chapter 6

Nonstationary Arrivals

If you were to think of the most frustrating, the most aggravating, and the most time-consuming sort of queue, there is a good chance that the evening rush-hour commute home from work would come to mind. The rush hour, both morning and evening, is the product of large numbers of people desiring to use the roads and highways at the same time. It is also an example of a nonstationary arrival pattern.

A ***nonstationary*** (also called *nonhomogeneous*) arrival pattern occurs when the customer arrival rate varies over time. It is a phenomenon that virtually all queueing systems experience to one extent or another. Restaurants have rush periods at lunchtime and dinnertime. Retail stores experience a rush period in the month before Christmas. Accountants experience rush periods prior to tax due dates. Rush periods are the consequence of the natural cycles in our lives. We orient ourselves toward daily, weekly, monthly, and yearly patterns. We tend to work at the same time, eat at the same time, sleep at the same time, shop at the same time, . . . And this puts a strain on queueing systems. It is far easier to serve customers when they arrive at an even rate than an uneven rate.

We have seen that when a queueing system operates in steady state, queues are produced by ***random variability*** in service times and arrival times. In steady state, there is no way to know when these queues will occur because they are totally random. A nonstationary arrival pattern presents an additional source of variability: ***predictable variability***. Predictable variability in the arrival pattern means that queues occur in a

predictable fashion, at the same time every day, or every week, and so on. Predictable queues (such as the evening rush hour) tend to be much larger and costlier than random queues.

This chapter provides several approaches for modeling a nonstationary arrival process and queueing system. It begins by defining the nonstationary version of the Poisson process. Next, a procedure is presented for using steady-state equations to model certain, lightly used queueing systems. This is followed by a description of how to simulate a nonstationary queueing process. Then a much simpler model, known as a fluid approximation, is provided. The chapter concludes with an alternative way to define the arrival process, in terms of desired departure times from the system.

6.1 THE NONSTATIONARY POISSON PROCESS

The nonstationary Poisson process is a Poisson process for which the arrival rate varies with time. More specifically, it can be defined as follows:

Definitions 6.1

The counting process $N(t)$ is a ***non-stationary Poisson process*** if:

A. The process has independent increments

B. $Pr\,[N(t + dt) - N(t) \begin{cases} = 0] = 1 - \lambda(t)dt \\ = 1] = \lambda(t)dt \\ > 1] = 0 \end{cases}$

where

$\lambda(t)$ = the arrival rate at time t

dt = a differential sized time interval

The definition is identical to the stationary Poisson process (Chap. 3), with the exception that the arrival rate, $\lambda(t)$, is now a function of time. As before, the arrival rate represents the *expected* number of customers to arrive per unit time. If λ (9:00) equals 10 per minute, then we would expect to see ten customers arrive in the 1-minute interval between 9:00 and 9:01 on average. The actual number of customers to arrive can be either smaller or larger than ten.

The arrival rate, having the dimensions *customers/time*, when integrated with respect to time, yields the expected number of customers to arrive over a *time interval*:

$$E(\text{arrivals between time } a \text{ and time } b) = \int_a^b \lambda(t)dt \qquad (6.1)$$

One can think of the stationary Poisson process as a special version of the nonstationary Poisson process. So, to take an example, if $\lambda(t) = \lambda = 10$ customers/hour, then the expected number of arrivals over a 1-hour period is

$$E(\text{arrivals over 1 hour}) = \int_0^1 10\,dt = 10t \Big|_0^1 = 10 \text{ customers}$$

And the expected number of arrivals over a half-hour period is

$$E(\text{arrivals over } \tfrac{1}{2} \text{ hour}) = \int_0^{.5} 10\,dt = 10t \Big|_0^{.5} = 5 \text{ customers}$$

As might be expected, these results are identical to what was found in Chap. 3 when the stationary Poisson process was presented. Of course, $\lambda(t)$ does not have to be constant. Suppose that customers arrive at a restaurant at the following rate:

$$\lambda(t) = 100\sin(t\pi/2) \qquad 0 \leq t \leq 2 \text{ (customers/hour)} \tag{6.2}$$

where time $t = 0$ is 11:30 A.M. and time $t = 2$ is 1:30 P.M., and the sine angle is measured in radians. With this function, customers arrive at the fastest rate at 12:30 (100 customers/hour) and the slowest rate at 11:30 and 1:30 (0 customers/hour).

Definition 6.2

$\Lambda(t)$ is the expected number of arrivals from time 0 to time t. $\Lambda(t)$ is calculated by integrating $\lambda(t)$ from 0 to time t:

$$\Lambda(t) = \int_0^t 100\sin(\pi\tau/2)d\tau = (200/\pi)[1 - \cos(\pi t/2)] \text{ customers} \tag{6.3}$$

Note that the variable used in the integrand (τ) must be distinguished from the variable used to bound the integral (t). Also note that $\Lambda(t)$ is measured in terms of customers, whereas $\lambda(t)$ is measured as a rate: customers per hour. $\Lambda(t)$ is plotted for the example in Fig. 6.1. The slope of $\Lambda(t)$ (that is, the derivative) is $\lambda(t)$. The figure shows that the arrival rate is largest at the center of the time interval and smallest at the ends, as already predicted. Remember that $\Lambda(t)$ represents the expected number of arrivals to occur by time t. The actual number of arrivals can be either larger or smaller than $\Lambda(t)$.

It is more common to base $\lambda(t)$ on ***interval counts*** than on an equation, as shown above. In the restaurant example, records might indicate that the average numbers of arrivals in each of four time periods are the following:

Time	Average arrivals
11:30–12:00	19
12:00–12:30	45
12:30–1:00	45
1:00–1:30	19

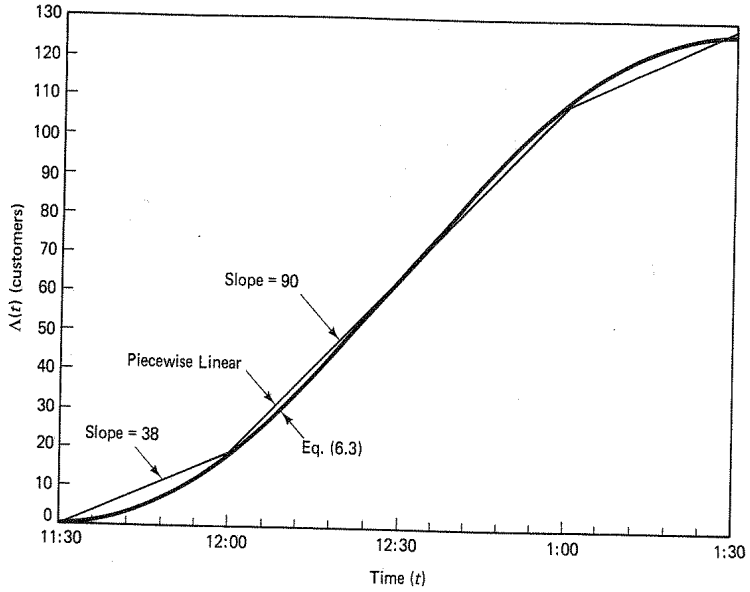

Figure 6.1 Expected cumulative arrivals at restaurant versus time.

These numbers can be translated into an arrival rate by dividing the number of arrivals by the size of the time interval. In all four cases, the time interval is one-half hour, so the arrival rate is

$$\lambda(t) = \begin{cases} 38 & 0 \leq t < .5 \\ 90 & .5 \leq t < 1.5 \\ 38 & 1.5 \leq t \leq 2 \end{cases} \text{ (customers/hour)} \tag{6.4}$$

The units here are important. If the interval counts are divided by a time unit measured in hours, then $\lambda(t)$ is measured in terms of arrivals per hour. $\Lambda(t)$ can, as usual, be found by integrating $\lambda(t)$:

$$\Lambda(t) = \begin{cases} 38t & 0 \leq t < .5 \\ 19 + 90(t - .5) & .5 \leq t < 1.5 \\ 109 + 38(t - 1.5) & 1.5 \leq t \leq 2 \end{cases} \tag{6.5}$$

$\Lambda(.5)$ is the average number of arrivals recorded for the first half-hour, $\Lambda(1)$ is $\Lambda(.5)$ plus the average number of arrivals recorded in the second half-hour, and so on. So $\Lambda(t)$ can actually be calculated by summing the interval counts and interpolating between the points. This new version of $\Lambda(t)$ is shown in Fig. 6.1 next to the plot of the equation for $\Lambda(t)$. For the interval counts, $\Lambda(t)$ is a piecewise linear curve, and $\lambda(t)$ (the slope of $\Lambda(t)$) is a step curve, with discontinuities at the ends of the time intervals. In reality, the true

arrival rate (the $\lambda(t)$ that generates the arrivals) would not have these discontinuities. The discontinuities are the unavoidable by-product of averaging the number of arrivals over time intervals.

6.1.1 Properties of the Non-Stationary Poisson Process

The nonstationary Poisson process does not possess the property that interarrival times are exponential random variables. Hence, it also does not possess the property that the time until the nth arrival is a gamma random variable. Yet it does have several properties in common with the stationary Poisson process. Most important of these is that the number of arrivals over any time interval is a Poisson random variable:

Property 1

The number of arrivals over the interval $[a,b]$ is Poisson with mean

$$E[A(b) - A(a)] = \int_a^b \lambda(t)dt = \Lambda(b) - \Lambda(a)$$

Example

The restaurant owner would like to determine the probability that three or fewer customers will arrive between 11:30 and 11:45, using the equation for $\lambda(t)$. The expected number of arrivals, from Eq. (6.3), is $\Lambda(.25) - \Lambda(0) = 4.85$ customers. The probability of n customers arriving is

$$P(n \text{ arrivals between 11:30 and 11:45}) = \frac{4.85^n}{n!} e^{-4.85} \quad n = 0, 1, \ldots$$

The probability of three or fewer arrivals is found by evaluating the above equation for $n = 0$, $n = 1, \ldots, n = 3$, which equals $.008 + .038 + .092 + .149 = .287$.

The event times within a time interval also have properties similar to the stationary Poisson process. As with the stationary Poisson process, the time of any event is independent of the time of any other event. The nonstationary process is also similar to the stationary process in that the probability distribution for the unordered event times is defined by $\Lambda(t)$:

Property 2

If $A(t)$ is the number of events in the interval $[0,\tau]$, the unordered event times are defined by $A(t)$ independent random variables with the probability distribution:

$$P(T \leq t) = \frac{\Lambda(t)}{\Lambda(\tau)} \tag{6.6}$$

where T is the random variable representing the event time.

With the stationary Poisson process, $\Lambda(t) = \lambda t$, so $P(T \leq t)$ is simply t/τ. This defines the uniform probability distribution over $[0,\tau]$. Again, the stationary Poisson process is a special case of the nonstationary Poisson process. More generally, the event times can have any distribution, as defined by the function $\Lambda(t)$. There is no reason to expect that it has any particular shape. The shape is determined from historical records of customer arrivals.

Example

The restaurant owner knows that five customers arrived between 11:30 and 11:45. He would now like to determine the likelihood that no one arrived before 11:35, using Eq. (6.3). The probability that any one of the five customers arrived before 11:35 is

$$P(\text{arrived before 11:35}) = \frac{\Lambda(1/12)}{\Lambda(.25)} \quad \frac{.545}{4.85} = .112$$

The probability that no one arrived before 11:35 is $(1 - .112)^5 = .551$.

Keep in mind that because the interarrival times are not exponential, the nonstationary Poisson process does not possess the memoryless property.

6.1.2 Goodness of Fit

The basic concept of checking for goodness of fit is the same for a nonstationary Poisson process as for a stationary Poisson process. As always, this begins with a check of plausibility. Does the probability that a customer arrives at any time depend on the times when other customers arrived? Do customers arrive one at a time? The answers to these questions must be affirmative for both the nonstationary and the stationary processes. The major difference in the plausibility check is that the arrival rate does not have to be constant. Thus, the conditions for the nonstationary process are not as strict. Many real arrival processes satisfy the conditions underlying the nonstationary Poisson process.

The quantitative goodness of fit tests are somewhat different, because the interarrival times do not have to be independent exponential random variables, and the arrival times within a time interval do not have to be uniform. The primary check is for the hypothesis:

H_1: The number of events in any time interval has a Poisson distribution.

This test requires large quantities of data, representing the numbers of arrivals over many recurring cycles. It is not enough to know how many customers arrived over each time interval of a single cycle (a day, for example). One must know the number of arrivals over each time interval of many cycles. Then the number of arrivals within each interval should be a Poisson random variable. This test can be carried out through a slight modification of the Kolmogorov-Smirnov test (see the statistics texts cited at the end of Chap. 3). Practically speaking, however, it is virtually impossible to obtain sufficient data to carry out the test, and, in the end, one must rely on the plausibility check. If you believe that arrivals are independent and you believe that customers arrive one at a time, then it should be safe to assume that the arrival process is nonstationary Poisson.

6.1.3 Parameter Estimation

The nonstationary Poisson process is not defined by the single parameter, λ, but by a function, $\lambda(t)$. This makes parameter estimation more complicated. The most straightforward approach is to base $\lambda(t)$ on interval counts, as was already illustrated in this chapter. Suppose that f_n is the average number of arrivals in interval n, from time a to time b, over I cycles. Then

$$\hat{\lambda}(t) = \frac{f_n}{b - a} \qquad a \leq t \leq b \tag{6.7}$$

A confidence interval for $\lambda(t)$ can be formed under the hypothesis that the number of arrivals in any interval is a Poisson random variable. Hence, the variance of the number of arrivals is the same as the expected number of arrivals. Because f_n is the average of a set of random variables, it must have a normal distribution if I is large (central limit theorem). This leads to the following confidence intervals:

95% Confidence ($I \geq 50$)

$$P[f_n/(b - a) - 1.96\sqrt{f_n/I}/(b - a) \leq \lambda(t) \leq f_n/(b - a) + 1.96\sqrt{f_n/I}/(b - a)] = .95 \tag{6.8}$$

99% Confidence ($I \geq 50$)

$$P[f_n/(b - a) - 2.58\sqrt{f_n/I}/(b - a) \leq \lambda(t) \leq f_n/(b - a) + 2.58\sqrt{f_n/I}/(b - a)] = .99 \tag{6.9}$$

Example

Suppose that the interval counts for the restaurant are based on 60 days of records. The 95% confidence interval for the 11:30 to 12:00 period, in which 19 arrivals were observed on average, is calculated as follows:

$$P(19/(.5) - 1.96\sqrt{19/60}/.5 \leq \lambda(t) \leq 19/.5 + 1.96\sqrt{19/60}/.5) = .95$$

$$P(38 - 2.2 \leq \lambda(t) \leq 38 + 2.2) = .95$$

In the example, the 95% confidence interval is fairly small, but this relies on 60 days of records. Clearly, obtaining precise estimates of $\lambda(t)$ requires, at a minimum, detailed data on dozens of cycles. Yet, even doing this may not suffice, for there may be no way of guaranteeing that the arrival rate will stay the same every day, every week, or every month. The pattern may never recur. This presents a problem with no clear resolution. No matter what approach is used, predictions based on estimates of $\lambda(t)$ will usually be imprecise.

Another decision to consider is how large the time intervals should be. It is certainly much easier to obtain a precise estimate for the number of arrivals over 1-hour intervals than 1-minute intervals, yet, if the arrival rate truly varies over the hour interval, the

variation will not be detected. As a rule, the intervals should be sufficiently small to detect any major changes in the arrival rate, but no smaller than necessary (unless there is an easy way to record arrival data). If a typical queue lasts 1 to 2 hours, then intervals of width 10 to 20 minutes should be sufficient. If a typical queue lasts an entire day, intervals of 1 hour should be sufficient; and if a typical queue lasts a week or more, then intervals of one day should be sufficient.

The alternative to interval counts is to derive $\lambda(t)$ from an estimate of $\Lambda(t)$. Suppose that $A_n(t)$ represents the cumulative arrivals to time t for cycle n of I total cycles. Then an estimate for $\Lambda(t)$ can be obtained as follows:

$$\hat{\Lambda}(t) = \frac{\sum_{n=1}^{I} A_n(t)}{I} = \bar{A}(t) \tag{6.10}$$

The natural estimate for $\lambda(t)$ would be the derivative of $\hat{\Lambda}(t)$. However, because $A_n(t)$ is a step function, $\hat{\Lambda}(t)$ must be too, meaning that the derivative of $\hat{\Lambda}(t)$ is undefined. As an alternative, $\hat{\Lambda}(t)$ can be set equal to a smooth approximation to the average of the arrival curves (a similar approach is shown in Fig. 4.4). The confidence interval for $\Lambda(t)$ is formulated in much the same way as the confidence interval for $\lambda(t)$:

95% Confidence ($I \geq 50$)

$$P[\hat{\Lambda}(t) - 1.96\sqrt{\hat{\Lambda}(t)/I} \leq \Lambda(t) \leq \hat{\Lambda}(t) + 1.96\sqrt{\hat{\Lambda}(t)/I}] = .95 \tag{6.11}$$

99% Confidence ($I \geq 50$)

$$P[\hat{\Lambda}(t) - 2.58\sqrt{\hat{\Lambda}(t)/I} \leq \Lambda(t) \leq \hat{\Lambda}(t) + 2.58\sqrt{\hat{\Lambda}(t)/I}] = .99 \tag{6.12}$$

The confidence interval is itself a function of t. The absolute width of the interval expands as t increases because the standard error for the estimator $\hat{\Lambda}(t)$ grows with t (but the relative width declines).

A third approach to estimating $\lambda(t)$ is to approximate the average of the cumulative arrival curves with an equation. The advantage of this approach is that it is much simpler to analyze an equation than a large data set. However, some loss in accuracy may result. The methodology for estimating such an equation is beyond the scope of this book, but can be found in texts on econometrics and statistical regression (see the end of this chapter).

An obvious disadvantage of the second and third approaches is that data on specific arrival times are required, whereas the interval approach only requires interval counts. The added data collection effort may not be justified in terms of increased accuracy.

Future arrival rates might also be partially predicted through a forecasting technique, of which there are many. A common approach is to base the shape of the curve $\Lambda(t)$ on the average of the historical arrival curves, but to scale the curve according to the forecast for the number of arrivals for a given cycle. That is, $\hat{\Lambda}(T)$ would equal a forecast for the number of arrivals during the cycle, and $\Lambda(t)$ would be scaled up or down from

$\Lambda(t)$ by the ratio $\hat{\Lambda}(T)/\bar{\Lambda}(T)$. (References on forecasting are provided at the end of this chapter.) Keep in mind that one's own judgment sometimes provides a good forecast, particularly when the arrival pattern is influenced by many external factors.

Finally, a nonstationary arrival process does not have to be cyclic; $\Lambda(t)$ can be any nondecreasing function. However, unless $\Lambda(t)$ is cyclic, it may be impossible to estimate $\Lambda(t)$, in which case the arrival pattern is not truly predictable. If the arrival pattern is not predictable, then it should not be modeled as a nonstationary Poisson process. The essence of the nonstationary Poisson process is *predictable* variability in the arrival process.

6.2 STEADY-STATE APPROXIMATION FOR A SLOWLY VARYING ARRIVAL RATE

A queueing system with a nonstationary arrival process will never enter steady state. The varying arrival rate constantly changes the probability distribution for the number of customers in the system. Yet this does not prevent the use of the steady-state equations to *approximate* the behavior of the system, particularly when the arrival rate is slowly changing and the system operates below capacity. When valid, the system will be said to be in ***quasi-steady state***.

Consider the performance of a single server queue, with exponential service times and a nonstationary Poisson arrival process.

Definitions 6.3

$\mu(t)$ = the service rate per server at time t

$\rho(t)$ = the absolute utilization at time t
$= \lambda(t)/\mu(t)$

$P_n(t)$ = the probability that n customers are in the system at time t

Then the steady-state approximation for $P_n(t)$ follows directly from the $M/M/1$ queueing equations:

$$P_n(t) \approx [1 - \rho(t)][\rho(t)]^n \qquad n = 0, 1, 2 \tag{6.13}$$

The validity of this approximation clearly depends on at least one factor: $\rho(t)$ *must be less than 1 for all* t. This factor in itself is quite restrictive, for systems with nonstationary arrival rates also tend to be overloaded from time to time. $\rho(t)$ must also be a slowly varying function. If $\rho(t)$ changes too quickly, then the system could not respond as fast as the steady-state model predicts. More precisely, the steady-state approximation can only be accurate if *the change in* $\rho(t)$ *during one relaxation time* (see Chap. 4) *is small in comparison to the average queue length*. Based on this principle, Newell (1982) provides the following rule for assessing the validity of the steady-state approximation:

$$\Delta = \left[\frac{1}{\mu(t)}\right]\left[\frac{1}{[1 - \rho(t)]^3}\right]\left|\frac{d\rho(t)}{dt}\right| \qquad \rho(t) < 1 \tag{6.14}$$

Single server approximation valid when $\Delta << 1$

One way to interpret Eq. (6.14) is that the amount that $\rho(t)$ changes during one service time ($d_\rho(t)/dt \mid \cdot 1/\mu$) should be small compared to the quantity $[1-\rho(t)]^3$.

Example

Consider the restaurant example again, with $\lambda(t) = 100\sin(t\pi/2)$. Suppose that the operator is concerned with queues of customers waiting to be seated by the maître d'hôtel. Hypothetically, the time to serve a customer has an exponential distribution with mean 20 seconds (.00556 hour). $\rho(t)$ is defined as follows:

$$\rho(t) = \frac{\lambda(t)}{\mu} = \frac{100\sin(\pi t/2)}{180} = .556\sin(\pi t/2)$$

To evaluate Eq. (6.14), the derivative of $\rho(t)$ is calculated first:

$$\frac{d\rho(t)}{dt} = .556\,\frac{\pi}{2}\cos(\pi t/2) = .873\cos(\pi t/2)$$

Δ can now be written as

$$\Delta = \frac{1}{180}\,\frac{|.873\cos(\pi t/2)|}{[1 - .556\sin(\pi t/2)]^3}$$

The above is calculated for various values of t

t	0	.25	.5	.75	.9	1	1.1	1.25	1.5	1.75	2
Δ	.0048	.0092	.0213	.0161	.0083	0	.0083	.0161	.0213	.0092	.0048

In this instance, the quasi-steady-state model seems appropriate (largely because the service time is very small) and the system quickly adapts to changes in the arrival rate.

Should the steady-state approximation be valid, then the performance measure equations from Chap. 5 can be used. For the example of an $M/M/1$ system

$$E[L_s(t)] \approx \frac{\rho(t)}{1 - \rho(t)} \tag{6.15}$$

Figure 6.2 shows $E[L_s(t)]$ for the restaurant example. The multiple server steady-state results (the $M/M/m$ and $M/G/1$ models, for example), can also be used if Eq. 6.14 is satisfied, *provided the following substitutions are made for* $\mu(t)$ *and* $\rho(t)$ *in Eq. 6.14, respectively*.

Definitions 6.4

$c(t)$ = the *combined* service capacity among all servers at time t

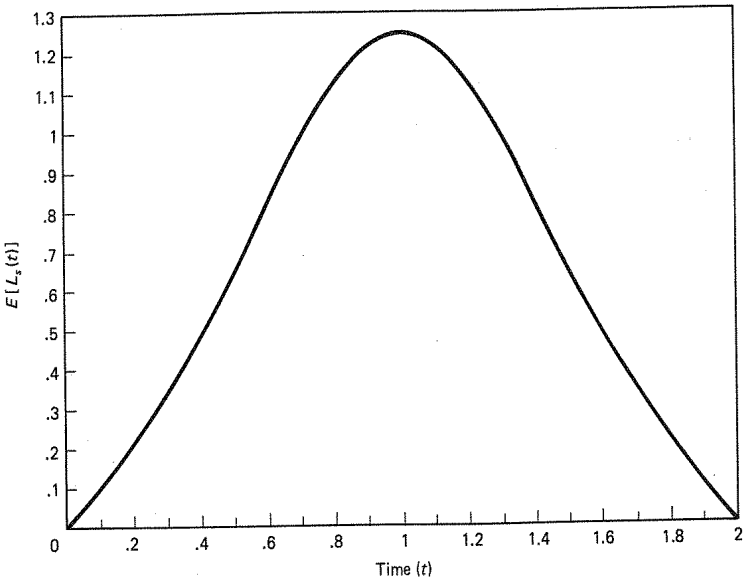

Figure 6.2 Expected customers in system for restaurant, determined by quasi-steady-state model.

$$\tilde{\rho}(t) = \text{proportional utilization at time } t$$
$$= \lambda(t)/c(t)$$

For example, if four servers work at the rate of ten customers per hour each, then $c(t) = 40$ customers/hour. Unfortunately, the steady-state equations have limited validity because they are not accurate when $\tilde{\rho}(t)$ is close to or exceeds 1, which invariably occurs from time to time when the arrival rate is not stationary.

6.3 SIMULATION OF A NONSTATIONARY POISSON PROCESS

If quasi-steady-state analysis is not applicable, one alternative is to simulate the queueing system. As mentioned in Chap. 4, simulation is a very robust technique that can be applied to a variety of situations. However, simulation does not always provide as meaningful results as does direct analysis. In this section, three techniques are presented for simulating a queueing system with a nonstationary Poisson arrival process. The first two are similar to techniques used in Chap. 4 for the stationary Poisson process and the third is new. A fourth technique is presented for the special case where $\Lambda(t)$ is piecewise linear. In all but the third technique, the basic approach is first to simulate arrival times, second simulate service times, and third combine the data to form the queue simulation.

The second and third steps are no different from those in Sec. 4.5, so they will not be repeated. The emphasis here is on simulating the arrival times.

6.3.1 Simulation Method 1

Recall that there are two ways to simulate a stationary Poisson process. The first of these is to simulate exponential interarrival times and sum them to obtain arrival times. Clearly, this approach will not work for a nonstationary arrival process; the interarrival times are not exponential random variables. However, a modification will work. Let

$$\lambda_{max} = \max_t \lambda(t)$$

The nonstationary Poisson process is simulated as follows:

1. Simulate a stationary Poisson process with rate λ_{max} by summing exponential interarrival times.
2. For each arrival simulated, generate a Bernoulli {0,1} random variable, with $p = \lambda(t)/\lambda_{max}$; from a U[0,1] random variable, U:

$$U \leq p \text{ denotes a success:accept arrival}$$
$$U > p \text{ denotes a failure:reject arrival}$$

Example

The restaurant is to be simulated over the time interval from 11:30 to 11:45($t = 0$ to .25, with Eq. (6.3)). The maximum arrival rate over this period occurs at $t = .25$, and equals

$$\lambda_{max} = 100\sin(.25\pi/2) = 38 \text{ customers/hour} \quad \text{(Eq. (6.2))}$$

Taking $U[0,1]$ random variables (U_{n1} and U_{n2}) from a random number table, the simulation is summarized in the table below:

n	U_{n1}	X_n	Y_n	$p = \lambda(t)/\lambda_{max}$	U_{n2}	Accept
1	.2188	.040	.04	.165	.8479	No
2	.4846	.019	.059	.243	.6108	No
3	.9586	.001	.060	.248	.5703	No
4	.4061	.024	.084	.346	.3113	Yes
5	.1037	.060	.144	.590	.3349	Yes
6	.5104	.018	.162	.662	.4038	Yes
7	.6088	.013	.175	.714	.9031	No
8	.0707	.070	.245	.988	.7986	Yes
9	.7919	.006	—			

The series Y_n represents a stationary Poisson process with rate 38. The nonstationary simulation accepts four of these arrivals, yielding the arrival times .084, .144, .162, and .245 hours.

Note that the acceptance probability varies in proportion to the arrival rate, so the simulation is truly nonstationary.

6.3.2 Simulation Method 2

The second approach presented in Chap. 4 can also be modified for a nonstationary Poisson process. The arrival process is simulated in these three steps.

1. Simulate a Poisson random variable $A(T)$ representing the number of arrivals over the time interval $[0,T]$.
2. Simulate $A(T)$ random variables representing the arrival times.
3. Sort the arrival times in ascending order to obtain the function $A(t)$.

The mean of the Poisson random variable equals $\Lambda(T)$. The simulation in the second step follows from the probability distribution for the arrival times defined by $\Lambda(t)$. Suppose that Y represents an arrival time. Then Y is found by solving the following, where U represents a uniform $[0,1]$ random variable:

$$U = \frac{\Lambda(Y)}{\Lambda(T)} \rightarrow Y = \Lambda^{-1}[U \cdot \Lambda(T)] \tag{6.16}$$

Example

The restaurant is to be simulated a second time over the interval from 11:30 to 11:45 ($t = 0$ to .25, with Eq. (6.3)). The expected number of arrivals over this interval equals 4.85. $A(.25)$ is found as follows:

a. $U = .435$ (from random number table)

b. From the Poisson distribution: $P[A(.25) \leq 3] = .287$
$P[A(.25) \leq 4] = .467 \rightarrow A(.25) = 4$

The arrival times are generated from solving the following:

$$U = \frac{\frac{200}{\pi}[1 - \cos(\pi Y/2)]}{4.85} \rightarrow Y = \frac{2}{\pi}\cos^{-1}\left[1 - U\frac{4.85\pi}{200}\right]$$

Taking $U[0,1]$ random variables from a random number table leads to the following values:

n	$U[0,1]$	Y
1	.8177	.226
2	.3677	.151
3	.2125	.115
4	.5474	.185

As in Chap. 4, the arrival times are sorted in ascending order to obtain $A(t)$.

The example can be visualized through Fig. 6.3, which shows how the arrival times are generated. The curve is $\Lambda(t)$, for t in the domain $[0,.25]$.

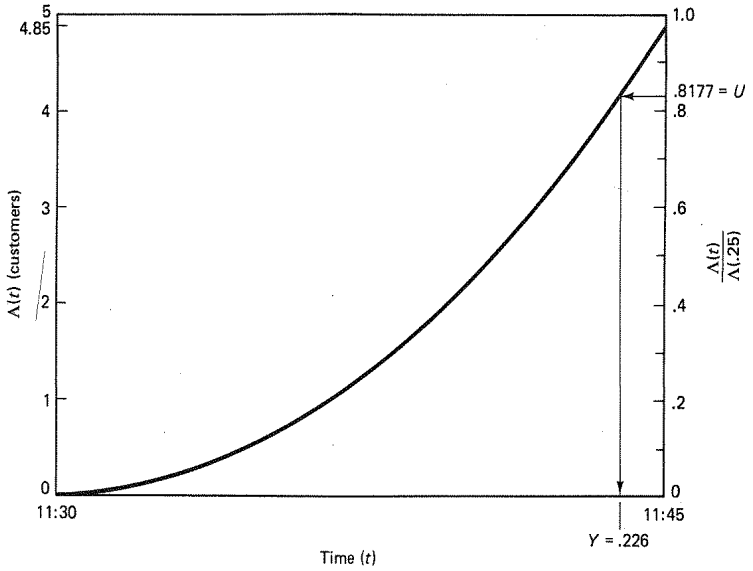

Figure 6.3 Simulation of a customer arrival time for a nonstationary process.

6.3.3 Simulation Method 3

An alternative simulation approach is to use an ***activity scanning*** procedure. This approach draws on the primary definition of the nonstationary Poisson process—that is, that the process has independent increments, and the probability of an arrival in a differential time interval dt is $\lambda(t)dt$. The simulation is an approximation to the Poisson process and is something like flipping coins whose probability of success $= \lambda(t)dt$. It is carried out in these steps:

1. Divide the time period $[0,T]$ into small time increments.
2. For each time increment, generate a $U[0,1]$ random variable.
3. For each increment simulate a Bernoulli random variable

If $U \leq \lambda(t)dt \rightarrow$ then an arrival occurred in the increment.
If $U > \lambda(t)dt \rightarrow$ then no arrival occurred in the increment.

The accuracy of the simulation depends on the size of the time increments. If these increments are very small, then the simulation will be indistinguishable from a nonstationary Poisson process; if the increments are large, then $A(T)$ will have a smaller variance than a nonstationary Poisson process (but usually the same mean). If the probability that a customer arrives in any time increment is no larger than .1, then the standard deviation of $A(T)$ will be within 5 percent of the standard deviation of the nonstationary Poisson

process, and if the probability that a customer arrives in any time increment is no larger than .05, then the standard deviation of $A(T)$ will be within 2.5 percent of the standard deviation of the Poisson process. Though it is usually easiest to use equal sized increments, nothing prevents the use of smaller time increments when the arrival rate is large and larger time increments when the arrival rate is small. In so doing, the probability of an arrival in an increment stays more or less constant.

It should be apparent that method 3 requires more computations than method 1 or 2, particularly if a high degree of accuracy is desired. Yet method 3 is not without merit. It is a more robust approach and applies to a greater variety of queueing characteristics because it allows the arrival and service simulations to be carried out simultaneously. For example, it can be used to simulate reneges, which do not necessarily occur when customers arrive or depart, but may occur at any point in time. The approach is particularly effective in an interactive computing environment, for it allows the user to see how the queueing system evolves over time (at a rapidly accelerated time scale). Many computer simulation programs use the activity scanning approach.

Example

The restaurant is to be simulated a third time over the interval [0,.25], this time with method 3. The maximum arrival rate over this interval is 38/hour. To keep the probability of an arrival less than .1, the time intervals should be no larger than 1/380 hours = 9.5 seconds. This has been rounded off to 10 seconds. Thus, a total of 90 time increments are simulated over the 15 minute period, as shown in Table 6.1.

The queue can be simulated by generating arrivals, generating service times, and combining the data, just as before. However, should one go to the effort of using method 3, an alternative approach would likely be used. At each time increment, the following steps would be performed:

1. Determine whether an arrival occurs.
2. For each customer in service, determine whether service is completed in the time increment.
3. Determine whether any customer enters service in the time increment.

In addition, extra steps can be added to account for reneging or other factors. Method 3 amounts to a dynamic simulation, as opposed to the alternatives, which are more of a batch simulation.

6.3.3 Special Case: $\Lambda(t)$ Is Piecewise Linear

A nonstationary arrival process is easiest to simulate when $\Lambda(t)$ is a piecewise linear function, meaning that the arrival rate stays constant over each of several time intervals. This special case is not all that unusual, for if $\Lambda(t)$ is based on interval counts, the arrival rate must be assumed to be constant over each interval. In reality, the arrival rate is likely some smooth function of time—it is just that data are not available to determine its exact

TABLE 6.1 RESTAURANT ARRIVALS SIMULATION: METHOD 3

Time (sec)	U	$\lambda(t)dt$	1 = Arr	$A(t)$	Time (sec)	U	$\lambda(t)dt$	1 = Arr	$A(t)$
10	0.172	0.001	0	0	460	0.523	0.055	0	1
20	0.944	0.002	0	0	470	0.876	0.057	0	1
30	0.556	0.004	0	0	480	0.175	0.058	0	1
40	0.036	0.005	0	0	490	0.891	0.059	0	1
50	0.282	0.006	0	0	500	0.154	0.060	0	1
60	0.380	0.007	0	0	510	0.405	0.061	0	1
70	0.205	0.008	0	0	520	0.980	0.062	0	1
80	0.455	0.010	0	0	530	0.062	0.064	1	2
90	0.279	0.011	0	0	540	0.690	0.065	0	2
100	0.515	0.012	0	0	550	0.390	0.066	0	2
110	0.158	0.013	0	0	560	0.655	0.067	0	2
120	0.437	0.015	0	0	570	0.438	0.068	0	2
130	0.121	0.016	0	0	580	0.957	0.070	0	2
140	0.015	0.017	1	1	590	0.234	0.071	0	2
150	0.422	0.018	0	1	600	0.742	0.072	0	2
160	0.809	0.019	0	1	610	0.788	0.073	0	2
170	0.666	0.021	0	1	620	0.623	0.074	0	2
180	0.225	0.022	0	1	630	0.899	0.075	0	2
190	0.469	0.023	0	1	640	0.229	0.077	0	2
200	0.600	0.024	0	1	650	0.572	0.078	0	2
210	0.943	0.025	0	1	660	0.909	0.079	0	2
220	0.105	0.027	0	1	670	0.088	0.080	0	2
230	0.123	0.028	0	1	680	0.440	0.081	0	2
240	0.463	0.029	0	1	690	0.081	0.082	1	3
250	0.774	0.030	0	1	700	0.012	0.084	1	4
260	0.297	0.031	0	1	710	0.164	0.085	0	4
270	0.802	0.033	0	1	720	0.726	0.086	0	4
280	0.434	0.034	0	1	730	0.991	0.087	0	4
290	0.050	0.035	0	1	740	0.958	0.088	0	4
300	0.296	0.036	0	1	750	0.984	0.089	0	4
310	0.530	0.037	0	1	760	0.300	0.090	0	4
320	0.644	0.039	0	1	770	0.640	0.092	0	4
330	0.207	0.040	0	1	780	0.882	0.093	0	4
340	0.970	0.041	0	1	790	0.056	0.094	1	5
350	0.393	0.042	0	1	800	0.828	0.095	0	5
360	0.827	0.043	0	1	810	0.558	0.096	0	5
370	0.703	0.045	0	1	820	0.962	0.097	0	5
380	0.905	0.046	0	1	830	0.715	0.098	0	5
390	0.352	0.047	0	1	840	0.645	0.100	0	5
400	0.753	0.048	0	1	850	0.067	0.101	1	6
410	0.531	0.049	0	1	860	0.504	0.102	0	6
420	0.320	0.051	0	1	870	0.213	0.103	0	6
430	0.600	0.052	0	1	880	0.691	0.104	0	6
440	0.926	0.053	0	1	890	0.979	0.105	0	6
450	0.083	0.054	0	1	900	0.887	0.106	0	6

shape. Hence, a piecewise linear $\Lambda(t)$ can be viewed as an approximation to the true $\Lambda(t)$, which cannot be derived from the available data.

For this special case, the interarrival times within each piece of the piecewise linear curve must have an exponential distribution, with mean $1/\lambda_j$, where λ_j is the arrival rate for piece j. Let T_j represent the time that piece j begins. The simulation proceeds as follows.

0. Set $j = 1$ to simulate arrivals over the first period.

1. Simulate a series of exponential random variables, mean $1/\lambda_j$: $X_1, X_2, \ldots$. Sum the exponential random variables to obtain the arrival times for period j: $Y_{1j} = T_j + X_1$, $Y_{2j} = Y_{1j} + X_2$, $Y_{3j} = Y_{2j} + X_3, \ldots$

2. Stop the simulation of piece j when $Y_{nj} > T_{j+1}$.
Increment j to $j + 1$, and return to step 1.

The simulation is completed by catenating the arrival times within the periods.

A key difference between this simulation and the simulation of a stationary Poisson process is that the first arrival within a piece does not equal X_1 plus the time of the previous arrival. Instead, it equals X_1 plus the time the piece began. The validity of this approach relies on the memoryless property of the exponential distribution. For piecewise linear arrival rates, the distribution of the time until the first arrival in a time interval does not depend on the elapsed time since the last arrival.

Example

According to historical records, customers arrive by a nonstationary Poisson process at a small post office at the rate of 5/hour between 11:00 and 12:00 and the rate of 10/hour between 12:00 and 1:00. The simulation proceeds as follows:

Piece 1: 11:00–12:00

n	1	2	3	4	5	6
U_n	.5528	.6105	.9726	.8983	.1614	.0180
X_n (min)	7.1	5.9	.3	1.3	21.9	48.2
Y_n (min)	7.1	13.0	13.3	14.6	36.5	—

Piece 2: 12:00–1:00

n	6	7	8	9	10	11	12	13	14	15	16
U_n	.7521	.8912	.1927	.1579	.7268	.6218	.4875	.5608	.5403	.3921	.0480
X_n (min)	1.7	.7	9.9	11.1	1.9	2.9	4.3	3.5	3.7	5.6	18.2
Y_n (min)	61.7	62.4	72.3	83.4	85.3	88.2	92.5	96.0	99.7	105.3	—

The simulation produced 15 customers.

6.4 FLUID APPROXIMATIONS: SHORT SERVICE TIME

A nonstationary Poisson process encounters two types of variation: random variation and predictable variation. The *predictable* variation is reflected in the function $\Lambda(t)$, which

gives the expected number of arrivals as a function of time. The *random* variation is reflected in the precise arrival times. Both are revealed in Fig. 6.4, which compares $\Lambda(t)$ to a simulation of $A(t)$ for the restaurant queue. While the simulation follows the same general pattern as $\Lambda(t)$, it is susceptible to minor perturbations reflecting random variations.

Because the number of arrivals in any time interval has a Poisson distribution, the mean, $\Lambda(t)$, must equal its variance. This means that the coefficient of variation (ratio of standard deviation to mean) is the following:

$$C[A(t)] = \frac{\sqrt{\Lambda(t)}}{\Lambda(t)} = \frac{1}{\sqrt{\Lambda(t)}} \tag{6.17}$$

Equation (6.17) states that the larger the value of $\Lambda(t)$, the smaller will be the random variations in the number of arrivals (in relation to the expected number of arrivals). For busy queueing systems, these random variations may be of minor importance relative to the predictable variations. To take an extreme example, a busy freeway toll plaza may have 8000 arrivals per hour, which would provide a coefficient of variation of just .011 for 1 hour. This means that a nonstationary Poisson arrival pattern can be accurately approximated with a ***deterministic*** model. The word "determinism" represents a belief that every event is the inevitable consequence of its antecedents. In the context of arrivals, determinism means that $A(t)$ is assumed to be known with certainty and equals $\Lambda(t)$. Though $\lambda(t)$ is not constant, the variations are entirely predictable.

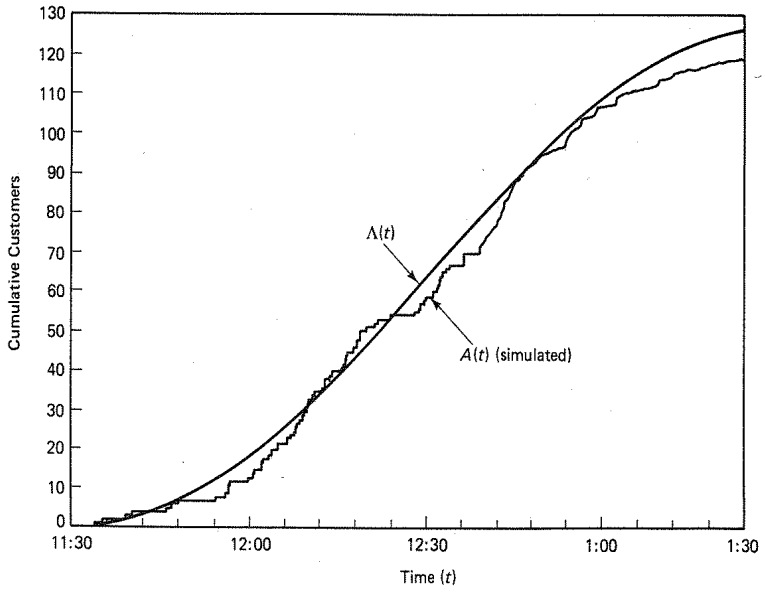

Figure 6.4 Comparison of arrival simulation to expected arrivals at restaurant.

Deterministic queueing models usually fall in the category of ***fluid approximations***. Whereas customers are discrete entities, fluids are not. If a unit of fluid is divided into any proportion, the result will be a smaller quantity of the same entity—it is still a fluid. The same cannot be said of customers, for if a group of customers is divided into proportions, the result may no longer be a group of customers. Half a customer is not a customer at all. Customers are not infinitely divisible. Nevertheless, a quantity of customers can be approximated by a continuous variable (particularly if the quantity is large) and modeled as a fluid. If $\Lambda(t)$ equals 155.3, little harm is caused if $A(t)$ is also assumed to equal 155.3.

Deterministic fluid models are much simpler to use than simulation and also provide more meaningful results. They highlight the important relationships between system attributes and system performance. They should be used when random variation is small relative to predictable variation.

A useful way to think of the fluid queueing model is in terms of the illustration in Fig. 6.5. A faucet deposits water into a tub, and a drain empties water from the tub. The tub represents the queue, and the water represents customers. The arrival rate is the rate at which the water flows out of the faucet into the tub, and the service rate is the capacity of the drain for emptying water from the tub. If water enters the tub faster than it is drained, then the water level in the tub rises. Analogously, if customers arrive faster than they are served, the queue grows. And, if water is drained faster than it enters, the water level will drop, until no water is left in the tub.

Suppose that vehicles arrive at a freeway toll plaza according to the curve $A(t)$ in Fig. 6.6. Based on a fixed service capacity, we would like to determine $D_q(t)$ and $D_s(t)$, cumulative departures from queue and from system, respectively. However, because service times are assumed to be small, $D_q(t) \approx D_s(t)$, and we will be concerned only with the former.

Definition 6.5

c = combined service capacity among all servers (constant rate over time)

Suppose that vehicles are served at the plaza at the rate of $c = 3600$ vehicles per hour. Then the deterministic fluid approximation for $D_q(t)$ has the appearance shown in

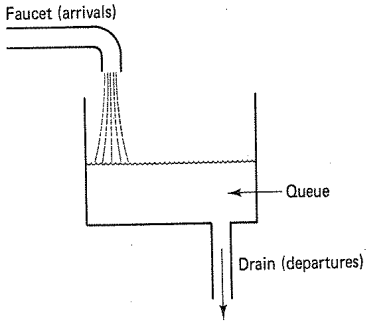

Figure 6.5 In a fluid model, the customers can be viewed as a liquid that accumulates in a tub. Queues increase when the fluid enters the tub faster than it leaves.

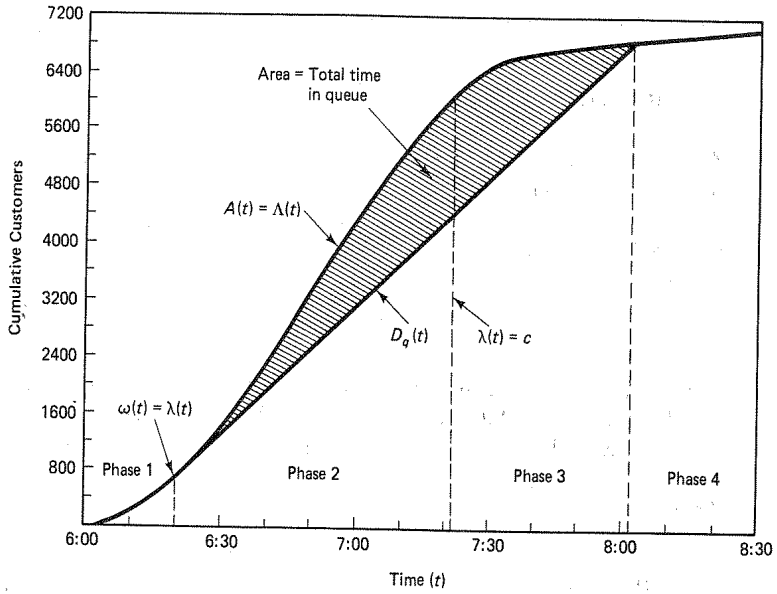

Figure 6.6 Cumulative diagram illustrating deterministic fluid model. When a queue exists, customers depart at a constant rate. Queues increase when the arrival rate exceeds the service capacity and decrease when the service capacity exceeds the arrival rate.

Fig. 6.6. Between the time 6:00 and 6:20 A.M., $D_q(t)$ is identical to $A(t)$ because vehicles can be served at a faster rate than they arrive. At 6:20 A.M., the arrival rate (slope of A(t)) has increased to the point where it exactly equals the service capacity. From this point on, the queue grows because vehicles arrive at a faster rate than they are served. The queue finally begins to decline when the arrival rate again equals the service capacity, which occurs at about 7:20. The queue eventually vanishes at about 8:00 A.M.

To draw $D_q(t)$, identify the point where $\lambda(t)$ first exceeds c (that is, the first point where the slope of $A(t) = c$, 6:20 in Fig. 6.6). From this point, draw a line tangent to $A(t)$, with slope c, forward until it again intersects $A(t)$ (8:00 in Fig. 6.6). From this second point onward, $A(t) = D_q(t)$ until such time that $\lambda(t)$ again exceeds c.

Over the period from 6:00 A.M. to 8:30 A.M., the system passes through four phases, which are identified as follows:

Phase 1: Stagnant

$$A(t) = D_q(t) \qquad \lambda(t) \leq c \qquad \frac{dL_q(t)}{dt} = 0 \tag{6.18}$$

Phase 1 represents the period from 6:00 to 6:20, when vehicles are served as fast as they arrive.

Phase 2: Queue Growth

$$A(t) > D_q(t) \qquad \lambda(t) > c \qquad \frac{dL_q(t)}{dt} = \lambda(t) - c > 0 \tag{6.19}$$

Phase 2 represents the period from 6:20 A.M. to 7:20 A.M., when vehicles arrive at a faster rate than they are served and the queue grows.

Phase 3: Queue Decline

$$A(t) > D_q(t) \qquad \lambda(t) < c \qquad \frac{dL_q(t)}{dt} = \lambda(t) - c < 0 \tag{6.20}$$

Phase 3 represents the period from 7:20 to 8:00 when customers arrive at a slower rate than they are served and the queue shrinks. Note that a queue can exist even when the arrival rate is smaller than the service capacity.

Phase 4: Stagnant

$$A(t) = D_q(t) \qquad \lambda(t) < c \qquad \frac{dL_q(t)}{dt} = 0 \tag{6.21}$$

The last phase begins at 8:00, when the queue finally vanishes.

In all four phases, the arrival rate and the service capacity determine the *rate* at which the queue grows or shrinks. When is the queue the largest? At the end of the phase 2, when the arrival rate equals the service capacity:

$$\text{Queue is largest when: } \lambda(t) = c \tag{6.22}$$

This is the time when the queue stops growing and begins shrinking. When does the queue vanish? When $A(t) = D_q(t)$. Note that the arrival rate can be much smaller than the service capacity when this happens.

Definition 6.6

$\omega(t)$ is the ***departure rate*** at time t

$$\omega(t) = \begin{cases} c & L_q(t) > 0 \\ \lambda(t) & L_q(t) = 0 \end{cases} = \frac{dD_q(t)}{dt}$$

The function $\omega(t)$ is the rate at which customers are departing at time t, whereas c is the *capacity* for serving customers. If there are zero customers in the queue, customers can depart no faster than the rate at which they arrive. $\omega(t)$ is also the slope (derivative) of $D_q(t)$. Figure 6.7 plots $\lambda(t)$ and $\omega(t)$ for the vehicle queue. In phases 1 and 4, $\omega(t) = \lambda(t)$, and in phases 2 and 3, $\omega(t) = c$. The division between phases 2 and 3 occurs when $\lambda(t) = \omega(t)$ (7:20), which signals the point where the queue begins to decline. By the time the queue vanishes (8:00), $\lambda(t)$ has dropped far below c. This creates a discontinuity in $\omega(t)$,

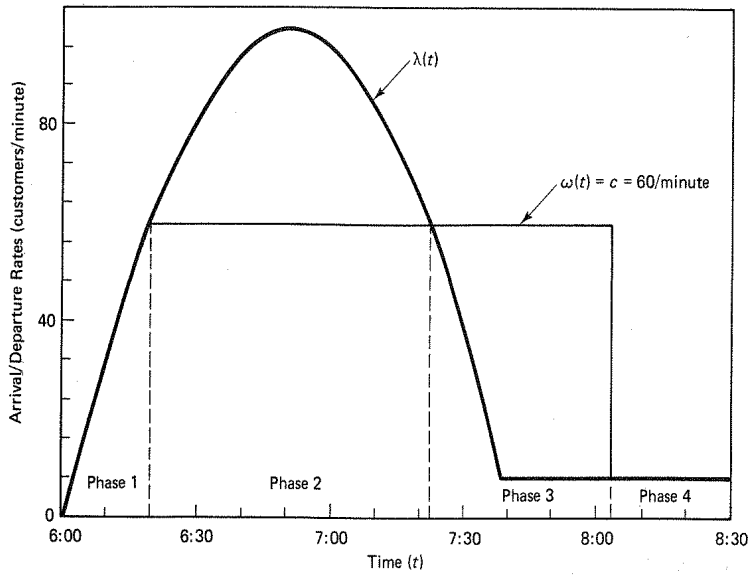

Figure 6.7 Arrival and departure rates versus time for a deterministic fluid model. When queue is at its maximum, the arrival and departure rates are equal. By the time the queue vanishes, the arrival rate is much less than the departure rate.

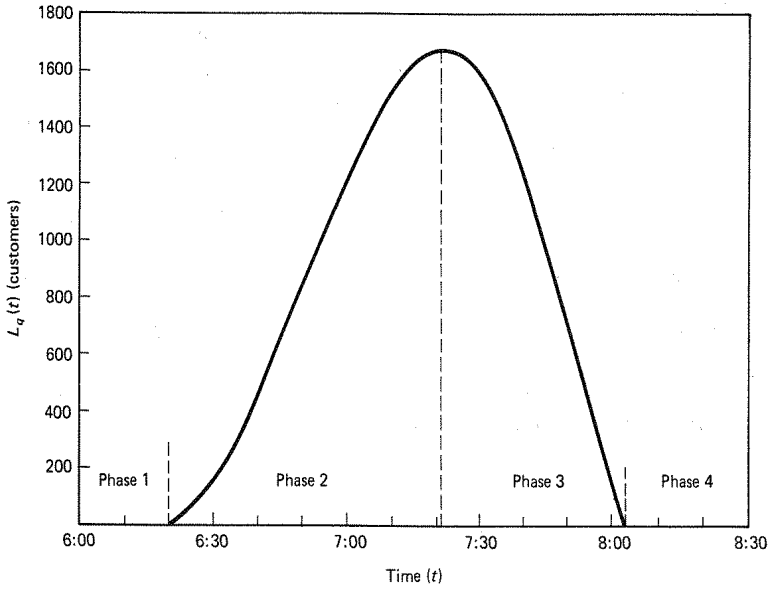

Figure 6.8 Customers in queue versus time for a deterministic fluid model. Queue size grows gradually at beginning, but declines rapidly at end.

as it shifts from the service capacity, c, to the arrival rate. These facts are further reflected in Fig. 6.8, which shows $L_q(t)$. Note that the queue grows gradually at first, but at the end, the decline is swift, as evidenced by the intersection of $L_q(t)$ with the horizontal axis.

These phenomena are in no way unique to vehicle queues. The features of queue growth when $\lambda(t) > c$, queue shrinkage when $\lambda(t) < c$, maximum queue when $\lambda(t) = c$, and discontinuity in $\omega(t)$ when the queue vanishes are universal. From the server's perspective, the end of the queue is quite dramatic, as the transition occurs from the service capacity to a very small departure rate.

The performance measures are calculated from the cumulative arrival and departure curves in the exact same manner as if they represented empirical observations (see Chap. 2). For example, the area between $A(t)$ and $D_q(t)$ is the total time spent in queue, and this area divided by the number of customers served is the average time in queue. The average number of customers in queue is the area between the curves divided by the length of time studied.

6.5 FLUID APPROXIMATIONS: LARGE SERVICE TIME

So far, no distinction has been made between departures from the system and departures from the queue. Earlier, we saw that $D_s^{-1}(n) = D_q^{-1}(n) + S(n)$, where $S(n)$ is the service time for customer n (that is, departure time from the system equals departure time from the queue plus the service time). With a deterministic approximation, $S(n)$ would be the inverse of the service capacity multiplied by the number of servers, m/c. This means that $D_s(t)$ is the same as $D_q(t)$, except that it is shifted to the right by m/c. In tollbooth, highway, and ticket window queues, the service time is very small (about 6 seconds), meaning virtually all the waiting time is spent in queue. Nevertheless, the queue is still caused by inadequate server capacity, and any effort to eliminate delay should be focused there. One should not try to eliminate the queue by expanding the capacity for storing queueing vehicles (this has been tried on highways).

Service times are not always small relative to time in queue, and $D_q(t)$ cannot always be assumed to equal $D_s(t)$. Figure 6.9 shows an arrival curve with the same shape as Fig. 6.6; however, the axes have been rescaled. The total number of arrivals over 2 1/2 hours is now 105. The service process consists of 20 servers, with a service time of 20 minutes per customer. Thus, the 20 servers can process customers at the rate of one per minute.

The procedure for constructing the departure curves is more complicated than it was for a short service time. It follows from the following two conditions, which must be satisfied at all times:

$$1.\ D_s(t + m/c) = D_q(t) \tag{6.23}$$

$$2.\ D_q(t) = \min\{A(t), D_s(t) + m\} \tag{6.24}$$

The fact that $D_s(t)$ depends on $D_q(t)$ and $D_q(t)$ depends on $D_s(t)$ suggests why the procedure is difficult. The procedure will be presented by way of an example. The segments below refer to portions of the curves $D_q(t)$ and $D_s(t)$, as indicated in Fig. 6.9.

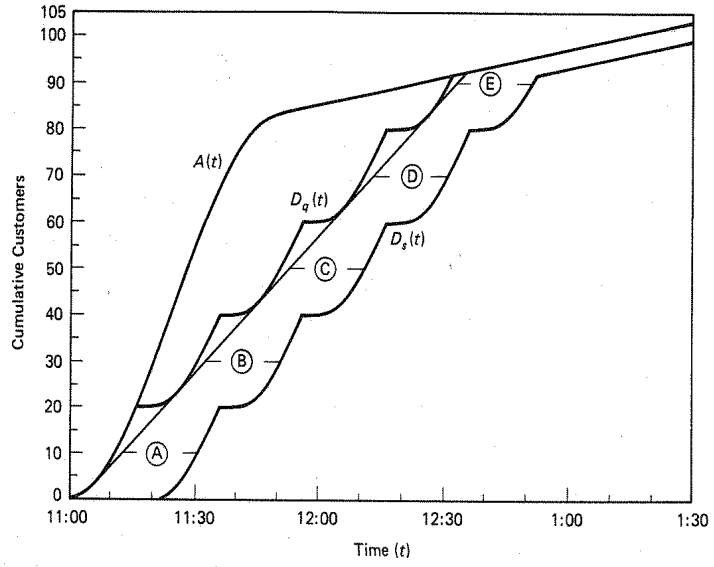

Figure 6.9 Deterministic fluid model with a long service time. Ripple pattern mirrors the arrival pattern immediately preceding the formation of a queue.

Segment A Until $A(t) = m$ ($t < 11{:}15$), customers enter service immediately and $D_q(t) = A(t)$. From condition 1, $D_s(t + m/c) = D_q(t) = A(t)$. That is, in segment A, $D_s(t)$ has the same shape as $A(t)$, but it is shifted right by 20 minutes.

Segment B At 11:15, 20 customers have arrived, but no customer has yet left the system, which means that a queue begins to form. $D_q(t)$ is now determined from the second part of condition 2: $D_q(t) = D_s(t) + m$. That is, segment B of $D_q(t)$ is segment A of $D_s(t)$ shifted upward by $m = 20$. From condition 1, segment B of $D_s(t)$ is found by shifting segment B of $D_q(t)$ to the right by 20 minutes.

The same procedure is repeated for segments C and D. Segment E is somewhat different.

Segment E From 12:15 until 12:30, $D_q(t) = D_s(t) + 20$. At 12:30, $D_q(t)$ intersects $A(t)$, meaning that the queue has vanished. From this point on, $D_q(t)$ is defined by the first part of condition 2: $D_q(t) = A(t)$. $D_s(t)$, as usual, is found by shifting $D_q(t)$ to the right by 20 minutes.

The ripple pattern in $D_q(t)$ and $D_s(t)$ is a distinct feature of the model. Customers are served in spurts, which parallel the arrival pattern in the first 15 minutes. These spurts alternate with 5-minute lulls, when $D_q(t)$ remains constant. The spurt/lull cycle occurs because a new batch of customers cannot begin service until the previous batch has

completed service. This contrasts with the short service time fluid model, shown as a straight line with the slope of 1 customer/minute in Fig. 6.9. Note that the short service time model approximates $D_q(t)$, except that the ripples are eliminated. The smaller the service time, the more similar the two departure curves will be.

In reality, random perturbations in service times will tend to smooth out the ripples in $D_q(t)$ and $D_s(t)$ over time, as is reflected in the queueing simulation provided in Fig. 6.10, based on the same set of data. The service time distribution is assumed to be normal with a mean of 20 minutes and a standard deviation of 5 minutes. The arrival process is nonstationary Poisson. The ripples are evident at the beginning of the simulation, but are later smoothed. Despite the absence of ripples, the general departure patterns of the simulation and fluid model are nearly the same.

6.6 ADJUSTMENTS TO DETERMINISTIC APPROXIMATION

Clearly, the validity of the deterministic approximation depends on the variability in service and interarrival times. Whereas the approximation predicts that queues do not begin to form until $\tilde{\rho}(t) = \lambda(t)/c(t) > 1$, we already know from steady-state analysis that random queues will exist when $\tilde{\rho}(t) < 1$. The approximation should somehow account for these random queues.

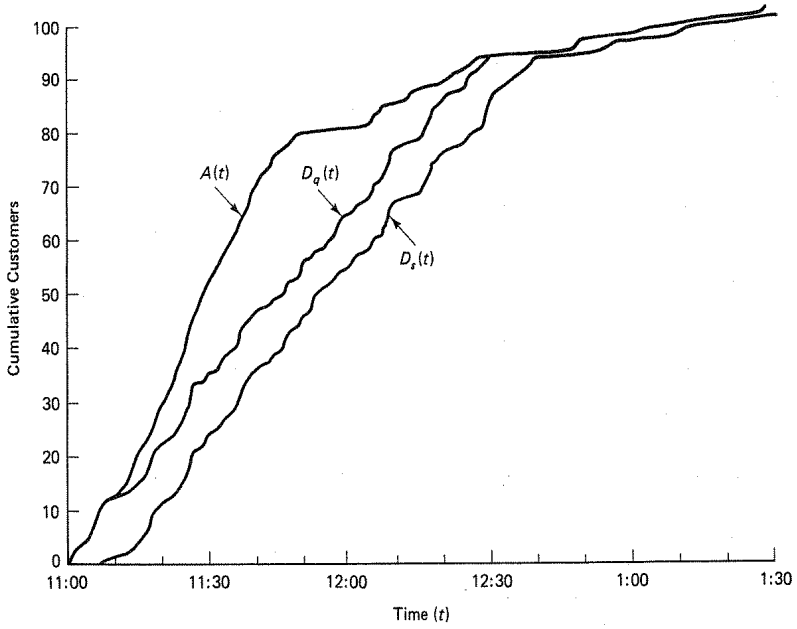

Figure 6.10 Simulation of a queueing system with long service time. Arrival and departure curves are similar to the deterministic fluid model (Fig. 6.9).

As $\tilde{\rho}(t)$ increases, from a value much less than 1 to a value much greater than 1, the expected queue length will pass through three stages:

Stage 1: $\tilde{\rho}(t) << 1$: In this stage, the quasi-steady-state model is valid, and provides a good prediction for queue length.

Stage 2: $\tilde{\rho}(t) \leqslant 1$, $(1 - \tilde{\rho}(t))$ is small: In this stage, queue lengths are difficult to predict. The quasi-steady-state model is not valid. Neither is the deterministic approximation, for it predicts zero queue lengths.

Stage 3: $\tilde{\rho}(t) > 1$: In this state, the *growth* in expected queue length is accurately predicted by the deterministic approximation.

In stage 2, two ways to predict queue length are ***diffusion models*** (see Newell, 1982, and the appendix to this chapter) and *simulation*. Neither approach is simple. In stage 3, queue length can be predicted from the *deterministic approximation*, provided that an estimate is made of the queue length when $\tilde{\rho}(t) = 1$. Newell provides the following approximation for nonstationary Poisson arrivals and independent service times:

$$E[L_q(t')] \approx \left[\left(\frac{1}{[1 + C^2(S)]^2}\right)\left(\frac{1}{c}\right)\left(\frac{d\tilde{\rho}(t')}{dt}\right)\right]^{-1/3} \tag{6.25}$$

where

$C(S)$ = the coefficient of variation in the service times
t' = the time when $\lambda(t) = c$

Example:

For the arrival curve in Fig. 6.7, $d\tilde{\rho}(t')/dt = (d\lambda(t)/dt)/c$, evaluated at 6:20, which is approximately 4/60 per minute ($c = 60$ customers/minute). If, for example, the coefficient of variation in the service times is .5, then

$$E[L_q(t')] \approx \left[\frac{1}{(1 + .25)^2}\ \frac{1}{60}\ \frac{4}{60}\right]^{-1/3} = 11.2$$

Thus, we would expect to have about 11 vehicles in the queue at time t', and $L_q(t)$, $t \geqslant t'$, would be shifted upward accordingly.

By comparison to the maximum queue length predicted by the deterministic model, $E[L_q(t')]$ appears to be small. The correction would have been much larger if the arrival rate changed slowly, which would have allowed the number of customers in the queue to approach the equilibrium steady-state distribution prior to t' (note that $E[L_q(t')]$ grows without bound as $d\tilde{\rho}(t')/dt$ approaches zero). Keep in mind that even though 11 vehicles is a relatively small number, its impact persists throughout the queueing period. It shifts the entire curve $L_q(t)$ upward by 11, not just a small portion. Thus, even a relatively small increase in queue length at or near t' can lead to large delays later on.

6.7 QUEUEING TO MEET A SCHEDULE

A good queueing model should not only replicate the behavior of a queueing system; it also should be capable of predicting the future behavior of the system should some change be made. If the service capacity is improved by adding servers, it should be able to predict the reduction in the waiting time. So far, we have assumed that the arrival pattern is a "given." This assumption deserves further examination.

For the sake of simplicity, it is worthwhile to consider two types of customers. The first customer will be labeled the *arrive when ready* customer. Take a post office queue as an example. To the postal customer, arrival time may not be influenced by queue lengths at various times of the day, because he or she arrives whenever he or she is ready to conduct business. The arrival pattern is fixed. Now consider a second type of customer. The *depart on schedule* customer can arrive at any time, but must make sure that he or she departs prior to a scheduled deadline. Take a commuter as an example. Commuters do not leave home "when ready"; commuters leave home at times that guarantee that they will arrive in time for work. Thus, the arrival time *is* influenced by the queue lengths encountered on the trip to work, for if queues become long, he or she will have to leave home earlier to guarantee that he or she "departs on schedule" (arrives at work on time).

By no means is this the only example of a customer that aims to depart on schedule. Perhaps the most pervasive example is in manufacturing, where companies place orders with their suppliers months in advance to ensure that components arrive when needed (that is, depart from the *supplier's* queue on schedule). Other examples include arrivals at sporting events and for airline flights.

A way to visualize the two customer types is that the "arrive when ready" customer is constrained to arrive at a certain time, and the "depart on schedule" customer is constrained to depart from the system at a certain time. The impact of a system improvement depends on which type (or types) of customer is using the system. In the former case, a change in service capacity will have no impact on customer arrival times. In the later case, a change in service capacity may encourage customers to change their arrival times.

Modeling customer behavior is a difficult thing. But, hypothetically, consider the consequences if the queue operator were to schedule "depart on schedule" customer arrivals with the objective of minimizing total waiting time. Then the cumulative arrival pattern would resemble Fig. 6.11. The operator would schedule arrivals at a constant rate, at times prior to the desired departure times. Although this pattern requires customers to depart from the system before desired, queues are eliminated (there is no way to serve all customers on time without having some depart from the system early).

Unfortunately, the optimal schedule is not how customers would actually arrive. Note in Fig. 6.11 that if any customer were to change unilaterally its arrival time, other customers would be forced to arrive late. Yet, one cannot deny that there is a strong incentive to do just that. A customer who arrives at 7:00 departs 30 minutes earlier than desired. Because the arrival pattern eliminates queues, this customer could easily change its arrival time to 7:30 and still depart on time. Therefore, if customers act individually in their own self-interest, then the operator's arrival pattern would be *unstable*. That is, the pattern could not be sustained because customers will change their departure times.

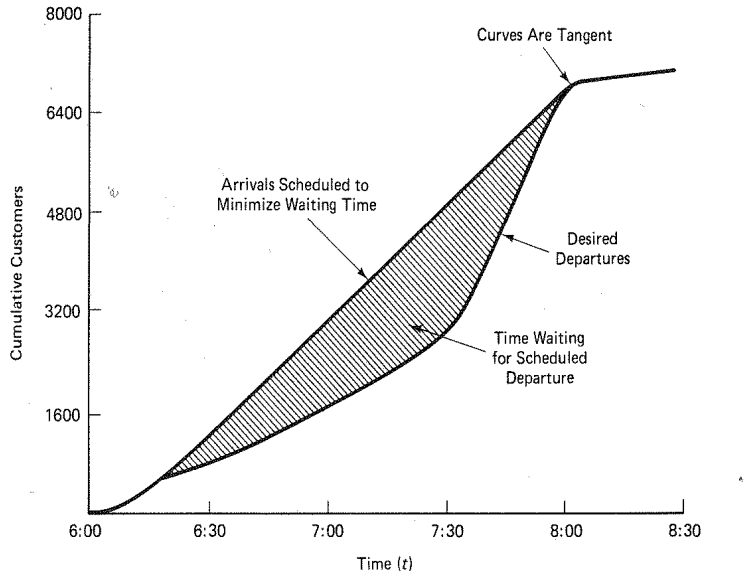

Figure 6.11 Customers must arrive early when they are required to depart from the queue on schedule. Ideally, customers should arrive at a constant rate and queues should not materialize.

A *stable* arrival pattern would be more like Fig. 6.12 (see Daganzo 1985; Hendrickson and Kocur 1981; and Newell 1987). Customers may at first choose to arrive later than the operator's schedule. But, with many customers changing their arrival times, queues will materialize and customers will be late. To compensate, the next time that customers arrive at the queue they will have to arrive earlier—even earlier than the original scheduled arrivals—and the end result: *Customers will not only depart from the queue earlier than desired, but they will also incur queueing delay.*

This is but one example of how queues are created by customers acting in their own self-interest, rather than working together for their common benefit. In Chap. 8, methods are presented for resolving such problems.

6.8 CHAPTER SUMMARY

The most severe queueing problems do not result from random variations in arrival times but, rather, from predictable variations in arrival rates. These predictable variations can be represented by the nonstationary Poisson process. Like the standard Poisson process, the nonstationary Poisson process possesses independent increments, and the probability of an arrival in a differential time increment equals $\lambda(t)dt$. Unlike the standard Poisson process, $\lambda(t)$ does not have to stay constant over time.

The arrival rate represents the expected number of arrivals to occur per unit time. The integral of $\lambda(\tau)$ from 0 to t, $\Lambda(t)$, equals the expected number of arrivals to occur by

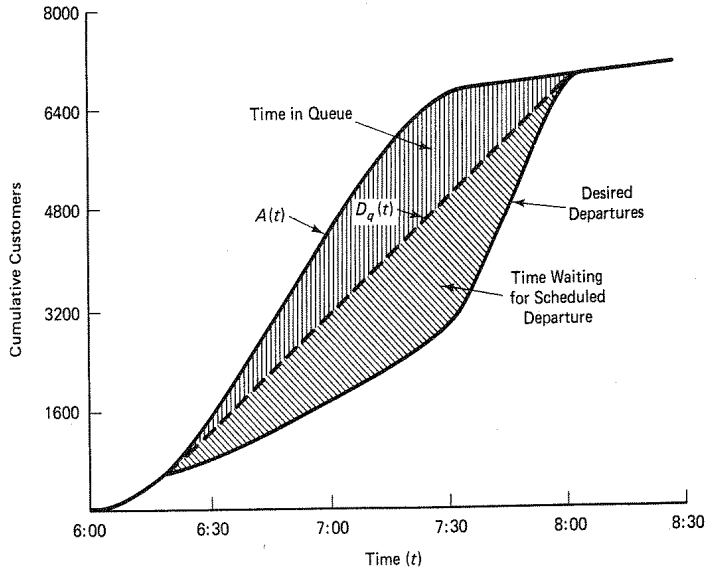

Figure 6.12 In actuality, when customers must "depart on schedule," customers will arrive earlier than necessary and queues will materialize.

time t. The actual number of arrivals to occur is a random variable, which can be more or less than $\Lambda(t)$. Like the standard Poisson process, this random variable has a Poisson distribution with mean $\Lambda(t)$. Nevertheless, the nonstationary Poisson process does not have exponential interarrival times and it is not memoryless. Among other things, this makes it difficult to determine whether a data set was or was not produced by a nonstationary Poisson process. Goodness of fit is primarily judged by plausibility.

A nonstationary Poisson process can be simulated in any of three ways: by simulating a stationary Poisson process and accepting arrivals with probability $\lambda(t)/\lambda_{max}$; by first generating a Poisson random variable representing the number of arrivals over a time interval and then simulating the arrival times; or by simulating Bernoulli random variables, representing whether or not arrivals occur in small time increments. If $\Lambda(t)$ is piecewise linear, arrivals can also be simulated by generating exponential random variables within each piece of the piecewise linear curve (see Sec. 6.3.3).

If the arrival rate is nonstationary, and much less than the service capacity, queueing behavior can be modeled with quasi-steady-state equations. When the function $\tilde{\rho}(t)$ is substituted for ρ/m, these equations provide an estimate for the queue's performance at any time t.

In most cases, the deterministic fluid model is the best way to visualize a nonstationary queueing system. The customer is represented by a fluid, which enters from a faucet (the arrival process) into a tub (the queue) and later empties from a drain (the service process). The queue grows whenever the arrival rate exceeds the service capacity, reaches a maximum when the arrival rate equals the service capacity, and declines when the arrival rate falls below the service capacity.

The fluid model does not account for random fluctuations in the arrival and service processes, just predictable fluctuations. These predictable fluctuations tend to overwhelm random fluctuations in arrival times. In Sec. 6.6, an adjustment factor was provided to estimate the size of the queue created by random fluctuations prior to the time $\tilde{\rho}(t)$ first exceeds 1. This adjustment factor is negligible when the rate of change in $\tilde{\rho}(t)$ is large.

Queues are most difficult to analyze when the arrival rate hovers at or near the service capacity, occasionally falling below or occasionally rising above. Neither steady-state analysis nor fluid approximations provide adequate predictions, the first overestimating queue lengths and the latter underestimating queue lengths. The most robust way to analyze such systems is through simulation.

FURTHER READING

Box, G. E. P. and G. M. Jenkins. 1970. *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

Daganzo, C. F. 1985. "The Uniqueness of a Time-Dependent Equilibrium Distribution of Arrivals at a Single Bottleneck," *Transportation Science*, 19, 29–37.

Fishman, G. S. 1973. *Concepts and Methods in Discrete Event Digital Simulation*. New York: John Wiley.

Hall, R. W. 1987. "Passenger Delay in a Rapid Transit Station," *Transportation Science*, 21, 279–292.

Hendrickson, C., and G. Kocur. 1981. "Schedule Delay and Departure Time Decisions in a Deterministic Model," *Transportation Science*, 15, 62–77.

Horonjeff, R. 1969. "Analyses of Passenger and Baggage Flows in Airport Terminal Buildings," *Journal of Aircraft*, 6, 446–451.

Kleinrock, L. K. 1976. *Queueing Systems, Vol. 2: Computer Applications*. New York: John Wiley.

Naddala, G. S. 1977. *Econometrics*, New York: McGraw-Hill.

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

———. 1987. "The Morning Commute for Nonidentical Travelers," *Transportation Science*, 21, 74–88.

Theil, H. 1971. *Principles of Econometrics*. New York: John Wiley.

PROBLEMS

1. Customers arrive at a cafeteria by a nonstationary Poisson process, with rates:

$$\lambda(t) = \begin{cases} 5/\text{hr} & 8{:}00\text{–}12{:}00 \\ 20/\text{hr} & 12{:}00\text{–}1{:}00 \\ 10/\text{hr} & 1{:}00\text{–}3{:}00 \end{cases}$$

(a) Write the function $\Lambda(t)$.
(b) What is the probability that exactly 30 customers arrived between 12:00 and 3:00?
(c) Suppose that 20 customers are known to have arrived between 8:00 and 1:00. What is the probability that no one arrived before 12:00? What is the probability that two customers arrived before 12:00?
(d) Suppose that the last customer to arrive before 12:00 arrived at 11:57. What is the probability that no customer arrived between 12:00 and 12:07?

2. The arrival rate in Prob. 1 is now defined by the function

$$\lambda(t) = 8 \sin(\pi t/7) \qquad 0 \leq t \leq 7$$

where $t = 0$ is equivalent to 8:00, time is measured in hours, and the angle is measured in radians. Repeat parts a–c from Prob. 1.

3. Figure 6.13 provides a cumulative arrival curve. From this curve, draw (approximately) $\lambda(t)$ on a piece of graph paper. At what time is the arrival rate the largest, and what is the largest arrival rate?

4. Figure 6.14 provides an arrival rate curve. From this curve, draw (approximately) $\Lambda(t)$ on a piece of graph paper.

5. The service time distribution at a single server queue is normal, with mean 5 minutes and variance 4 minutes^2. The arrival process is nonstationary Poisson, with the following arrival rate:

$$\lambda(t) = 50e^{.05t}/(100 + e^{.05t})^2 \qquad \lambda(t) \text{ in cust/min, } t \text{ in min.}$$

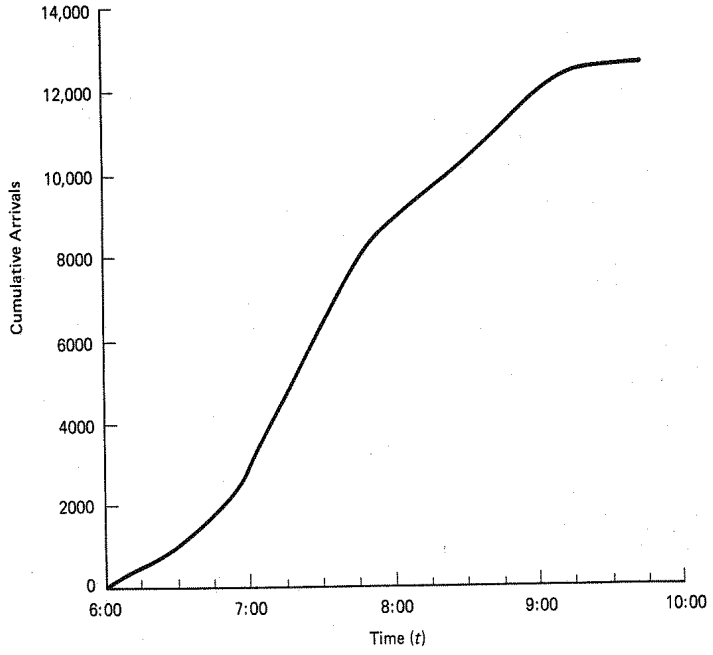

Figure 6.13 Example cumulative arrival curve.

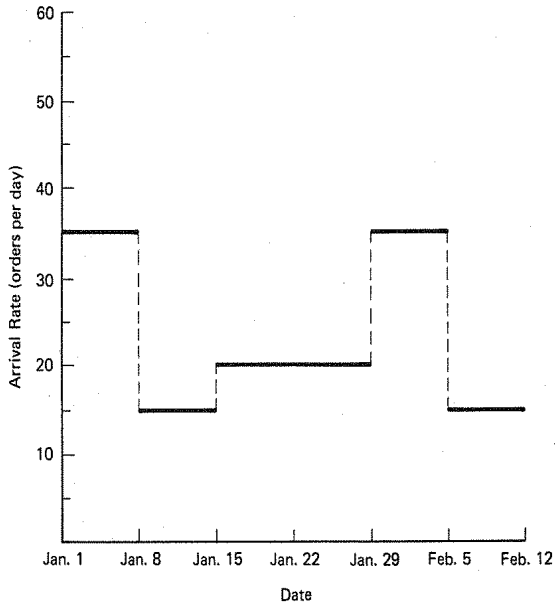

Figure 6.14 Example arrival rate curve.

(a) Use the quasi-steady-state approximation to write the function $L_q(t)$.
*(b) Using a computer, plot the function $L_q(t)$ over the interval from $t = 0$ to 200 minutes. At what time is $L_q(t)$ the largest, and what is the largest value of $L_q(t)$?

***6.** For Prob. 5, determine whether the quasi-steady-state approximation is valid at all times (a computer may be helpful on this problem). If not, describe how the true performance would differ from the predicted performance.

7. Repeat Prob. 5, using the arrival rate function in Prob. 1 (over 8:00 to 3:00 interval) and two servers, each with exponential service time distribution and mean service time of 5 minutes. In words, discuss whether you believe your approximation is valid.

8. Suppose that the function given in Prob. 1 is based on the average number of arrivals over 20 days. Give 95% confidence intervals for the arrival rate in each interval.

9. Using a computer, simulate the arrival process in Prob. 1 by each of the following means:
(a) By generating a Poisson random variable, then generating the individual arrival times
(b) By generating exponential random variables
(c) By simulating a stationary Poisson process and randomly accepting or rejecting arrivals

***10.** Using a computer, simulate the queueing system described in Prob. 5, for one 200-minute period. Plot queue length as a function of time and compare your result to $L_q(t)$. Are your results reasonable?

*Difficult problem

11. Figure 6.14 gives a cumulative arrival curve. Copy the figure on a piece of graph paper. Using a deterministic fluid approximation, with service rate of 4800 customers/hour and short service time, do the following:
- **(a)** Draw the cumulative departure curve.
- **(b)** Indicate when the queue is largest, and the largest queue size.
- **(c)** Indicate when the queue is growing at the fastest rate. Estimate the maximum rate of growth.
- **(d)** Indicate when the queue begins and vanishes.
- **(e)** On a separate graph, show queue length as a function of time.
- **(f)** Estimate total waiting time among all customers.

***12.** A forms processing center has a constant service time of 15 minutes, and 12 servers operate. Copy Fig. 6.14 on a piece of graph paper, and rescale the vertical axis by dividing by 100. Based on the service process and the arrival curve, repeat parts a–f from Prob. 11.

***13.** For Prob. 11, suppose that the coefficient of variation in the service time is .5.
- **(a)** Using the adjustment factor in Sec. 6.6, estimate the queue size at the time $\lambda(t) = c$. Assume that $d\lambda(t)/dt = 20{,}000$ customers/hour^2 at the time $\lambda(t) = c$.
- **(b)** Draw a graph representing $L_q(t)$ for the period after $\lambda(t)$ first equals c. Compare and contrast your result to Prob. 11.
- **(c)** Identify (approximately) the time period over which a quasi-steady-state model would be valid.
- **(d)** Using the quasi-steady-state model, and assuming an $M/G/1/\infty$ queue, plot $L_q(t)$ over the period over which it is valid.
- **(e)** Discuss how your answers to parts b and d might be used to estimate $L_q(t)$ over the entire time range.

***14.** In Prob. 11, suppose that on one out of ten days the server malfunctions and can only process customers at the rate of 3600/hour. Estimate the total waiting time among all customers.

15. Figure 6.14 represents the cumulative number of jobs that a printer must deliver by time t. The company would like to begin processing jobs as late as possible, yet still deliver them on time. Jobs can be processed at the rate of 5500 per hour.
- **(a)** On a piece of graph paper, draw the curve representing cumulative job completions.
- **(b)** At what time is the number of jobs that have been completed, but not delivered, the largest?

16. Customers at a convenience market have a tendency to renege if their wait in line is sufficiently long.

$$P(\text{renege if wait is } t \text{ or less}) = \begin{cases} t/30 & 0 \leq t \leq 30 \text{ minutes} \\ 1 & t > 30 \text{ minutes} \end{cases}$$

Put another way, for each minute the customer waits (up to 30), there is a 1/30 chance of reneging. To take an example, if a customer has to wait 10 minutes, there is a one-third chance of reneging.

Describe in words how your simulation in Prob. 10 could be modified to account for this behavior.

QUEUE EXPERIMENT: FOR IN-CLASS DISCUSSION

The queue experiment can be completed in the privacy of your own home in about 20 minutes. You will need an ordinary bathroom sink (preferably with a screw-type water valve), a ruler, and a

*Difficult problem

timing device. Record your observations, but do not hand them in. Read the entire description before beginning.

Setup. Open the cold water valve to maximum flow, counting the number of revolutions of the handle. Record this number.

Push the drain control to its lowest height (open drain). Open the cold water valve to one-half the maximum flow. Now raise the drain control until you begin to see the water level rising in the sink. Leave the drain control at a level that maintains a constant water level. Place the ruler in the sink in a way that allows you to measure the depth of the water.

Experiment. Throughout the following, record the height of the water at 10-second intervals.

Time (sec)	Action
0	Open the cold-water valve to 3/4 maximum flow
10	Open the cold-water valve to maximum flow
20	Reduce the cold-water flow to 3/4 maximum flow
30	Reduce the cold-water flow to 1/2 maximum flow
40	Reduce the cold-water flow to 1/4 maximum flow
50	Turn the water off

Continue recording the water level until no water is left in the sink. (Absolute accuracy is not essential in this experiment.)

Questions

Based on the flow *only*, plot $A(t)$ and $D_q(t)$ on a piece of graph paper.
Based on the water height recordings *only*, plot $L_q(t)$ on a separate piece of graph paper.
When is $L_q(t)$ the largest?
When does the queue vanish?
Describe the function $\omega(t)$.
Are your data consistent? Explain why or why not.

EXERCISE: NONSTATIONARY MODELING

The purpose of this exercise is to learn how to simulate a nonstationary Poisson process and to use deterministic fluid approximations.

1. One of the local restaurants has been collecting data over the last year on the arrival pattern of customers during lunchtime. The data below provide the number of arrivals from the opening time (11:30) until 2:00.

Time period	Average number of arrivals
11:30–11:45	5
11:45–12:00	30
12:00–12:15	30
12:15–12:30	15
12:30–12:45	5
12:45–1:00	20
1:00–1:15	10
1:15–1:30	5
1:30–1:45	6
1:45–2:00	4
Total	130

Each arrival represents a "party" of customers. The restaurant has 35 tables. The service time (the time from when a party is seated at its table until the table is available to seat the next party) has an exponential distribution, with mean of 30 minutes.

1. Simulate the arrival pattern for one lunchtime (11:30–2:00), based on a nonstationary Poisson process. Plot $A(t)$ on a piece of graph paper.
2. Simulate the departure pattern for one lunchtime, based on the exponential distribution. Plot $D_q(t)$ and $D_s(t)$ on the same paper as $A(t)$.
3. Determine the average number of customers in the system and the average time in system. Answer the following:
   - **(a)** At what time is the queue the longest?
   - **(b)** At what time does queueing begin?
   - **(c)** At what time does the queue vanish?
   - **(d)** At what time is the wait (not counting service time) the longest?
   - **(e)** What is the longest waiting time?
4. Now assume that the arrival pattern is deterministic, as defined by the average arrival rates. Also assume that all customers are served in exactly 30 minutes.
   - **(a)** Draw $A(t)$, $D_q(t)$, and $D_s(t)$ on a single piece of graph paper. (Be careful. $D_q(t)$ and $D_s(t)$ are not easy to draw because of the long service time. Make sure that $D_q(t) - D_s(t) \leq 35$ for all t.)
   - **(b)** Answer all the questions from part 3 for these new curves.
   - **(c)** Explain why, or why not, your answers to part 4b are different from your answers to part 3.
   - **(d)** Under what conditions would the deterministic approximation be more accurate? Explain.

APPENDIX: DIFFUSION MODELS

The term *diffuse* means "to cause to spread or disperse, as a gas or a liquid." The term *diffusion* means the process of diffusing. Diffusion models are used in physics to represent the molecular diffusion of fluids. But diffusion models are also applicable to queues, particularly in the analysis of the stochastic behavior of nonstationary queueing systems. Exact analysis of these systems is extremely difficult, and diffusion models provide both

simple and robust results. As already seen in the chapter text, deterministic fluid models can be used to approximate queue behavior. Stochastic diffusion models can too: The rate at which a fluid diffuses across a region boundary is analogous to the transition rate across a cordon line in a transition rate diagram (see Chap. 5).

The following will refer to two types of diffusion models. The ***diffusion equation*** is a differential equation, first developed for molecular diffusion, but also applicable to queueing systems. A ***diffusion process*** (also called Brownian motion) is a stochastic process in which the interevent times are independent normal random variables. As applied to queueing, the fundamental assumption of the diffusion equation is that the number of arrivals and number of departures behave like diffusion processes and are mutually independent, whenever the queue size is positive. However, just because the arrival and departure processes behave like a diffusion process does not mean that they are diffusion processes. As we will see, other stochastic processes, including the Poisson process, can be approximated by the diffusion process.

The derivation of the diffusion equation, and its role in developing nonstationary results, are beyond the scope of this book. Interested readers are referred to Newell (1982). However, basic concepts can be illustrated with the following example, based on the diffusion process. In the example, it will be assumed that the probability of the queue being size zero is negligible. This might represent a situation where the queue size is initially large and the arrival rate exceeds the service rate. The following analysis is similar to that in Newell (1982).

First, consider the arrival process. Assume that at time 0, no arrivals have yet occurred. Then the time of the nth arrival, $A^{-1}(n)$, must equal the sum of n interarrival times, A_i:

$$A^{-1}(n) = \sum_{i=1}^{n} A_i \tag{6.A1}$$

From the central limit theorem, we know that if n is large and A_i are independent identically distributed, $A^{-1}(n)$ must (approximately) have a normal distribution. For large values of k, this implies that $A^{-1}(ki) - A^{-1}[k(i-1)]$ must also be approximately normally distributed. Because dependencies should be weak (at most) for large k, $A^{-1}(ki)$ should behave like a diffusion process. (This does not mean that $A^{-1}(n)$ behaves like a diffusion process.) A similar argument can be made for the departure process, $D_s^{-1}(ki)$, based on the assumption that the queue does not vanish. $D_s^{-1}(n)$ is the sum of n service times. If the service times are identically distributed, $D_s^{-1}(ki)$ will behave like a diffusion process. For an FCFS queue, the fact that $D_s^{-1}(n)$ and $A^{-1}(n)$ are approximately normal means that confidence intervals can be obtained for the waiting time of the nth customer.

Of greater interest than the processes $D_s^{-1}(ki)$ and $A^{-1}(ki)$ are the processes $D_s(t)$ and $A(t)$. The latter are related to the former. In fact, it is possible to describe events in either of two ways:

$$A^{-1}(n) \leq \tau \leftrightarrow A(\tau) \geq n \tag{6.A2}$$

The left side states that the arrival curve intersects the horizontal line, passing through the point n, at or to the left of time τ in Fig. 6.15. The right side states that the arrival curve

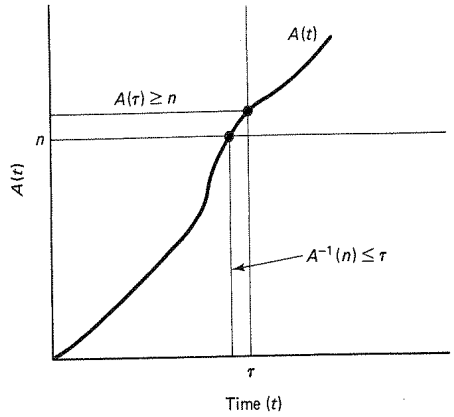

Figure 6.15 The event that $a^{-1}(n) \leq \tau$ is equivalent to the event $A(\tau) \geq n$.

intersects the vertical line, passing through the time τ, at or above the point n. As can be seen in Fig. 6.13, the statements are equivalent.

There is also an important relationship between the two points of intersection, $A^{-1}(n)$ and $A(t)$. Specifically, if $\tau\lambda \approx n$:

$$\lambda[\tau - A^{-1}(n)] + n \approx \lambda[2\tau - A^{-1}(n)] \approx \eta \qquad (6.A3)$$

One can conclude (for large n) that if $A^{-1}(n)$ is approximately normally distributed, $A(t)$ must also be approximately normally distributed, with mean and variance defined by

$$E[A(t)] = \lambda t \qquad V[A(t)] \approx \lambda^2 V[A^{-1}(\lambda t)]$$
$$\approx \lambda^2(\lambda t)V(A) \quad \text{(independent interarrival times)}$$
$$\approx \lambda t C^2(A) \qquad (6.A4)$$

where $V(A)$ is the variance of the interarrival time and $C(A)$ is the coefficient of variation of the interarrival time. If the arrival process is Poisson, $V(A)$ is just $1/\lambda^2$, and $V[A(t)]$ equals λt, the variance of a Poisson distribution with mean λt. Similar relationships hold for $D_s(t)$, which is also approximately normally distributed:

$$E[D_s(t)] = ct \qquad V[D_s(t)] \approx c^2 V[D_s^{-1}(ct)]$$
$$\approx c^2(ct)V(S) \quad \text{(independent service times)}$$
$$\approx ctC^2(S) \qquad (6.A5)$$

where $V(S)$ is the variance of the service time distribution and $C(S)$ is the coefficient of variation.

It is now possible to derive the probability distribution for the number of customers in the system, $L_s(t)$. Because $L_s(t)$ is the difference between two (approximately) normal random variables, $A(t)$ and $D_s(t)$, it must also be normally distributed. Assuming service

times are independent, that customers arrive by a Poisson process, and that arrival and departure processes are independent:

Poisson Arrivals, Independent Service Times

$$E[L_s(t)] = [A(0) - D(0)] + (\lambda - c)t \qquad (6.A6a)$$
$$V[L_s(t)] = t[\lambda + cC^2(S)] \qquad (6.A6b)$$

Equation (6.A6) can be used both to predict $L_s(t)$ and to obtain a confidence interval for $L_s(t)$, based on the normal distribution. It is valid for large values of t, provided that the probability that the queue vanishes is negligible. This is to say

Approximation Valid When

$$E[L_s(t)] - 2\sqrt{V[L_s(t)]} > 0 \qquad t \geq 0 \qquad (6.A7)$$

In reality, a queue will not instantaneously attain a large queue size, as assumed; it must first make the "transition through saturation." At or below saturation, the departure and arrival processes are certainly not independent. From time to time, the queue will vanish, and service must stop. Equation (6.A6a) will then *underestimate* queue size. Fortunately, there is an alternative model: the diffusion *equation*. Through the use of *boundary conditions*, the diffusion equation can be used to approximate the queue size distribution, even when the queue vanishes from time to time. The diffusion equation was the basis for Eq. (6.25), which gives an estimate for the queue size at the time when the queue makes the transition through saturation (that is, $\lambda = c$).