prentice Hall.

NEWELL, G. F. 1982. *Applications of Queueing Theory*. London: Chapman and Hall.

NIEBEL, B. W. 1982. *Motion and Time Study*. Homewood, Ill.: Richard D. Irwin.

WONNACOTT, T. H., and R. J. WONNACOTT. 1977. *Introductory Statistics*. New York: John Wiley.

PROBLEMS

1. The data below represent jobs that arrived at a printing shop between the opening time, 9:00 A.M., and 12:00 noon. Calculate (a) average and standard deviation of time in queue; (b) average and standard deviation of time in system; and (c) average and standard deviation of tardiness.

Customer	Arrival time	Depart queue	Depart system	Due
1	9:03	9:03	9:23	10:00
2	9:10	9:23	9:55	10:00
3	9:40	9:55	10:35	11:00
4	10:05	10:35	10:50	11:00
5	10:15	10:50	11:20	11:00
6	10:40	11:20	11:35	12:00
7	11:50	11:50	12:20	1:00

2. For the data in Prob. 1, plot the cumulative arrivals, cumulative departures from queue, and cumulative departures from system. For the time interval from 9:00 to 12:00, determine (a) average and standard deviation of customers in queue and (b) average and standard deviation of customers in system.
3. From your answers to Probs. 1 and 2, verify Little's formula for customers in queue. Also, check Little's formula for customers in system. Explain why the formula is not verified for customers in system and give a remedy.

4. Repeat Probs. 1–3 for the data set below.

Customer	Arrival time	Depart queue	Depart system	Due
1	9:05	9:05	9:30	12:00
2	9:10	9:30	10:05	12:00
3	9:20	10:05	10:35	12:00
4	10:35	10:35	11:10	12:00
5	11:20	11:20	11:45	12:00
6	11:40	11:45	12:15	12:00

5. Determine the throughput and utilization for the data in Prob. 1 over the 9:00–12:00 period.

*** 6.** The following are arrival and service times for work that was in a small machine shop during January. The system has parallel servers.

Customer	Arrival time	Depart queue	Depart system
1*			Jan. 4
2*			Jan. 6
3	Jan. 1	Jan. 1	Jan. 15
4	Jan. 3	Jan. 4	Jan. 14
5	Jan. 7	Jan. 7	Jan. 18
6	Jan. 12	Jan. 14	Jan. 29
7	Jan. 16	Jan. 16	Feb. 2
8	Jan. 16	Jan. 18	Jan. 27
9	Jan. 22	Jan. 27	Feb. 7
10	Jan. 24	Jan. 29	Feb. 20
11	Jan. 26	Feb. 2	Feb. 15

*Arrived in December.

(a) How many servers are there?
(b) Plot cumulative arrivals, cumulative departures from queue, and cumulative departures from system.
(c) Customers are served FCFS, but the system is not first-in, first-out. Explain why.
(d) Among jobs that arrived in January, calculate average time in queue and average time in system. From Little's formula, estimate average customers in queue and average customers in system.
(e) Will your estimate for average customers in queue from part d equal the average customers in queue over January? Why or why not?

7. The printer in Prob. 1 has decided to shift to an LCFS discipline. As before, one server operates. Assume that arrival and service times are the same as in Prob. 1.
(a) Draw a Gantt diagram for the customers, indicating arrival time, departure time, and due date.
(b) Plot cumulative arrivals, cumulative departures from system, and cumulative departures from queue.
(c) Calculate average time in system and the standard deviation of time in system.

8. The printer in Prob. 1 would now like to serve customers with the shortest service time first. Using this discipline, repeat parts a–c in Prob. 7. Can you see any advantages or disadvantages of this discipline over LCFS? Over FCFS?

*Difficult problem

9. These service times were recorded at a fast-food counter (in seconds):

9 57 63 30 18 7 76 97 53 59 37 54 88 39 94

(a) Draw the empirical probability distribution function.
(b) Using Eq. (2.16), along with one of the methods in the appendix, calculate the average service time.
(c) Confirm your answer to part b by summing the service times and dividing by 15.

***10.** The following data set includes some reneges. Plot the cumulative arrival/departure diagram, along with a cumulative reneging diagram. Determine all meaningful measures of performance.

Customer	Arrival time	Depart queue	Depart system	Renege time
1	9:05	9:05	9:30	
2	9:10	9:30	10:05	
3	9:20	10:05	10:35	
4	9:25			9:25
5	9:40			9:50
6	9:49			9:50
7	10:35	10:35	11:10	
8	11:20	11:20	11:45	
9	11:40	11:45	12:15	

11. Suppose that the data in Prob. 1 contain a reaction/movement time in addition to a service time. If the reaction/movement time is 30 seconds when a customer is already in the system and zero otherwise, what is the average time that the customer is actually being served (not counting the reaction/movement time)?

12. A telephone system began the day with no calls being served. The data below give the number of calls received and departed in each of the next 8 minutes.

	1	2	3	4	5	6	7	8
Arrived	10	15	12	8	16	15	11	10
Departed queue	10	13	12	10	13	13	12	13
Departed system	3	13	12	12	11	13	12	13

(a) Draw the cumulative diagrams and derive estimates for average time in queue and in system.
(b) Can anything be said about the standard deviation of these values? Explain.
(c) How many servers are operating?

13. Describe three ways for estimating the expected time in queue for the following queueing systems. In each case, describe what values you would measure and how you would measure them.
(a) Patients in a doctor's office
(b) Customers at a delicatessen
(c) Computer users on a time-shared system
(d) Autos at a traffic signal
(e) Checks waiting to be cleared by a bank

14. How would you modify your answers to Prob. 13 to estimate expected queue length?

15. How would you modify your answers to Prob. 13 to estimate standard deviation of time in queue?

*Difficult problem

16. How would you modify your answers to Prob. 13 to estimate reneges/hour?
17. Before patrons can check out books at a library, they must fill out a charge card for each book. Unfortunately, some of the patrons forget to complete the cards before reaching the clerk. In these cases, the clerk asks the patron to step aside and fill out the card, after which the patron receives priority over other customers waiting in line. Describe how you would measure the performance of this queueing system. Also, discuss how you might modify your cumulative diagram and Gantt chart.

EXERCISE: QUEUE OBSERVATION AND MEASUREMENT

The purpose of this exercise is to observe and record the behavior of a queueing system. These recordings will be used in later chapters as the basis for queue models.

In a two-person team, observe a queue (your instructor will provide direction) for *exactly one hour* and record the following (to the second):

1. Time that each person arrived at the queue
2. Time that each person began service
3. Time that each person completed service
4. Time that each person reneged (if applicable)

In addition, record the number of people in the queue and the number of people in service at the beginning of the hour and at the end of the hour.

Based on the recordings, each team should turn in the following:

1. A graph showing cumulative arrivals, cumulative departures from queue, and cumulative departures from system
2. A graph showing the empirical probability distribution for the interarrival times
3. If applicable, a cumulative graph for reneging customers
4. A graph showing the empirical probability distribution for the service times
5. Numerical estimates of the following (show work):
   - **(a)** Average and standard deviation of queue length and average number of customers in the system
   - **(b)** Average and standard deviation of time in queue and time in system
   - **(c)** Average and standard deviation of the interarrival time
   - **(d)** Average and standard deviation of the service time
   - **(e)** Average arrival rate.
6. Verify the equations $L_q = \lambda W_q$ and $L_s = \lambda W_s$.

APPENDIX—DEFINITE INTEGRALS

The *definite integral* of a non-negative single-valued function

$$\int_a^b f(x)\,dx \tag{2.A1}$$

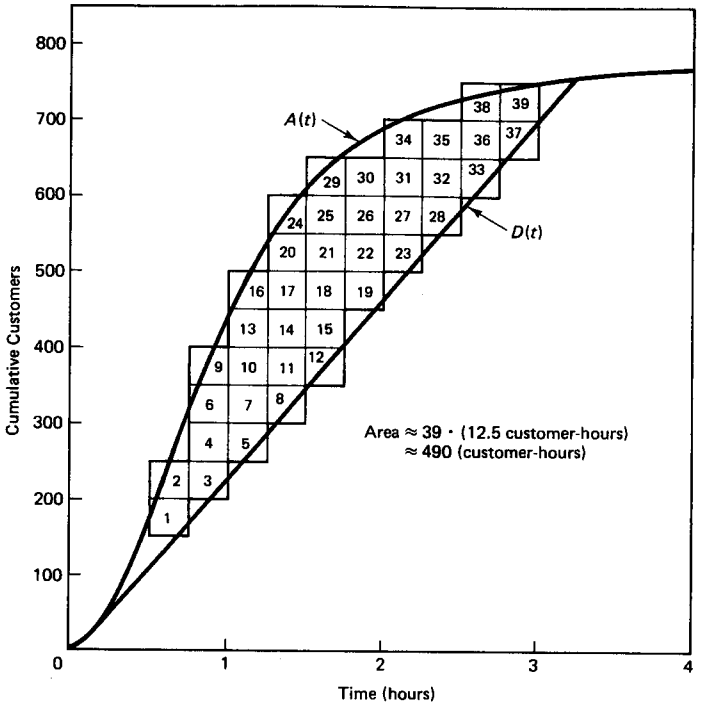


Figure 2.14 Approximation of definite integral by counting squares.

equals the area between the x-axis and $f(x)$ over the range from a to b. When $f(x)$ is expressed as an equation, a definite integral can by calculated by one of the methods of numerical integration found in any calculus text. For some functions, the definite integral can also be determined by evaluating the indefinite integral of $f(x)$ at b and a and taking the difference.

A definite integral can also be calculated when a function is not expressed as an equation but is plotted on a piece of graph paper.

Method 1: Weights. This is the simplest method but requires special equipment. From the piece of graph paper, cut out the section of paper bounded by the vertical lines through points a and b, the x-axis, and the function. Weigh this section of paper on an accurate scale (W_f). Cut out a large square, of known dimensions x by y, from the remaining section of graph paper and weigh it (W_s). The integral is approximately

$$\int_a^b f(x)dx \approx \frac{W_f}{W_s}\,(x \cdot y) \tag{2.A2}$$

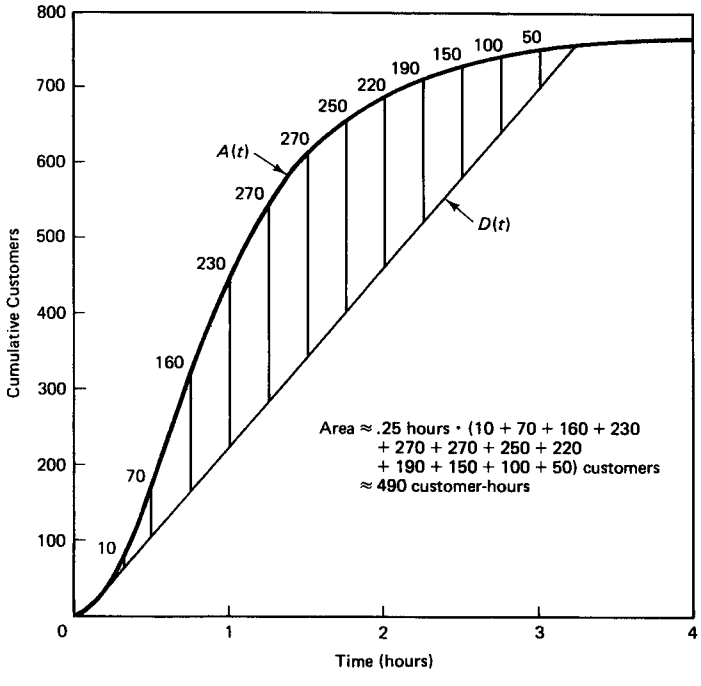


Figure 2.15 Approximation of definite integral by measuring lines.

Method 2: Squares. This approach is simple but a bit tedious. Determine the size of a single square on the graph paper (A_s). If the majority of a square's area falls in the desired region, shade it a light color. Count the number of shaded squares (N). The integral is approximately (see Fig. 2.14):

$$\int_a^b f(x)dx \approx N \cdot A_s \tag{2.A3}$$

The smaller the square is, the more accurate the approximation will be. Of course, this requires more effort.

Method 3: Lines. This approach requires more than just counting but can be more accurate than method 2. On a piece of graph paper, draw a series of equidistant vertical lines (spaced d_x apart) over the range from a to b. Calculate the height of each line, then sum the data to obtain a total height, h_t. As shown in Fig. 2.15, the integral is approximately

$$\int_a^b f(x)dx \approx d_x \cdot h_t \tag{2.A4}$$

As with method 2, accuracy increases when the spacing decreases.

Method 4: Geometry. This approach is less general. Some definite integrals represent the calculation of ordinary geometric shapes, such as rectangles, triangles, and trapezoids. This is true of cumulative diagrams, because of the discrete nature of customer counts. Partition the area of interest into rectangular, triangular, and trapezoidal regions. Calculate the area of each region and then sum the areas. Provided that the areas can be accurately measured, the result should be the exact integral. An example is shown in Fig. 2.5a.