Queuing Analysis

Queueing analysis, or queueing theory, is a mathematical approach to studying and optimizing systems where entities wait in line (queue) for service. Let's break down this definition with an example:

Imagine you're at a popular coffee shop during rush hour. Customers arrive at the shop, join a queue to order their drinks, wait for their turn, get served by the barista, and then leave. This scenario can be analyzed using queueing theory.

Some definition :

a. **Customer** : One who requires service is called a customer.

b. **Server** : One who provides service is called server.

c. **Queue (Waiting line)** : A group of customers at some place to receive service is called the queue.

Component of Queue System :

(a) The input or arrival pattern

(b) Queue or waiting line

(c) The service discipline

(d) The service machanism or service pattern

(e) The output or departure

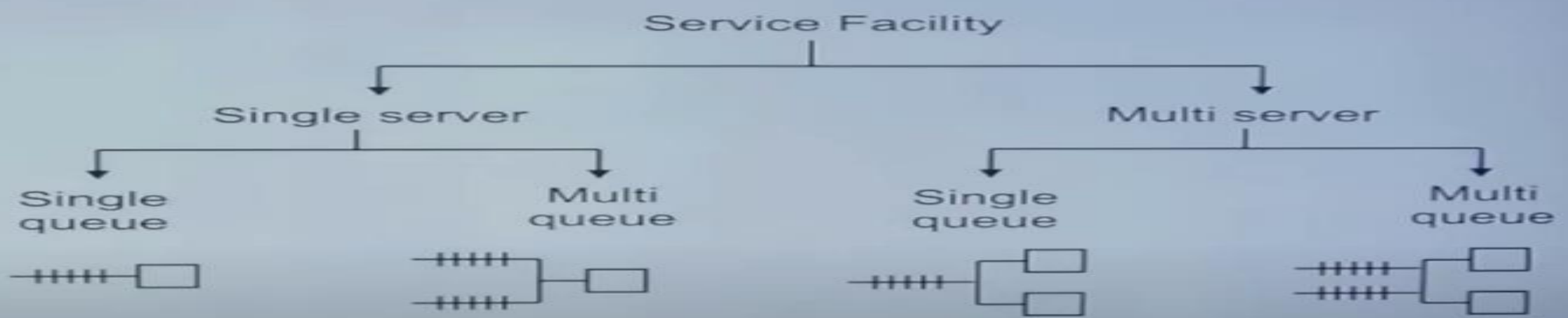
a. **Input or Arrival Pattern :**

(i) Balking : If the customer return back without getting service, because of long queue is called balking.

(ii) Reneging : This occur when the waiting customer leaves the queue, due to impatient.

(iii) Jocklying : If there be more than one queue and customer leave one queue and join other is called Jocklying.

b. **Service pattern :**



SOME IMPORTANT DEFINITION IN QUEUE THEORY

(a) Queue Length : The number of person in queue at any time.

(b) Servicing Time : Time taken for servicing one unit of the queue.

(c) Mean Arrival Rate : The number of expected customers in one unit time is known as mean arrival rate. It is denoted by λ.

(d) Mean servicing rate : The expected complete service in one unit time is called mean servicing rate. It is denoted by μ.

$$\text{Traffic intensity } (\rho) = \frac{\text{Mean arrival rate}}{\text{Mean servicing rate}} = \frac{\lambda}{\mu}$$

(e) Idle Period : The time interval in between the completion of the service and the new arrival is called the idle time.

Example : The time spent by repairman on his jobs has an exponential distribution with mean 30 minutes. If he repair sets in the order in which they come in and if the arrival of sets is approximately Poisson with an average rate of 10 per 8 hour day, then find traffic intensity ρ.

Solution :

Case - 1 : Given

$$\mu = \frac{1}{30} \times 60 = 2 \text{ sets per hour} \quad \& \quad \lambda = \frac{10}{8} = \frac{5}{4} \text{ per house}$$

Then $\rho = \frac{\lambda}{\mu} = \frac{5}{4 \times 2} = \frac{5}{8}$

CLASSIFICATION OF QUEUE MODELS

Some Important Notations

(i) λ - Mean arrival rate

(ii) μ - Mean service rate

(iii) L_s - Expected service length

(iv) L_q - Expected queue length

(v) W_s - Expected waiting time per customer in service

(vi) W_q - Expected waiting time per customer in queue

(vii) L_n - Expected length of non-empty queue.

CLASSIFICATION OF QUEUE MODEL

1. Single Server Queuing Models

Model - I (M/M/1) : (∞/FIFO)

Here First M : Poisson Input

Second M : Poisson Output

1 : Single Server

∞ : Infinite Capacity of System

FIFO : First In First Out

Model - II (M/M/1) : (N/FIFO)

Here the arrival and service rate depend upon the length of line.

Model - I (M/M/1) : (∞/FIFO)

Queue model with Poisson arrival, Poisson service, Single server channel with infinite capacity and FIFO service discipline.

Arrival rate = λ/hr

Service rate = μ/hr

Traffic intensity $\rho = \lambda/\mu$

and $P_n = \rho^n (1 - \rho)$

1. Average number of customer in service $L_s = \frac{\lambda}{\mu - \lambda}$

2. Average number of customer in queue $L_q = \frac{\lambda^2}{\mu(\mu - \lambda)}$

3. Average waiting time of customer in queue $W_q = \frac{\lambda}{\mu(\mu - \lambda)}$

4. Average waiting time of customer in service $W_s = \frac{1}{\mu - \lambda}$

5. Probability of queue size exceeds n

   $P(\text{queue size} \geq n) = \rho^n$

6. Expected length of non-empty queue $L_n = \frac{\mu}{\mu - \lambda} = \frac{1}{1 - \rho}$

7. Probability that no customer at the counter

$$P_0 = 1 - \rho = 1 - \frac{\lambda}{\mu} = \frac{\mu - \lambda}{\mu}$$

8. Probability of waiting time in queue $\geq n$

$$P(\text{waiting time} \geq n) = \int_n^\infty \lambda\left(1 - \frac{\lambda}{\mu}\right) e^{-(\mu-\lambda)t} dt$$

9. Probability that an arrival will have to wait for more than N minutes

$P(\text{Waiting time + Service time} \geq N) = \rho^N$

$$\Rightarrow \quad \rho^N = \int_N^\infty (\mu - \lambda) e^{-(\mu-\lambda)t} dt$$

Example : Customers arrive at a box office with one ticket window according to a Poisson's input process with mean rate of 30 per hour. The time required to serve a customer has an exponential distribution with mean 90 seconds.

Find the average

(a) Length of service (L_s)

(b) Queue length (L_q)

(c) Waiting time in queue (W_q)

(d) Time spent by a customer in the system

Solution : Mean arrival rate $\lambda = 30$ customers/hour

Mean service rate $\mu = \dfrac{1}{3/2 \text{ minutes}} = \dfrac{2}{3} \times 60 = 40$ services/hour

The time spent by a repairman on his jobs has an exponential distribution with mean 30 minutes. If he repairs sets in the order in which they come in, and if the arrival of sets is approximately Poisson with an average rate of 10 per 8 hour day, what is the repairman's expected idle time each day? How many jobs are ahead of the average set just brought in?

Arrivals at a telephone booth are considered to be Poisson, with an average time of 10 minutes between one arrival and the next. The length of a phone call assumed to be distributed exponentially, with mean 3 minutes. Find the following :

(a) What is the probability that a person arriving at the booth will have to wait?

(b) What is the average length of the queues that form from time to time?

(c) The telephone department will install a second booth when convinced that an arrival would expect to have to wait at least three minutes for the phone. by how much must the flow of arrivals he increased in order to justify a second booth?

Model - II (M/M/1) : (N/FIFO)

Queue model with Poisson arrival, Poisson service, Single server channel with finite capacity (N) and FIFO service discipline.

Arrival rate = λ/hr

Service rate = μ/hr

Traffic intensity $\rho = \lambda/\mu$

Example : Patients arrive at a clinic according to Poisson distribution at a rate of 30 patient per hour. The waiting room can not accomodate more than 14 patients. Examination time per patient is exponential with mean rate of 20 per hr.

(i) Find the effective arrival rate at the clinic.

(ii) What is the probability that an arriving patient will not wait?

(iii) What is the expected waiting time until a patient is discharged from the clinic?

Solution : Given $\lambda = 30$ patient / hr and $\mu = 20$ patient/hr

Also, $\rho = \frac{\lambda}{\mu} = \frac{30}{20} = 1.5 \neq 1$

Hence, $P_0 = \frac{1-\rho}{1-\rho^{N+1}} = \frac{1-1.5}{1-(1.5)^{16}} = \frac{0.5}{655.8408} = 0.000762$

(i) Hence effective arrival rate is

$$\mu(1-P_0) = 20(1-0.000762) = 19.98/\text{hr.}$$

(ii) P(patient will not wait) = $P_0 = 0.000762$

(iii) $$L_s = \frac{\rho}{1-\rho^{N+1}}\left(\frac{1-\rho^N}{1-\rho} - N\rho^N\right)$$

$$= \frac{1.5}{1-(1.5)^{16}}\left(\frac{1-(1-5)^{15}}{1-1.5} - 15(1.5)^{15}\right)$$

$$= \frac{1.5}{-655.84}(-5694.62) \cong 13.02$$

$$W_s = \frac{L_s}{\mu(1-P_0)} \Rightarrow W_s = \frac{13.02}{19.98} = 0.65 \text{ hrs.} \cong 39 \text{ minutes}$$

CHAPTER 17

Queueing Theory

Queues (waiting lines) are a part of everyday life. We all wait in queues to buy a movie ticket, make a bank deposit, pay for groceries, mail a package, obtain food in a cafeteria, start a ride in an amusement park, etc. We have become accustomed to considerable amounts of waiting, but still get annoyed by unusually long waits.

However, having to wait is not just a petty personal annoyance. The amount of time that a nation's populace wastes by waiting in queues is a major factor in both the quality of life there and the efficiency of the nation's economy.

Great inefficiencies also occur because of other kinds of waiting than people standing in line. For example, making *machines* wait to be repaired may result in lost production. *Vehicles* (including ships and trucks) that need to wait to be unloaded may delay subsequent shipments. *Airplanes* waiting to take off or land may disrupt later travel schedules. Delays in *telecommunication* transmissions due to saturated lines may cause data glitches. Causing *manufacturing jobs* to wait to be performed may disrupt subsequent production. Delaying *service jobs* beyond their due dates may result in lost future business.

Queueing theory is the study of waiting in all these various guises. It uses *queueing models* to represent the various types of *queueing systems* (systems that involve queues of some kind) that arise in practice. Formulas for each model indicate how the corresponding queueing system should perform, including the average amount of waiting that will occur, under a variety of circumstances.

Therefore, these queueing models are very helpful for determining how to operate a queueing system in the most effective way. Providing too much service capacity to operate the system involves excessive costs. But not providing enough service capacity results in excessive waiting and all its unfortunate consequences. The models enable finding an appropriate balance between the cost of service and the amount of waiting.

After some general discussion, this chapter presents most of the more elementary queueing models and their basic results. Section 17.10 discusses how the information provided by queueing theory can be used to design queueing systems that minimize the total cost of service and waiting, and then Chap. 26 (on the book's website) elaborates considerably further on the application of queueing theory in this way.

17.1 PROTOTYPE EXAMPLE

The emergency room of COUNTY HOSPITAL provides quick medical care for emergency cases brought to the hospital by ambulance or private automobile. At any hour there is always one doctor on duty in the emergency room. However, because of a growing tendency for emergency cases to use these facilities rather than go to a private physician, the hospital has been experiencing a continuing increase in the number of emergency room visits each year. As a result, it has become quite common for patients arriving during peak usage hours (the early evening) to have to wait until it is their turn to be treated by the doctor. Therefore, a proposal has been made that a second doctor should be assigned to the emergency room during these hours, so that two emergency cases can be treated simultaneously. The hospital's management engineer has been assigned to study this question.

The management engineer began by gathering the relevant historical data and then projecting these data into the next year. Recognizing that the emergency room is a queueing system, she applied several alternative queueing theory models to predict the waiting characteristics of the system with one doctor and with two doctors, as you will see in the latter sections of this chapter (see Tables 17.2 and 17.3).

17.2 BASIC STRUCTURE OF QUEUEING MODELS

The Basic Queueing Process

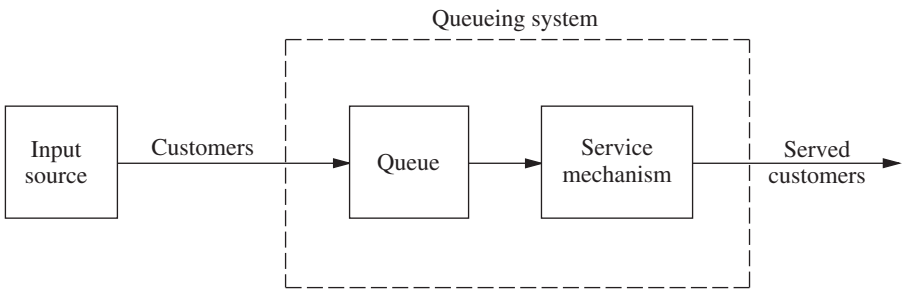
The basic process assumed by most queueing models is the following. *Customers* requiring service are generated over time by an *input source.* These customers enter the *queueing system* and join a *queue.* At certain times, a member of the queue is selected for service by some rule known as the *queue discipline.* The required service is then performed for the customer by the *service mechanism,* after which the customer leaves the queueing system. This process is depicted in Fig. 17.1.

Many alternative assumptions can be made about the various elements of the queueing process; they are discussed next.

Input Source (Calling Population)

One characteristic of the input source is its size. The *size* is the total number of customers that might require service from time to time, i.e., the total number of distinct potential customers. This population from which arrivals come is referred to as the **calling population.** The size may be assumed to be either *infinite* or *finite* (so that the input source also is said to be either *unlimited* or *limited*). Because the calculations are far easier for the infinite case, this assumption often is made even when the actual size is some relatively

FIGURE 17.1
The basic queueing process.



large finite number; and it should be taken to be the implicit assumption for any queueing model that does not state otherwise. The finite case is more difficult analytically because the number of customers in the queueing system affects the number of potential customers outside the system at any time. However, the finite assumption must be made if the rate at which the input source generates new customers is significantly affected by the number of customers in the queueing system.

The statistical pattern by which customers are generated over time must also be specified. The common assumption is that they are generated according to a *Poisson process;* i.e., the number of customers generated until any specific time has a Poisson distribution. As we discuss in Sec. 17.4, this case is the one where arrivals to the queueing system occur randomly but at a certain fixed mean rate, regardless of how many customers already are there (so the *size* of the input source is *infinite*). An equivalent assumption is that the probability distribution of the time between consecutive arrivals is an *exponential* distribution. (The properties of this distribution are described in Sec. 17.4.) The time between consecutive arrivals is referred to as the **interarrival time.**

Any unusual assumptions about the behavior of arriving customers must also be specified. One example is *balking,* where the customer refuses to enter the system and is lost if the queue is too long.

Queue

The queue is where customers wait *before* being served. A queue is characterized by the maximum permissible number of customers that it can contain. Queues are called *infinite* or *finite,* according to whether this number is infinite or finite. The assumption of an *infinite queue* is the standard one for most queueing models, even for situations where there actually is a (relatively large) finite upper bound on the permissible number of customers, because dealing with such an upper bound would be a complicating factor in the analysis. However, for queueing systems where this upper bound is small enough that it actually would be reached with some frequency, it becomes necessary to assume a *finite queue.*

Queue Discipline

The queue discipline refers to the order in which members of the queue are selected for service. For example, it may be first-come-first-served, random, according to some priority procedure, or some other order. First-come-first-served usually is assumed by queueing models, unless it is stated otherwise.

Service Mechanism

The service mechanism consists of one or more *service facilities,* each of which contains one or more *parallel service channels,* called **servers.** If there is more than one service facility, the customer may receive service from a sequence of these (*service channels in series*). At a given facility, the customer enters one of the parallel service channels and is completely serviced by that server. A queueing model must specify the arrangement of the facilities and the number of servers (parallel channels) at each one. Most elementary models assume one service facility with either one server or a finite number of servers.

The time elapsed from the commencement of service to its completion for a customer at a service facility is referred to as the **service time** (or *holding time*). A model of a particular queueing system must specify the probability distribution of service times for each server (and possibly for different types of customers), although it is common to assume the *same* distribution for all servers (all models in this chapter make this assumption). The service-time distribution that is most frequently assumed in practice (largely because it is far more tractable than any other) is the *exponential* distribution discussed in Sec. 17.4,

and most of our models will be of this type. Other important service-time distributions are the *degenerate* distribution (constant service time) and the *Erlang* (gamma) distribution, as illustrated by models in Sec. 17.7.

An Elementary Queueing Process

As we have already suggested, queueing theory has been applied to many different types of waiting-line situations. However, the most prevalent type of situation is the following: A single waiting line (which may be empty at times) forms in the front of a single service facility, within which are stationed one or more servers. Each customer generated by an input source is serviced by one of the servers, perhaps after some waiting in the queue (waiting line). The queueing system involved is depicted in Fig. 17.2.

Notice that the queueing process in the prototype example of Sec. 17.1 is of this type. The input source generates customers in the form of emergency cases requiring medical care. The emergency room is the service facility, and the doctors are the servers.

A server need not be a single individual; it may be a group of persons, e.g., a repair crew that combines forces to perform simultaneously the required service for a customer. Furthermore, servers need not even be people. In many cases, a server can instead be a machine, a vehicle, an electronic device, etc. By the same token, the customers in the waiting line need not be people. For example, they may be items waiting for a certain operation by a given type of machine, or they may be cars waiting in front of a tollbooth.

It is not necessary that there actually be a physical waiting line forming in front of a physical structure that constitutes the service facility. The members of the queue may instead be scattered throughout an area, waiting for a server to come to them, e.g., machines waiting to be repaired. The server or group of servers assigned to a given area constitutes the service facility for that area. Queueing theory still gives the average number waiting, the average waiting time, and so on, because it is irrelevant whether the customers wait together in a group. The only essential requirement for queueing theory to be applicable is that changes in the number of customers waiting for a given service occur just as though the physical situation described in Fig. 17.2 (or a legitimate counterpart) prevailed.

Except for Sec. 17.9, all the queueing models discussed in this chapter are of the elementary type depicted in Fig. 17.2. Many of these models further assume that all

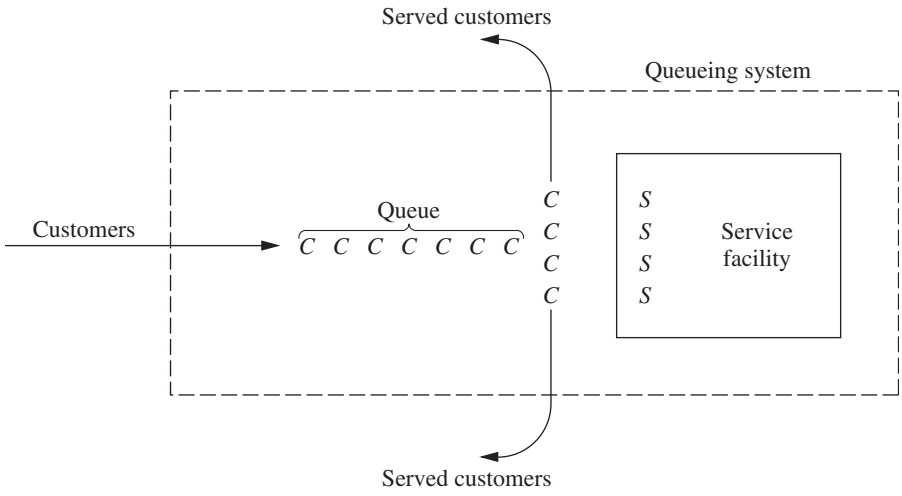
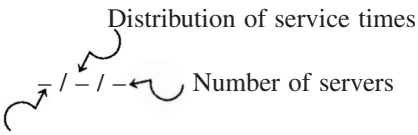


FIGURE 17.2
An elementary queueing system (each customer is indicated by a *C* and each server by an *S*).

interarrival times are independent and identically distributed and that all *service times* are independent and identically distributed. Such models conventionally are labeled as follows:



where M = exponential distribution (Markovian), as described in Sec. 17.4,
D = degenerate distribution (constant times), as discussed in Sec. 17.7,
E_k = Erlang distribution (shape parameter = k), as described in Sec. 17.7,
G = general distribution (any arbitrary distribution allowed),[1] as discussed in Sec. 17.7.

For example, the $M/M/s$ model discussed in Sec. 17.6 assumes that both interarrival times and service times have an exponential distribution and that the number of servers is s (any positive integer). The $M/G/1$ model discussed again in Sec. 17.7 assumes that interarrival times have an exponential distribution, but it places no restriction on what the distribution of service times must be, whereas the number of servers is restricted to be exactly 1. Various other models that fit this labeling scheme also are introduced in Sec. 17.7.

Terminology and Notation

Unless otherwise noted, the following standard terminology and notation will be used:

State of system = number of customers in queueing system.
Queue length = number of customers waiting for service to begin.
= state of system *minus* number of customers being served.
$N(t)$ = number of customers in queueing system at time t ($t \geq 0$).
$P_n(t)$ = probability of exactly n customers in queueing system at time t, given number at time 0.
s = number of servers (parallel service channels) in queueing system.
λ_n = mean arrival rate (expected number of arrivals per unit time) of new customers when n customers are in system.
μ_n = mean service rate for overall system (expected number of customers completing service per unit time) when n customers are in system. *Note:* μ_n represents *combined* rate at which all *busy* servers (those serving customers) achieve service completions.
λ, μ, ρ = see following paragraph.

When λ_n is a constant for all n, this constant is denoted by λ. When the mean service rate *per busy server* is a constant for all $n \geq 1$, this constant is denoted by μ. (In this case, $\mu_n = s\mu$ when $n \geq s$, that is, when all s servers are busy.) Under these circumstances, $1/\lambda$ and $1/\mu$ are the *expected interarrival time* and the *expected service time,* respectively. Also, $\rho = \lambda/(s\mu)$ is the **utilization factor** for the service facility, i.e., the expected fraction of

[1]When we refer to interarrival times, it is conventional to replace the symbol G by GI = general independent distribution.

time the individual servers are busy, because $\lambda/(s\mu)$ represents the fraction of the system's service capacity ($s\mu$) that is being *utilized* on the average by arriving customers (λ).

Certain notation also is required to describe *steady-state* results. When a queueing system has recently begun operation, the state of the system (number of customers in the system) will be greatly affected by the initial state and by the time that has since elapsed. The system is said to be in a **transient condition.** However, after sufficient time has elapsed, the state of the system becomes essentially independent of the initial state and the elapsed time (except under unusual circumstances).[2] The system has now essentially reached a **steady-state condition,** where the probability distribution of the state of the system remains the same (the *steady-state* or *stationary* distribution) over time. Queueing theory has tended to focus largely on the steady-state condition, partially because the transient case is more difficult analytically. (Some transient results exist, but they are generally beyond the technical scope of this book.) The following notation assumes that the system is in a *steady-state condition:*

P_n = probability of exactly n customers in queueing system.

L = expected number of customers in queueing system $= \sum_{n=0}^{\infty} nP_n$.

L_q = expected queue length (excludes customers being served) $= \sum_{n=s}^{\infty} (n - s)P_n$.

$\mathcal{W}$ = waiting time in system (includes service time) for each individual customer.

$W = E(\mathcal{W})$.

$\mathcal{W}_q$ = waiting time in queue (excludes service time) for each individual customer.

$W_q = E(\mathcal{W}_q)$.

Relationships between L, W, L_q, and W_q

Assume that λ_n is a constant λ for all n. It has been proved that in a steady-state queueing process,

$$L = \lambda W.$$

(Because John D. C. Little provided the first rigorous proof, this equation sometimes is referred to as **Little's formula.**) Furthermore, the same proof also shows that

$$L_q = \lambda W_q.$$

If the λ_n are not equal, then λ can be replaced in these equations by $\bar{\lambda}$, the *average* arrival rate over the long run. (We shall show later how $\bar{\lambda}$ can be determined for some basic cases.)

Now assume that the mean service time is a constant, $1/\mu$ for all $n \geq 1$. It then follows that

$$W = W_q + \frac{1}{\mu}.$$

These relationships are extremely important because they enable all four of the fundamental quantities—L, W, L_q, and W_q—to be immediately determined as soon as

[2]When λ and μ are defined, these unusual circumstances are that $\rho \geq 1$, in which case the state of the system tends to grow continually larger as time goes on.

one is found analytically. This situation is fortunate because some of these quantities often are much easier to find than others when a queueing model is solved from basic principles.

17.3 EXAMPLES OF REAL QUEUEING SYSTEMS

Our description of queueing systems in Sec. 17.2 may appear relatively abstract and applicable to only rather special practical situations. On the contrary, queueing systems are surprisingly prevalent in a wide variety of contexts. To broaden your horizons on the applicability of queueing theory, we shall briefly mention various examples of real queueing systems that fall into several broad categories. We then will describe queueing systems in several prominent companies (plus one city) and the award-winning studies that were conducted to design these systems.

Some Classes of Queueing Systems

One important class of queueing systems that we all encounter in our daily lives is **commercial service systems,** where outside customers receive service from commercial organizations. Many of these involve person-to-person service at a fixed location, such as a barber shop (the barbers are the servers), bank teller service, checkout stands at a grocery store, and a cafeteria line (service channels in series). However, many others do not, such as home appliance repairs (the server travels to the customers), a vending machine (the server is a machine), and a gas station (the cars are the customers).

Another important class is **transportation service systems.** For some of these systems the vehicles are the customers, such as cars waiting at a tollbooth or traffic light (the server), a truck or ship waiting to be loaded or unloaded by a crew (the server), and airplanes waiting to land or take off from a runway (the server). (An unusual example of this kind is a parking lot, where the cars are the customers and the parking spaces are the servers, but there is no queue because arriving customers go elsewhere to park if the lot is full.) In other cases, the vehicles, such as taxicabs, fire trucks, and elevators, are the servers.

In recent years, queueing theory probably has been applied most to **internal service systems,** where the customers receiving service are *internal* to the organization. Examples include materials-handling systems, where materials-handling units (the servers) move loads (the customers); maintenance systems, where maintenance crews (the servers) repair machines (the customers); and inspection stations, where quality control inspectors (the servers) inspect items (the customers). Employee facilities and departments servicing employees also fit into this category. In addition, machines can be viewed as servers whose customers are the jobs being processed. A related example is a computer laboratory, where each computer is viewed as the server.

There is now growing recognition that queueing theory also is applicable to **social service systems.** For example, a judicial system is a queueing network, where the courts are service facilities, the judges (or panels of judges) are the servers, and the cases waiting to be tried are the customers. A legislative system is a similar queueing network, where the customers are the bills waiting to be processed. Various health-care systems also are queueing systems. You already have seen one example in Sec. 17.1 (a hospital emergency room), but you can also view ambulances, X-ray machines, and hospital beds as servers in their own queueing systems. Similarly, families waiting for low- and moderate-income housing, or other social services, can be viewed as customers in a queueing system.

Although these are four broad classes of queueing systems, they still do not exhaust the list. In fact, queueing theory first began early in the 20th century with applications to telephone engineering (the founder of queueing theory, A. K. Erlang, was an employee of the Danish Telephone Company in Copenhagen), and telephone engineering still is an important application. Furthermore, we all have our own personal queues—homework assignments, books to be read, and so forth. However, these examples are sufficient to suggest that queueing systems do indeed pervade many areas of society.

Some Award-Winning Studies to Design Queueing Systems

The prestigious *Franz Edelman Awards for Management Science Achievement* are awarded annually by the Institute of Operations Research and the Management Sciences (INFORMS) for the year's best applications of OR. A rather substantial number of these awards have been given for innovative applications of queueing theory to the design of queueing systems.

Two of these award-winning applications of queueing theory are described in application vignettes later in this chapter (Secs. 17.6 and 17.9). The selected references at the end of the chapter also include a sampling of articles describing some other award-winning applications. (A link to all these articles, including for the application vignettes, is provided on the book's website.) We briefly describe a few of these other applications of queueing theory below.

As described in Selected Reference A1, one of the early first-prize winners of the Edelman competition was the *Xerox Corporation.* The company had recently introduced a major new duplicating system that was proving to be particularly valuable for its owners. Consequently, these customers were demanding that Xerox's tech reps reduce the waiting times to repair the machines. An OR team then applied queueing theory to study how to best meet the new service requirements. This resulted in replacing the previous one-person tech rep territories by larger three-person tech rep territories. This change had the dramatic effect of both substantially reducing the average waiting times of the customers and increasing the utilization of the tech reps by over 50 percent. (Chapter 11 of Selected Reference 9 presents a case study that is based on this application of queueing theory by the Xerox Corporation.)

L.L. Bean, Inc., the large telemarketer and mail-order catalog house, relied mainly on queueing theory for its award-winning study of how to allocate its telecommunications resources that is described in Selected Reference A4. The telephone calls coming in to its call center to place orders are the customers in a large queueing system, with the telephone agents as the servers. The key questions being asked during the study were the following.

1. How many telephone trunk lines should be provided for incoming calls to the call center?
2. How many telephone agents should be scheduled at various times?
3. How many hold positions should be provided for customers waiting for a telephone agent? (Note that the limited number of hold positions causes the system to have a finite queue.)

For each interesting combination of these three quantities, queueing models provide the measures of performance of the queueing system. Given these measures, the OR team carefully assessed the cost of lost sales due to making some customers either incur a busy signal or be placed on hold too long. By adding the cost of the telemarketing resources,

the team then was able to find the combination of the three quantities that minimizes the expected total cost. This resulted in cost savings of $9 to $10 million per year.

Another first prize in the Edelman competition was won by AT&T for a study that combined the use of queueing theory and simulation (the subject of Chap. 20). As described in Selected Reference A2, the queueing models are of both AT&T's telecommunication network and the call center environment for the typical business customers of AT&T that have such a center. The purpose of the study was to develop a user-friendly PC-based system that AT&T's business customers can use to guide them in how to design or redesign their call centers. Since call centers comprise one of the United States' fastest-growing industries, this system had been used about 2,000 times by AT&T's business customers by the time of the article. This resulted in more than $750 million in annual profit for these customers.

Hewlett-Packard (HP) is a leading multinational manufacturer of electronic equipment. Some years ago, the company installed a mechanized assembly-line system for manufacturing ink-jet printers at its plant in Vancouver, Washington, to meet the exploding demand for such printers. It soon became apparent that the system installed would not be fast enough or reliable enough to meet the company's production goals. Therefore, a joint team of management scientists from HP and the Massachusetts Institute of Technology (MIT) was formed to study how to redesign the system to improve its performance.

As described in Selected Reference A3 for this award-winning study, the HP/MIT team quickly realized that the assembly-line system could be modeled as a special kind of queueing system where the customers (the printers to be assembled) go through a series of servers (assembly operations) in a fixed sequence. A special queueing model for this kind of system quickly provided the analytical results that were needed to determine how the system should be redesigned to achieve the required capacity in the most economical way. The changes included adding some buffer storage space at strategic points to better maintain the flow of work to the subsequent stations and to dampen the effect of machine failures. The new design increased productivity about 50 percent and yielded incremental revenues of approximately $280 million in printer sales as well as additional revenue from ancillary products. This innovative application of the special queueing model also provided HP with a new method for creating rapid and effective system designs subsequently in other areas of the company.

17.4 THE ROLE OF THE EXPONENTIAL DISTRIBUTION

The operating characteristics of queueing systems are determined largely by two statistical properties, namely, the probability distribution of *interarrival times* (see "Input Source" in Sec. 17.2) and the probability distribution of *service times* (see "Service Mechanism" in Sec. 17.2). For real queueing systems, these distributions can take on almost any form. (The only restriction is that negative values cannot occur.) However, to formulate a queueing theory *model* as a representation of the real system, it is necessary to specify the assumed form of each of these distributions. To be useful, the assumed form should be *sufficiently realistic* that the model provides *reasonable predictions* while, at the same time, being *sufficiently simple* that the model is *mathematically tractable.* Based on these considerations, the most important probability distribution in queueing theory is the *exponential distribution.*

Suppose that a random variable T represents either interarrival or service times. (We shall refer to the occurrences marking the end of these times—arrivals or service

completions—as *events*.) This random variable is said to have an *exponential distribution with parameter* α if its probability density function is

$$f_T(t) = \begin{cases} \alpha e^{-\alpha t} & \text{for } t \geq 0 \\ 0 & \text{for } t < 0, \end{cases}$$

as shown in Fig. 17.3. In this case, the cumulative probabilities are

$$\begin{aligned} P\{T \leq t\} &= 1 - e^{-\alpha t} \\ P\{T > t\} &= e^{-\alpha t} \end{aligned} \qquad (t \geq 0),$$

and the expected value and variance of T are, respectively,

$$E(T) = \frac{1}{\alpha},$$

$$\text{var}(T) = \frac{1}{\alpha^2}.$$

What are the implications of assuming that T has an exponential distribution for a queueing model? To explore this question, let us examine six key properties of the exponential distribution.

Property 1: $f_T(t)$ is a strictly *decreasing* function of t ($t \geq 0$).

One consequence of Property 1 is that

$$P\{0 \leq T \leq \Delta t\} > P\{t \leq T \leq t + \Delta t\}$$

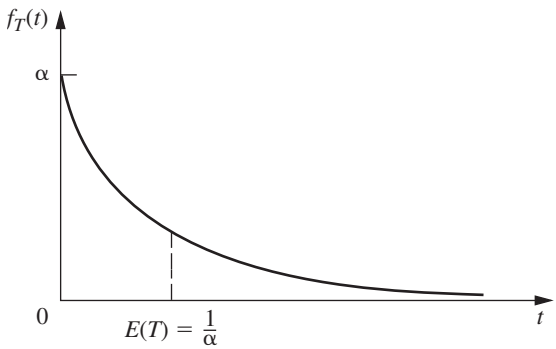
for any strictly positive values of Δt and t. [This consequence follows from the fact that these probabilities are the area under the $f_T(t)$ curve over the indicated interval of length Δt, and the average height of the curve is less for the second probability than for the first.] Therefore, it is not only possible but also relatively likely that T will take on a small value near zero. In fact,

$$P\left\{0 \leq T \leq \frac{1}{2}\,\frac{1}{\alpha}\right\} = 0.393$$

whereas

$$P\left\{\frac{1}{2}\,\frac{1}{\alpha} \leq T \leq \frac{3}{2}\,\frac{1}{\alpha}\right\} = 0.383,$$

so that the value T takes on is more likely to be "small" [i.e., less than half of $E(T)$] than "near" its expected value [i.e., no further away than half of $E(T)$], even though the second interval is twice as wide as the first.



■ **FIGURE 17.3**
Probability density function for the exponential distribution.

Is this really a reasonable property for T in a queueing model? If T represents *service times,* the answer depends upon the general nature of the service involved, as discussed next.

If the service required is essentially identical for each customer, with the server always performing the same sequence of service operations, then the actual service times tend to be near the expected service time. Small deviations from the mean may occur, but usually because of only minor variations in the efficiency of the server. A small service time far below the mean is essentially impossible, because a certain minimum time is needed to perform the required service operations even when the server is working at top speed. The exponential distribution clearly does not provide a close approximation to the service-time distribution for this type of situation.

On the other hand, consider the type of situation where the specific tasks required of the server differ among customers. The broad nature of the service may be the same, but the specific type and amount of service differ. For example, this is the case in the County Hospital emergency room problem discussed in Sec. 17.1. The doctors encounter a wide variety of medical problems. In most cases, they can provide the required treatment rather quickly, but an occasional patient requires extensive care. Similarly, bank tellers and grocery store checkout clerks are other servers of this general type, where the required service is often brief but must occasionally be extensive. An exponential service-time distribution would seem quite plausible for this type of service situation.

If T represents *interarrival times,* Property 1 rules out situations where potential customers approaching the queueing system tend to postpone their entry if they see another customer entering ahead of them. On the other hand, it is entirely consistent with the common phenomenon of arrivals occurring "randomly," described by subsequent properties. Thus, when arrival times are plotted on a time line, they sometimes have the appearance of being clustered with occasional large gaps separating clusters, because of the substantial probability of small interarrival times and the small probability of large interarrival times, but such an irregular pattern is all part of true randomness.

Property 2: Lack of memory.

This property can be stated mathematically as

$$P\{T > t + \Delta t \,|\, T > \Delta t\} = P\{T > t\}$$

for any positive quantities t and Δt. In other words, the probability distribution of the *remaining* time until the event (arrival or service completion) occurs always is the same, regardless of how much time (Δt) already has passed. In effect, the process "forgets" its history. This surprising phenomenon occurs with the exponential distribution because

$$\begin{aligned} P\{T > t + \Delta t \,|\, T > \Delta t\} &= \frac{P\{T > \Delta t,\, T > t + \Delta t\}}{P\{T > \Delta t\}} \\ &= \frac{P\{T > t + \Delta t\}}{P\{T > \Delta t\}} \\ &= \frac{e^{-\alpha(t+\Delta t)}}{e^{-\alpha \Delta t}} \\ &= e^{-\alpha t} \\ &= P\{T > t\}. \end{aligned}$$

For *interarrival times,* this property describes the common situation where the time until the next arrival is completely uninfluenced by when the last arrival occurred. For *service times,* the property is more difficult to interpret. We should not expect it to hold in a situation where the server must perform the same fixed sequence of operations for each customer, because then a long elapsed service should imply that probably little

remains to be done. However, in the type of situation where the required service operations differ among customers, the mathematical statement of the property may be quite realistic. For this case, if considerable service has already elapsed for a customer, the only implication may be that this particular customer requires more extensive service than most.

Property 3: The *minimum* of several independent exponential random variables has an exponential distribution.

To state this property mathematically, let $T_1, T_2, \ldots, T_n$ be *independent* exponential random variables with parameters $\alpha_1, \alpha_2, \ldots, \alpha_n$, respectively. Also let U be the random variable that takes on the value equal to the *minimum* of the values actually taken on by $T_1, T_2, \ldots, T_n$; that is,

$$U = \min \{T_1, T_2, \ldots, T_n\}.$$

Thus, if T_i represents the time until a particular kind of event occurs, then U represents the time until the *first* of the n different events occurs. Now note that for any $t \geq 0$,

$$\begin{aligned} P\{U > t\} &= P\{T_1 > t, T_2 > t, \ldots, T_n > t\} \\ &= P\{T_1 > t\}P\{T_2 > t\} \cdots P\{T_n > t\} \\ &= e^{-\alpha_1 t}e^{-\alpha_2 t} \cdots e^{-\alpha_n t} \\ &= \exp\left(-\sum_{i=1}^{n} \alpha_i t\right), \end{aligned}$$

so that U indeed has an exponential distribution with parameter

$$\alpha = \sum_{i=1}^{n} \alpha_i.$$

This property has some implications for interarrival times in queueing models. In particular, suppose that there are several (n) *different* types of customers, but the interarrival times for *each* type (type i) have an exponential distribution with parameter α_i ($i = 1, 2, \ldots, n$). By Property 2, the *remaining* time from any specified instant until the next arrival of a customer of type i has this same distribution. Therefore, let T_i be this remaining time, measured from the instant a customer of *any* type arrives. Property 3 then tells us that U, the interarrival times for the queueing system as a whole, has an exponential distribution with parameter α defined by the last equation. As a result, you can choose to ignore the distinction between customers and still have exponential interarrival times for the queueing model.

However, the implications are even more important for *service times* in multiple-server queueing models than for interarrival times. For example, consider the situation where all the servers have the same exponential service-time distribution with parameter μ. For this case, let n be the number of servers *currently* providing service, and let T_i be the *remaining* service time for server i ($i = 1, 2, \ldots, n$), which also has an exponential distribution with parameter $\alpha_i = \mu$. It then follows that U, the time until the *next* service completion from any of these servers, has an exponential distribution with parameter $\alpha = n\mu$. In effect, the queueing system *currently* is performing just like a *single*-server system where service times have an exponential distribution with parameter $n\mu$. We shall make frequent use of this implication for analyzing multiple-server models later in the chapter.

When using this property, it sometimes is useful to also determine the probabilities for *which* of the exponential random variables will turn out to be the one which has the minimum value. For example, you might want to find the probability that a particular server j will finish serving a customer first among n busy exponential servers. It is fairly straightforward (see Prob. 17.4-9) to show that this probability is proportional to the

parameter α_j. In particular, the probability that T_j will turn out to be the smallest of the n random variables is

$$P\{T_j = U\} = \frac{\alpha_j}{\sum_{i=1}^{n} \alpha_i}, \qquad \text{for } j = 1, 2, \ldots, n.$$

Property 4: Relationship to the Poisson distribution.

Suppose that the *time* between consecutive occurrences of some particular kind of event (e.g., arrivals or service completions by a continuously busy server) has an exponential distribution with parameter α. Property 4 then has to do with the resulting implication about the probability distribution of the *number* of times this kind of event occurs over a specified time. In particular, let $X(t)$ be the number of occurrences by time t ($t \geq 0$), where time 0 designates the instant at which the count begins. The implication is that

$$P\{X(t) = n\} = \frac{(\alpha t)^n e^{-\alpha t}}{n!}, \qquad \text{for } n = 0, 1, 2, \ldots;$$

that is, $X(t)$ has a Poisson distribution with parameter αt. For example, with $n = 0$,

$$P\{X(t) = 0\} = e^{-\alpha t},$$

which is just the probability from the exponential distribution that the *first* event occurs after time t. The mean of this Poisson distribution is

$$E\{X(t)\} = \alpha t,$$

so that the expected number of events *per unit time* is α. Thus, α is said to be the *mean rate* at which the events occur. When the events are counted on a continuing basis, the counting process $\{X(t); t \geq 0\}$ is said to be a **Poisson process** with parameter α (the mean rate).

This property provides useful information about *service completions* when service times have an exponential distribution with parameter μ. We obtain this information by defining $X(t)$ as the number of service completions achieved by a *continuously busy* server in elapsed time t, where $\alpha = \mu$. For *multiple*-server queueing models, $X(t)$ can also be defined as the number of service completions achieved by n continuously busy servers in elapsed time t, where $\alpha = n\mu$.

The property is particularly useful for describing the probabilistic behavior of *arrivals* when interarrival times have an exponential distribution with parameter λ. In this case, $X(t)$ is the *number* of arrivals in elapsed time t, where $\alpha = \lambda$ is the *mean arrival rate.* Therefore, arrivals occur according to a **Poisson input process** with parameter λ. Such queueing models also are described as assuming a *Poisson input.*

Arrivals sometimes are said to occur *randomly,* meaning that they occur in accordance with a Poisson input process. One intuitive interpretation of this phenomenon is that every time period of fixed length has the *same* chance of having an arrival regardless of when the preceding arrival occurred, as suggested by the following property.

Property 5: For all positive values of t, $P\{T \leq t + \Delta t \mid T > t\} \approx \alpha\, \Delta t$, for small Δt.

Continuing to interpret T as the time from the last event of a certain type (arrival or service completion) until the next such event, we suppose that a time t already has elapsed without the event's occurring. We know from Property 2 that the probability that the event will occur within the next time interval of fixed length Δt is a *constant* (identified in the next paragraph), regardless of how large or small t is. Property 5 goes further to say that when the value of Δt is small, this constant probability can be approximated very closely by $\alpha\, \Delta t$. Furthermore, when considering different small

values of Δt, this probability is essentially *proportional* to Δt, with proportionality factor α. In fact, α is the *mean rate* at which the events occur (see Property 4), so that the *expected number* of events in the interval of length Δt is *exactly* $\alpha\ \Delta t$. The only reason that the probability of an event's occurring differs slightly from this value is the possibility that *more than one* event will occur, which has negligible probability when Δt is small.

To see why Property 5 holds mathematically, note that the constant value of our probability (for a fixed value of $\Delta t > 0$) is just

$$\begin{aligned} P\{T \le t + \Delta t \,|\, T > t\} &= P\{T \le \Delta t\} \\ &= 1 - e^{-\alpha\ \Delta t}, \end{aligned}$$

for any $t \ge 0$. Therefore, because the series expansion of e^x for any exponent x is

$$e^x = 1 + x + \sum_{n=2}^{\infty} \frac{x^n}{n!},$$

it follows that

$$\begin{aligned} P\{T \le t + \Delta t \,|\, T > t\} &= 1 - 1 + \alpha\ \Delta t - \sum_{n=2}^{\infty} \frac{(-\alpha\ \Delta t)^n}{n!} \\ &\approx \alpha\ \Delta t, \qquad \text{for small } \Delta t,^3 \end{aligned}$$

because the summation terms become relatively negligible for sufficiently small values of $\alpha\ \Delta t$.

Because T can represent either interarrival or service times in queueing models, this property provides a convenient approximation of the probability that the event of interest occurs in the next small interval (Δt) of time. An analysis based on this approximation also can be made exact by taking appropriate limits as $\Delta t \to 0$.

Property 6: Unaffected by aggregation or disaggregation.

This property is relevant primarily for verifying that the *input process* is *Poisson.* Therefore, we shall describe it in these terms, although it also applies directly to the exponential distribution (exponential interarrival times) because of Property 4.

We first consider the aggregation (combining) of several Poisson input processes into one overall input process. In particular, suppose that there are several (n) *different* types of customers, where the customers of each type (type i) arrive according to a *Poisson input process* with parameter λ_i ($i = 1, 2, \ldots, n$). Assuming that these are *independent* Poisson processes, the property says that the *aggregate* input process (arrival of all customers without regard to type) also must be Poisson, with parameter (mean arrival rate) $\lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_n$. In other words, having a Poisson process is *unaffected by aggregation.*

This part of the property follows directly from Properties 3 and 4. The latter property implies that the interarrival times for customers of type i have an exponential distribution with parameter λ_i. For this identical situation, we already discussed for Property 3 that it implies that the interarrival times for all customers also must have an exponential distribution, with parameter $\lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_n$. Using Property 4 again then implies that the aggregate input process is Poisson.

The second part of Property 6 ("unaffected by disaggregation") refers to the reverse case, where the *aggregate* input process (the one obtained by combining the input processes

[3]More precisely,

$$\lim_{\Delta t \to 0} \frac{P\{T \le t + \Delta t \,|\, T > t\}}{\Delta t} = \alpha.$$

for several customer types) is known to be Poisson with parameter λ, but the question now concerns the nature of the *disaggregated* input processes (the individual input processes for the individual customer types). Assuming that each arriving customer has a *fixed* probability p_i of being of type i ($i = 1, 2, \ldots, n$), with

$$\lambda_i = p_i\lambda \qquad \text{and} \qquad \sum_{i=1}^{n} p_i = 1,$$

the property says that the input process for customers of type i also must be Poisson with parameter λ_i. In other words, having a Poisson process is *unaffected by disaggregation.*

As one example of the usefulness of this second part of the property, consider the following situation. Indistinguishable customers arrive according to a Poisson process with parameter λ. Each arriving customer has a fixed probability p of *balking* (leaving without entering the queueing system), so the probability of entering the system is $1 - p$. Thus, there are two types of customers—those who balk and those who enter the system. The property says that each type arrives according to a Poisson process, with parameters $p\lambda$ and $(1 - p)\lambda$, respectively. Therefore, by using the latter Poisson process, queueing models that assume a Poisson input process can still be used to analyze the performance of the queueing system for those customers who enter the system.

Another example in the Worked Examples section of the books'website illustrates the application of several of the properties of the exponential distribution presented in this section.

17.5 THE BIRTH-AND-DEATH PROCESS

Most elementary queueing models assume that the inputs (arriving customers) and outputs (leaving customers) of the queueing system occur according to the *birth-and-death process.* This important process in probability theory has applications in various areas. However, in the context of queueing theory, the term **birth** refers to the *arrival* of a new customer into the queueing system, and **death** refers to the *departure* of a served customer. The *state* of the system at time t ($t \geq 0$), denoted by $N(t)$, is the number of customers in the queueing system at time t. The birth-and-death process describes *probabilistically* how $N(t)$ changes as t increases. Broadly speaking, it says that *individual* births and deaths occur *randomly,* where their mean occurrence rates depend only upon the current state of the system. More precisely, the assumptions of the birth-and-death process are the following:

Assumption 1. Given $N(t) = n$, the current probability distribution of the *remaining* time until the next *birth* (arrival) is *exponential* with parameter λ_n ($n = 0, 1, 2, \ldots$).

Assumption 2. Given $N(t) = n$, the current probability distribution of the *remaining* time until the next *death* (service completion) is *exponential* with parameter μ_n ($n = 1, 2, \ldots$).

Assumption 3. The random variable of assumption 1 (the remaining time until the next *birth*) and the random variable of assumption 2 (the remaining time until the next *death*) are mutually independent. The next transition in the state of the process is either

$n \rightarrow n + 1$ (a single birth)

or

$n \rightarrow n - 1$ (a single death),

depending on whether the former or latter random variable is smaller.

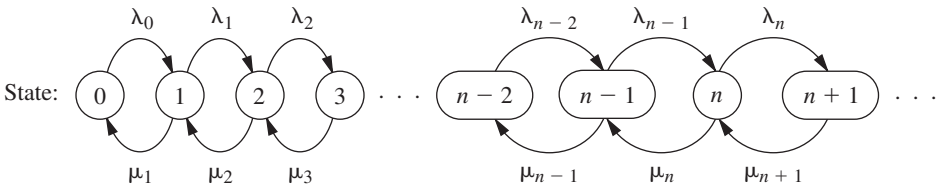


FIGURE 17.4
Rate diagram for the birth-and-death process.

For a queueing system, λ_n and μ_n respectively represent the *mean arrival rate* and the *mean rate of service completions,* when there are n customers in the system. For some queueing systems, the values of the λ_n will be the same for all values of n, and the μ_n also will be the same for all n except for such small n (e.g., $n = 0$) that a server is idle. However, the λ_n and the μ_n also can vary considerably with n for some queueing systems.

For example, one of the ways in which λ_n can be different for different values of n is if potential arriving customers become increasingly likely to *balk* (refuse to enter the system) as n increases. Similarly, μ_n can be different for different n because customers in the queue become increasingly likely to *renege* (leave without being served) as the queue size increases. **Another example** in the Worked Examples section of the books'website illustrates a queueing system where both balking and reneging occur. This example then demonstrates how the general results for the birth-and-death process lead directly to various measures of performance for this queueing system.

Analysis of the Birth-and-Death Process

Because of its assumptions, the birth-and-death process is a special type of *continuous time Markov chain.* (See Sec. 16.8 for a description of continuous time Markov chains and their properties, including an introduction to the general procedure for finding steady-state probabilities that will be applied in the remainder of this section.) Queueing models that can be represented by a continuous time Markov chain are far more tractable analytically than any other.

Because Property 4 for the exponential distribution (see Sec. 17.4) implies that the λ_n and μ_n are mean rates, we can summarize these assumptions by the rate diagram shown in Fig. 17.4. The arrows in this diagram show the only possible *transitions* in the state of the system (as specified by assumption 3), and the entry for each arrow gives the mean rate for that transition (as specified by assumptions 1 and 2) when the system is in the state at the base of the arrow.

Except for a few special cases, analysis of the birth-and-death process is very difficult when the system is in a *transient* condition. Some results about the probability distribution of $N(t)$ have been obtained, but they are too complicated to be of much practical use. On the other hand, it is relatively straightforward to derive this distribution *after* the system has reached a *steady-state* condition (assuming that this condition can be reached). This derivation can be done directly from the rate diagram, as outlined next.

Consider any particular state of the system n ($n = 0, 1, 2, \ldots$). Starting at time 0, suppose that a count is made of the number of times that the process enters this state and the number of times it leaves this state, as denoted below:

$E_n(t)$ = number of times that process enters state n by time t.

$L_n(t)$ = number of times that process leaves state n by time t.