QUEUEING THEORY

TELETRAFFIC THEORY

NETWORK OF QUEUES

All of us, either directly or through the use of various machines that we have become dependent upon, wait for service in a variety of lines on a regular basis. Customers wait in lines at banks to be served by a bank teller; drivers wait in their cars in traffic jams or at toll booths; patients wait in doctors' waiting rooms; electronic messages wait in personal computers to be delivered over communication networks; telephone calls are put on hold to be answered by operators; computer programs are stored in computer memory to be executed by a time-sharing computer system; and so on. In many situations, scarce resources are to be shared among a collection of users who require the use of these resources at unspecified times. They also require the use of these resources for random periods of time. This probabilistic nature of requests causes these requests to arrive while the resources are in use by other members of the user community. A mechanism must be put in place to provide an orderly access to the resources requested. The most common mechanism is to put the user requests in a waiting line or "queue." "Queueing theory" deals with the study of the behavior and the control of waiting lines. It provides us with the necessary mathematical structure and probability tools to model, analyze, study, evaluate, and simulate systems involving waiting lines and queues. It is a branch of applied mathematics, applied probability theory, and operations research. It is known under various names such as: queueing theory, theory of stochastic server systems, theory of systems of flows, traffic or teletraffic theory, congestion theory, and theory of mass service. Standard texts on queueing theory include Refs. 1–31. For a summary of many of the most important results in queueing theory, the reader is referred to a survey paper by Cooper (7). For a bibliography of books and survey papers on queueing theory see Refs. 8, 29. For nontechnical articles explaining queueing theory for the layman the reader is referred to Refs. 9, 26.

A typical queueing system can be described as one where customers arrive for service, wait for service, and, leave the system after being served. The service requests occur according to some stochastic process, and the time required for the server(s) to service a request is also probabilistically distributed. In general, arrivals and departures (i.e., service completions) cannot be synchronized, so waiting time may result. It is, therefore, critical to be able to characterize waiting time and many other important performance measures of a queueing system. For a typical queueing system, one is interested in answering questions such as: How long does a typical customer have to wait? What is the number of customers in the system at any given point in time? How large should the waiting room be to accommodate certain percentage of potential customers? How many servers are needed to keep the waiting time below a certain limit? What are subjective and economical advantages and disadvantages of modifying various parameters of the systems such as the number of servers or the size of the waiting room? How often is the server busy? Queueing theory attempts to answer these and other related questions through detailed mathematical analysis and provides us with the necessary tools to evaluate related performance measures.

The purpose of this article is to provide an introductory overview of the fundamental notions of queueing theory. The remaining sections of this article will discuss the following topics: a brief history of the development of queueing theory; applications of queueing theory; specification and characterization of queueing systems; notions of probability theory of importance to queueing theory; modeling and analysis of elementary queueing systems; references to more advanced topics; and a list of references.

HISTORY OF THE DEVELOPMENT OF QUEUEING THEORY

The English word "queue" is borrowed from the French word "queue" which itself is taken from the Latin word "cauda" meaning "tail." Most researchers and scientists in the field prefer the spelling "queueing" over "queuing." However, many American dictionaries and software spell checkers prefer the spelling "queuing." For further discussion of "queueing" vs. "queuing" spelling, see Refs. 27, 28. Queueing theory has been under development since the early years of this century. It has since progressed considerably, and today it is based upon a vast collection of results, methods, techniques, and voluminous literature. A good summary of the early history of queueing theory can be found in Ref. 6, pp. 20–25.

Historically, queueing theory originated as a very practical subject. It was developed to provide models to predict the behavior of systems that attempt to provide service for randomly arising demands. Much of the early work was developed in relation with problems in telephone traffic engineering. The pioneering work of Agner Krarup Erlang, from 1909 to 1929, laid the foundations of modern teletraffic and queueing theory. Erlang, a Danish mathematician and engineer who worked for the Copenhagen Telephone Exchange, published his first article in 1909 on the application of probability theory to telephone traffic problems (10). Erlang's work soon drew the attention of other probability theorists such as T. C. Fry and E. C. Molina in the 1920s, who expanded much of Erlang's work on the application of the theory to telephone systems. Telephony remained one of the principal applications until about 1950.

In the years immediately following World War II, activity in the fields of probability theory and operations research (11, 12) grew rapidly, causing a new surge of interest in the subject of queueing theory. In the late 1950s, queueing theory became one of the most popular subjects within the domains of applied mathematics and applied probability theory. This popularity, however, was fueled by its mathematical challenges and not by its applications. Clever and elegant mathematical techniques has enabled researchers (such as Pollaczek, Kolmogorov, Khin-

J. Webster (ed.), *Wiley Encyclopedia of Electrical and Electronics Engineering*. Copyright © 2007 John Wiley & Sons, Inc.

chine, Crommelin, and Palm) to derive exact solutions for a large number of mathematical problems associated with models of queueing systems. Regrettably, in the period of 1950–1970, queueing theory, which was originated as a very practical subject, had become of little direct practical value.

Since the 1970s there has seen a rebirth and explosion of queueing theory activities with an emphasis on practical applications. The performance modeling and analysis of computer systems and data transmission networks opened the way to investigate queues characterized by complex service disciplines and interconnected systems. Most of the theoretical advances since the 1970s are directly attributable to developments in the area of computer systems performance evaluation as represented in Refs. 13–16.

APPLICATIONS OF QUEUEING THEORY

Interest in queueing theory has often been stimulated by practical problems and real world situations. Queueing theory concepts have applications in many disciplines such as telephone systems traffic engineering, migration and population models in biology, electrical and fluid flow models, clustering models in chemistry, manufacturing systems, computer systems, digital data transmission systems, flow through communication networks, inventory control, time sharing and processor sharing computer systems, telecommunications, machine repair, taxi stands, aircraft landing, loading and unloading ships, scheduling patients in hospitals, factory production flow, intelligent transportation systems, call centers, and so on. There are many other important applications of the queueing theory as presented in Refs. 1–6 and 13–16. We elaborate further on only two of these applications in this section.

Queueing theory has played a major role in the study of both packet switching and circuit switching communication networks. Queueing arises naturally in packet switching networks where user messages are broken into small units of transmission called packets. Packets arriving at various intermediate network nodes, on the way to their final destination, are buffered in memory, processed to determine the appropriate outgoing route, and then are transmitted on the chosen outgoing link when their time for transmission comes up. If, for example, the chosen outgoing link is in use when it is time for a given packet to be transmitted, then that packet must be kept in the memory (i.e., queued) until the link becomes available. The time spent in the buffer waiting for transmission is an important measure of system performance. This waiting time depends on various parameters such as nodal processing power, transmission link speed, packet lengths, traffic rates in terms of packets per second, and so on. Queueing theory provides the necessary mathematical tools to model and analyze such queueing configurations.

For another example of application of queueing theory consider a typical bank and the mechanism that bank management has put in place to direct incoming customers to the available bank tellers. In some banks, each teller has his or her own queue and incoming customers are free to join the waiting line of any of the tellers based on some personal preferences. Some customers often join the shortest queue, and some join the queue of a particular teller that they personally know, whereas others may join the queue of the teller that is perceived to be the fastest. On the other extreme, some banks (via the use of various directional signs and/or ropes) direct all the incoming customers into a single waiting line that feeds all the tellers. The customer at the head of this queue is then served by the next available bank teller. The question now becomes which one of these two modes of operation is more appropriate. The answer strongly depends on such parameters as the performance measures that the bank management is interested in optimizing, the number and the speed of the tellers, the type of banking transactions, and the number of incoming customers visiting the bank in a typical day. Similar issues arise in other cases such as supermarket checkout counters, fast-food restaurants, airport landing and departure schedules, and multiprocessor computer systems. Queueing theory methods enable us to model, analyze, and decide on the best strategy for such applications.

SPECIFICATION AND CHARACTERIZATION OF QUEUEING SYSTEMS

Figure 1 represents the basic elements of a queueing system. As shown in Fig. 1, a basic queueing system is one where members of a population (i.e., customers or entities of some kind) arrive at a service station to receive service of some type. After receiving service, the units depart the service facility. A "queue" or waiting line is developed whenever the service facility cannot service all the units requiring service. Although many queueing systems may be represented by similar diagrams, an accurate representation of such a system requires a detailed characterization of the underlying parameters and processes.

Key Parameters and Varieties of Queueing Systems

To fully describe a queueing system analytically, various aspects and parameters of the system must be known. The most important of them are presented here.

The Arrival Pattern. Let the successive customers arrive to the system at times $t_1, t_2, t_3, \ldots$, where $0 \leq t_1 < t_2 < t_3 < \cdots < t_n < \cdots$. Then we define $y_i = t_{i+1} - t_i$, where $i = 1, 2, 3, \ldots$, as the interarrival times of the customers. We normally assume that arrival times form a stochastic process and that the interarrival times, y_i, are independent and identically distributed (iid) according to probability distribution function $A(\cdot)$, where $A(\tau) = P(y_i \leq \tau)$. Function $A(\cdot)$ is then referred to as the interarrival time distribution or simply the arrival distribution. Additional information such as whether each arrival event contains one or a group of customers of fixed or random size (i.e., "bulk arrivals") can also be specified if applicable.

Customer Population and Behavior. The customer population, or the source of the customers, can either be finite or infinite. Infinite customer populations are normally easier to describe mathematically and analyze their performance analytically. This is because in a finite population source

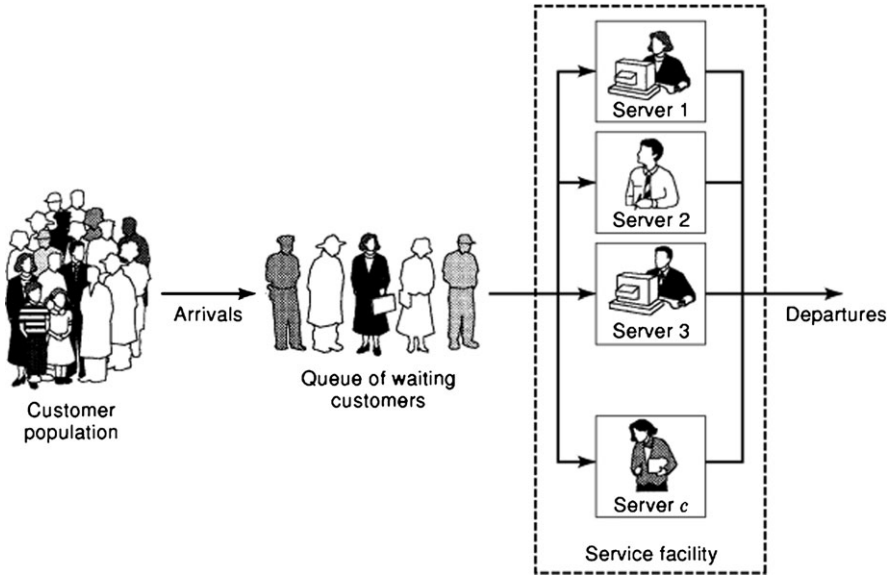

Figure 1. Basic elements of a typical queueing system.

model, the number of customers in the system affects the arrival rate which in turn makes the analysis more difficult. In addition to the properties of the entire customer population, behavior of individual customers could also be of importance and, therefore, must be formally specified. For example, if a customer decides not to join the system after seeing the size of the queue, it is said that the customer has "balked." Or, for example, a customer is said to have "reneged" if, after having waited in the queue for some time, he or she becomes impatient and leaves the system before his service begins. Customers, if allowed, may "jockey" from one queueing system to another (with a perceived shorter waiting time, for example).

The Service Mechanism. The queue's service mechanism is described by specifying the number of servers, c, and the stochastic characterization of the service times. It is normally assumed that the service times of successive customers, $x_1, x_2, x_3, \ldots$, are iid with probability distribution $B(\cdot)$, where $B(\tau) = P(x_i \leq \tau)$, and are also independent of the interarrival times $y_1, y_2, y_3, \ldots$. Additional information such as whether the customers are served individually or in groups (of fixed or random size) can also be specified if applicable.

The Queueing Discipline. The queueing discipline is the description of the mechanism for determining which of the waiting customers gets served next, along with the associated rules regarding formation of the queue. The most basic queueing disciplines are listed and described below:

1. First-Come First Served (*FCFS*) or First-In First-Out (*FIFO*) The waiting customers are served in the order of their arrival times.
2. Last-Come First-Served (*LCFS*) or Last-In First-Out (*LIFO*) The customer who has arrived last is chosen as the one who gets served when a server becomes available.
3. Service in Random Order (*SIRO*) or Random Selection for Service (*RSS*) The customer to be served next is chosen stochastically from the waiting customers according to a uniform probability distribution. In general, the probability distribution used to choose the next customer could be any discrete probability distribution.
4. Priority (*PR* or *PRI*) There could also be some notion of priority in the queueing system where the customer population is divided in two or more priority classes. Any waiting member of a higher priority class is chosen to be served before any customer from a lower priority class. Queueing systems with priority classes are divided into two types. Under a "preemptive priority" discipline, whenever a higher priority customer arrives while a lower priority customer is in service, the lower priority customer is preempted and is taken out of service without having his service completed. In this case, the preempted customer is placed back in the queue ahead of all customers of the same class. Under the "non-preemptive priority" discipline, once the service of any customer is started, it is allowed to be completed regardless of arrivals from higher priority classes. Moreover, the preemptive priority queueing systems can further be divided into two types. Under the discipline of "preemptive resume," whenever a preempted customer reenters service he simply continues his service where he left off. Under "preemptive repeat," a preempted customer draws a new value of service time from the service time distribution each time it reenters service.

Maximum Number of Customers Allowed. In many systems the capacity of queueing system is assumed to be infinite, which implies that every arriving customer is allowed to join the queue and wait until served. However, in many real-life situations, the queueing systems have either no or only a finite amount of capacity for customers to wait.

In a queueing system with no room for customers to wait, whenever all the servers are busy, any additional arriving customer is turned away; this type of system is referred to as "loss systems." Loss systems have been used to model the behavior of many dial-up telephone systems and telephone switching equipment. Queueing systems with a positive but finite waiting room have been deployed to characterize the performance of various computing and telecommunications systems where the finite waiting room models the finite amount of memory or buffer present in such real-world systems.

Number of Servers. In general a queueing system can have either one, or finitely many, or an infinite number of servers. "Single-server systems" are the simplest ones where a maximum of one user can be served at any given point in time. A "multiserver system" contains c servers, where $0 < c < \infty$, and can serve up to c simultaneous customers at any given point in time. An "infinite-server system" is one in which each arriving customer is immediately provided with an idle server.

Performance Measures

In any queueing system there are many performance tradeoffs to be considered. For example, if the number of servers in the system is so large that queues rarely form, then the servers are likely to be idle a large fraction of time, resulting in wasting of resources and extra expense. On the other hand, if almost all customers must join long queues, and servers are rarely idle, there might be customer dissatisfaction and possibly lost customers which again has negative economical consequences. Queueing theory provides the designer the necessary tools to analyze the system and ensure that the proper level of resources are provided in the system while avoiding excessive cost. The designer can accomplish this, for example, by considering several alternative system architectures and by evaluating each by queueing theory methods. In addition, the future performance of an existing system can also be predicted so that upgrading of the system can be achieved in a timely and economical fashion. For example, an analytical queueing model of a computer communication network might indicate that, in its present configuration, it cannot adequately support the expected traffic load two years in the future. The model may make it possible to evaluate different alternatives for increased capacity such as increasing the number of nodes in the network, increasing the computing power of existing nodes, providing more memory and buffer space in the network nodes, increasing the transmission speeds of the communication links, or increasing the number of communication links. Determining the most appropriate solution can be done through careful evaluation of various performance measures of the queueing systems.

The following performance measures represent some of the most common and important aspects of queueing systems which are normally investigated:

1. The Queue Length This performance measure is related to the number of customers waiting in the system. Some authors use this term to represent only the number of customers in the queue proper (i.e., not including the one or more customers who are being served), and others use it to represent the total number of customers in the system. In the former case it is often referred to as the "queue length," and in the latter case it is often referred to as the "number in the system."
2. The Waiting Time This performance measure is related to the amount of time spent by a customer in the system. This term is used in two different ways. Some authors use the term to refer to the total time spent by a customer in the queueing system, which is the sum of the time spent by the customer in the waiting line before service and the service time itself. Others define it as only the time spent in the queue before the service. In the former case it is often referred to as the "system time," and in the latter case it is often referred to as the "queueing time."
3. The Busy Period This is the length of time during which the server is continuously busy. Any busy period begins when a customer arrives at an empty system, and it ends when the number of customers in the system reaches zero. The time period between two successive busy periods is referred to as the "idle period" for obvious reasons.

Kendall's Notation for Queueing Systems

It is a common practice to use a short-hand notation of the form $A/B/c/K/m/Z$ to denote various aspects of a queueing system. This notation is referred to as *Kendall's notation*. This type of short-hand was first developed by Kendall (17) and later extended by Lee (18). It defines some of the basic parameters which must be known about a queue in order to study its behavior and analyze its performance. In Kendall's notation $A/B/c/K/m/Z$, A describes the interarrival time distribution, B describes the service time distribution, c is the number of (parallel) servers, K is the maximum number of customers allowed in the system (waiting plus in service), m is the size of the customer population, and Z describes the queue discipline. The traditional symbols used in the first and second positions of Kendall's notation, and their meanings, are:

M	Exponentially distributed interarrival time or service time distribution
D	Deterministic (i.e., constant) interarrival time or service time distribution
E_k	k-stage Erlangian (Erlang-k) interarrival time or service time distribution
H_k	k-stage Hyperexponential interarrival time or service time distribution
G	General interarrival or service time distribution

The third, fourth, and fifth positions in Kendall's notation could be any positive integer. The traditional symbols used in the last position of Kendall's notation are: FCFS, FIFO, LCFS, LIFO, SIRO, RSS, PR, and PRI, as described earlier in this section; and also GD, which refers to a general queue discipline.

As an example of Kendall notation, an M/D/2/50/∞/SIRO queueing system is one with exponential interarrival time, constant service time, 2 parallel servers, a system capacity of 50 (i.e., a maximum of 48 in the queue and 2 in service), a customer population that is infinitely large, and the waiting customers are served in a random order.

Whenever the last three elements of Kendall's notation are omitted, it is meant that $K = \infty$, $m = \infty$, and Z = FCFS (i.e., there is no limit to the queue size, the customer source is infinite, and the queue discipline is FCFS). As an example of the shorter version of Kendall's notation, an $M/M/1$ queue has Poisson arrivals, exponential service time, and 1 server, there is no limit to the queue size, the customer source is infinite, and the queue discipline is FCFS.

It should be noted that although Kendall's notation is quite useful and very popular, it is not meant to characterize all possible models and configurations of queueing systems. For example, Kendall's notation is normally not used to indicate bulk arrivals, or queues in series, and so on.

NOTIONS OF PROBABILITY THEORY OF IMPORTANCE TO THE STUDY OF QUEUES

Probability theory has a major and fundamental role in the study and analysis of queueing models. As mentioned earlier, queueing theory is considered a branch of applied probability theory. It is assumed here that the reader is familiar with the basic notions of elementary probability theory such as notions of events, probability, statistical independence, distribution and density functions, and expectations or averages. The reader is referred to Ref. 19 for a complete treatment of probability theory. Here we discuss a few aspects of probability notions which are of great importance to the study of queues.

Probability Distributions of Importance to Queueing Theory

As is indicative of Kendall's notation, queueing theory deals with a large number of different types of probability distributions to mathematically model the behavior of customer interarrival times and the customer service times. In the rest of this section, we briefly describe some of the most important probability distributions that are used often in various queueing theory analysis.

Exponential Probability Distribution. The probability distribution most commonly assumed for customer interarrival time and for customer service times in queueing models is the exponential distribution. This popularity is due to its pleasant mathematical properties which often result in much simplification of the analytical work. A continuous random variable X has an exponential distribution with parameter $\lambda > 0$ if its density function $f(\cdot)$ is defined by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geqslant 0 \\ 0 & \text{for } x < 0 \end{cases} \tag{1}$$

It distribution function is given by

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{for } x \geqslant 0 \\ 0 & \text{for } x < 0 \end{cases} \tag{2}$$

Both its mean and its standard deviation are equal to $1/\lambda$.

The exponential distribution is unique among the continuous distributions because it has the so-called "memoryless property" or "Markov property." The memoryless property is that if we know that a random variable has an exponential distribution, and we know that the value of the random variable is at least some value, say t, then the distribution for the remaining value of the variable (i.e., the difference between the total value and t) has the same exponential distribution as the total value. That is,

$$P(X > t + h | X > t) = P(X > h) \quad \text{for } t > 0, h > 0 \tag{3}$$

Another interpretation of Eq. (3) is that, if X is the waiting time until a particular event occurs and t units of time have produced no event, then the distribution of further waiting time is the same as it would be if no waiting time had passed; that is, the system does not "remember" that t time units have produced no "arrival."

Poisson Probability Distribution and Poisson Random Process. Poisson random variable is used in many applications where we are interested in counting the number of occurrences of an event (such as arrivals to a queueing system) in a certain time period or in a region of space. Poisson random variables also appear naturally in many physical situations. For example, the Poisson probability mass function gives an accurate prediction for the relative frequencies of the number of particles emitted by a radioactive mass during a fixed time period. A discrete random variable X is said to have a Poisson distribution with parameter $\lambda > 0$ if X has a probability mass function of the form

$$P(k; \lambda) = P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad \text{for } k = 0, 1, 2, 3, \ldots \tag{4}$$

Both the mean and the standard deviation of the Poisson random variable are equal to λ.

Now consider a situation in which events occur at random instants of time at an average rate of λ events per second. For example, an event could represent the arrival of a customer to a service station or the breakdown of a component in some system. Let $N(t)$ be the number of event occurrences in the time interval $[0, t]$. $N(t)$ is then a nondecreasing, integer-valued, continuous-time random process. Such a random process is said to be a Poisson process if the number of event occurrences in the time interval $[0, t]$ has a Poisson distribution with mean λt. That is,

$$P(N(t) = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \quad \text{for } k = 0, 1, 2, 3, \ldots \tag{5}$$

Like the exponential distribution, Poisson process also has a number of unique properties which has made it very at-

tractive for analytical studies of queueing systems. In particular, Poisson process has a "memoryless property"; occurrence of events during a current interval of time is independent of occurrences of events in previous intervals. In other words, events occurring in nonoverlapping intervals of time are independent of each other. Furthermore, the interevent times (i.e., interarrival times in case of queueing system) in a Poisson process from an iid sequence of exponential random variables with mean $1/\lambda$.

Erlang-k Probability Distribution. A. K. Erlang (10) used a special class of gamma random variables (19), now often called "Erlang-k" or "k-stage Erlangian," in his study of delays in telephone traffic. A random variable, T, is said to be an Erlang-k random variable with parameter λ or to have an Erlang distribution with parameters k and λ, if T is gamma random variable with the density function f given by

$$f(x) = \begin{cases} \dfrac{\lambda k(\lambda k x)^{k-1}}{(k-1)!} e^{-\lambda k x} & \text{for } x \geqslant 0 \\ 0 & \text{for } x < 0 \end{cases} \tag{6}$$

The mean and variance of Erlang-k random variable are $1/\lambda$ and $1/(k\lambda^2)$, respectively. An Erlang-k random variable can be obtained by adding k independent exponentially distributed random variables each with parameter λk. The physical model that Erlang had in mind was a service facility consisting of k identical independent service substations connected in series one after another, each with an exponential distribution of service time. He wanted this special facility to have the same average service time as a single service facility whose service time was exponential with parameter λ. Thus the service time, T, for the facility with k stages could be written as the sum of k exponential random variables, each with parameter λk.

Hyperexponential Probability Distribution. If the service time of a queueing system has a large standard deviation relative to the mean value, it can often be approximated by a hyperexponential distribution. The model representing the simplest hyperexponential distribution is one with two parallel stages in the facility; the top one having exponential service with parameter μ_1, and the bottom stage having exponential service with parameter μ_2. A customer entering the service facility chooses the top stage with probability α_1 or the bottom stage with probability α_2, where $\alpha_1 + \alpha_2 = 1$. After receiving service at the chosen stage, with the service time being exponentially distributed with average service rate μ_i, the customer leaves the service facility. A new customer is not allowed to enter the facility until the original customer has completed service. The probability density function for the service time, the probability distribution function, mean, and variance are given by

$$f_x(t) = \alpha_1\mu_1 e^{-\mu_1 t} + \alpha_2\mu_2 e^{-\mu_2 t} \qquad \text{for } t \geq 0 \tag{7}$$

$$F_x(t) = P(x \leq t) = 1 - \alpha_1 e^{-\mu_1 t} - \alpha_2 e^{-\mu_2 t} \quad \text{for } t \geq 0 \tag{8}$$

$$E[x] = \frac{\alpha_1}{\mu_1} + \frac{\alpha_2}{\mu_2} \tag{9}$$

$$\text{Var}[x] = \frac{2\alpha_1}{\mu_1^2} + \frac{2\alpha_2}{\mu_2^2} - \left(\frac{\alpha_1}{\mu_1} + \frac{\alpha_2}{\mu_2}\right)^2 \tag{10}$$

The two-stage hyperexponential distribution described above can be generalized to k stages for any positive integer greater than 2.

Notions of Transient and Steady State

Analysis of a queueing system often involves the study of the system's characteristics over time. A system is defined to be in "transient state" if its behavior and associated performance measures are dependent on time. This usually occurs at the early stages of the operation of the system where its behavior is heavily dependent on the initial state of the system. A system is said to be in "steady state" or "equilibrium" when the behavior of the system becomes independent of time. This usually occurs after the system has been in operation for a long time, and the influence of initial conditions and of the time since start-up have diminished. In steady state, the number of customers in the system and in the queue are independent of time.

A necessary condition for a queueing system to reach steady state is that the elapsed time since the start of the operation is mathematically long enough (i.e., the limit as time tends to infinity). However, this condition is not sufficient to guarantee that a queueing system is in steady state. In addition to elapsed time, particular parameters of the queueing system itself will have an effect on whether and when the system reaches steady state. For example, if the average arrival rate of customers is higher than the overall average service rate of the system, then the queue length will continue to grow forever and steady state will never be reached. Although many authors have studied the transient behavior of queueing systems, the majority of the key results and existing literature deal with steady-state behavior of queueing systems.

Random Variables of Interest

In this section we define and list the key random variables and associated notations used in queueing theory and in the rest of this article. Some of the primary random variables and notations are graphically illustrated in Fig. 2 and many more are listed in Table 1. Clearly, there are some obvious relationships between some of the random variables listed in Fig. 2 and/or Table 1. For example, with respect to the number of customers in the system, we must have

$$N(t) = N_q(t) + N_s(t) \tag{11}$$

and

$$N = N_q + N_s \tag{12}$$

In Eq. (12), it is assumed that the queueing system has reached the steady state. It should, however, be noted that although the system is in steady state, quantities N, N_q, and N_s are random variables; that is, they are not constant and have probability distributions associated with them. In other words, "steady state" means that the probabilities are independent of time but not that the system becomes deterministic.

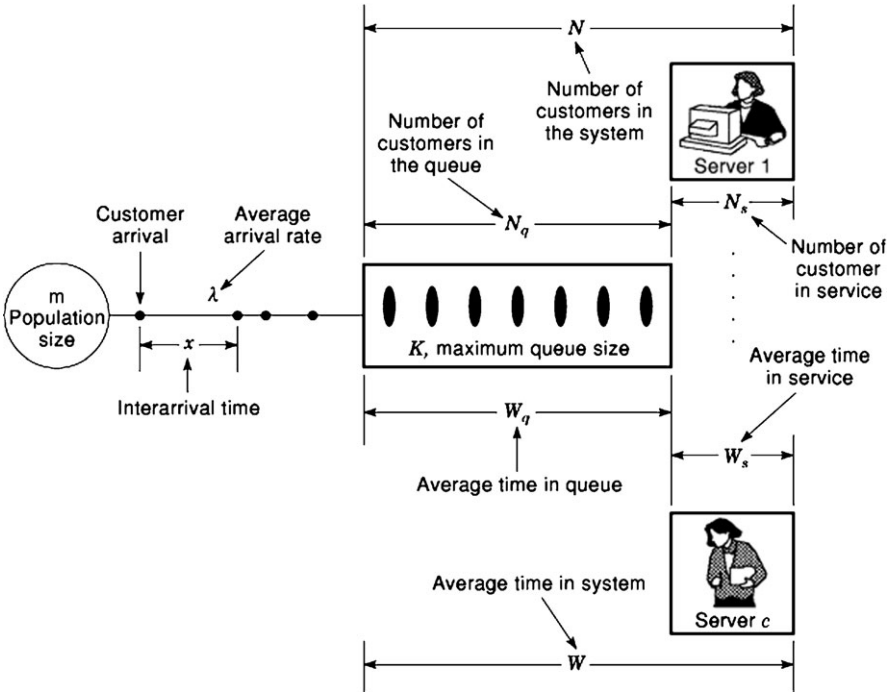

Figure 2. Graphical representation of some variables of importance to queueing theory.

Table 1. Definition of Some of the Key Variables of Importance to Queueing Theory

y	Random variable representing customer interarrival times
λ	Average arrival rate of customers to the system
$1/\lambda$	Average interarrival time of customers to the system
x	Random variable representing customer service times
μ	Average service time per customer
$1/\mu$	Average service rate per customer
c	Number of servers
ρ	Traffic intensity or offered load
$N(t) \to N$	Random variables representing number of customers in the system at time t and in steady state
$N_q(t) \to N_q$	Random variables representing number of customers in the queue at time t and in steady state
$N_s(t) \to N_s$	Random variables representing number of customers in service at time t and in steady state
$P_n(t) \to P_n$	Transient and steady-state probability of having exactly n customers in the system
L	Average steady-state number of customers in the system
L_q	Average steady-state number of customers in the queue
L_s	Average steady-state number of customers in service
$w(x)$	Random variable representing steady state probability distribution of the waiting time in the system
W	Average steady-state time spent by a customer in the system
W_q	Average steady-state time spent by a customer in the queue
W_s	Average steady-state time spent by a customer in service

Applying expectations operation to both sides of Eq. (12), we get

$$L = L_q + L_s \tag{13}$$

There are similar obvious relationships between some of the random variables related to waiting times. For example, the total time in the queueing system for any customer is the sum of his waiting time in the queue and his service time, that is,

$$W = W_q + W_s \tag{14}$$

We are clearly interested in studying relationships between other random variables and parameters of the interest which might not be as obvious as those given in Eqs. (11)–(14). Development of such relationships are a major byproduct of modeling and analysis of queueing systems, as will be discussed in the next section.

MODELING AND ANALYSIS OF ELEMENTARY QUEUEING SYSTEMS

In this section we present, in some detail, some of the key techniques used by queueing theory community to model and analyze some of the elementary queueing models. In particular, we will illustrate the application of birth-and-death stochastic processes to the analysis of these models.

Little's Formula

Little's formula (which is also known as "Little's result" and "Little's theorem") is one of the most fundamental and often used results in queueing theory. It provides a simple, but very general, relationship between the average waiting time and the average number of customers in a queueing system. Its first rigorous proof in its general form was given by J. D. C. Little (20). Its validity and proofs of some special cases, however, were known to researchers prior to Little's proof. Consider an arbitrary queueing system in

steady state. Let L, W, and λ be the average number of customers in the system, average time spent by customers in the system, and average number of customer arrivals per unit time, respectively. Little's theorem states that

$$L = \lambda W \tag{15}$$

regardless of the interarrival and service time distributions, the service discipline, and any dependencies within the system.

Rigorous proof of Little's theorem is given in every standard queueing theory text (1–6). What follows is an intuitive justification of Little's result given in Ref. 12. Suppose that the system receives a reward (or penalty) of 1 for every unit of time that a customer spends in it. Then the total expected reward per unit time is equal to the average number of customers in the system, L. On the other hand, the average number of customers coming into the system per unit time is λ; the expected reward contributed by each customer is equal to his average residence time, W. Since it does not matter whether the reward is collected on arrival or continuously, we must have $L = \lambda W$. A different interpretation of Little's result is obtained by rewriting it as $\lambda = L/W$. Since a customer in the system remains there for an average time of W, his average rate of departure is $1/W$. The total average departure rate is, therefore, L/W. Thus, the relation holds if the average arrival rate is equal to the average departure rate. But the latter is clearly the case since the system is in equilibrium.

It is important to note that we have not even specified what constitutes "the system," nor what customers do there. It is just a place where customers (entities) arrive, remain for some time, and then depart after having received service. The only requirement is that the processes involved should be stationary (i.e., system should be in steady state). Therefore, we can apply Little's theorem not only to the entire queueing system [as represented by Eq. (15)], but also to particular subsections of it. For example, applying Little's theorem to only the waiting line portion of a $G/G/c$ queueing system, where $1 \leq c \leq \infty$, results in

$$L_q = \lambda W_q \tag{16}$$

where L_q and W_q are as defined in Table 1. Now consider another situation, where the "system" is defined as the "set of c servers" in a $G/G/c$ queueing system, where $1 \leq c \leq \infty$. Since every incoming customer enters a server eventually, the rate of arrivals into the "set of c servers" is also λ. The average time a customer spends in the system here is simply $1/\mu$. According to Little's theorem, the average number of customers in the system is therefore λ/μ. Thus in any $G/G/c$ or $G/G/\infty$ system in steady state, the average number of busy servers is equal to the traffic intensity, ρ. When $c = 1$, the average number of busy servers is equal to the probability that the server is busy. Therefore, in any single-server system in the steady state we have

$$P(\text{there are customers in the system}) = \rho \tag{17}$$

$$P(\text{system is idle}) = 1 - \rho \tag{18}$$

Birth-and-Death Process

Most elementary queueing models assume that the inputs (i.e., arriving customers) and outputs (i.e., departing customers) of the queueing system occur according to the so-called "birth-and-death process." This important process in probability theory has application in other areas also. However, in the context of queueing theory, the term "birth" refers to the arrival of a new customer and the term "death" refers to the departure of a served customer. The state of the system at time t, for $t \geq 0$, is given by random variable $N(t)$ defined as the number of customers in the system at time t. Thus the birth-and-death process describes probabilistically how $N(t)$ changes at t increases.

Formally speaking, a stochastic process is a birth-and-death process if it satisfies the following three assumptions: (1) Given $N(t) = n$, the current probability distribution of the remaining time until the next birth is exponentially distributed with parameter λ_n for $n = 0, 1, 2, \ldots$; (2) given $N(t) = n$, the current probability distribution of the remaining time until the next death is exponentially distributed with parameter μ_n for $n = 0, 1, 2, \ldots$; and (3) only one birth or death can occur at a time. Figure 3, which shows the state transition diagram of a birth-and-death process, graphically summarizes the three assumptions just described. The arrows in this diagram show the only possible transitions in the state of the system, and the label for each arrow gives the mean rate for the transition 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. On the other hand, it is relatively easy to derive the probability distribution of the number of customers in the system in steady state. In steady state, the probability of finding the system in a given state does not change with time. In particular, the probability of there being more than k customers in the system is constant. The transition from state k to state $k + 1$ increases this probability, and the transition from state $k + 1$ to state k decreases it. Therefore, these two transitions must occur at the same rate. If this were not so, the system would not be in steady state. This yields to the following key principle: In equilibrium, the average rate into any state is equal to the average rate out of that state. This basic principle can be used to generate a set of equations called the "balance equations." After constructing the balance equations for all the states in terms of the unknown probabilities P_n, this system of equations can then be solved to find these probabilities. As shown in Fig. 3, there are only two transitions associated with state zero which result in the following balance equation for that state:

$$\mu_1 P_1 = \lambda_0 P_0 \tag{19}$$

There are four transitions associated with state 1 resulting in the following balance equation for that state:

$$\lambda_0 P_0 + \mu_2 P_2 = (\lambda_1 + \mu_1) P_1 \tag{20}$$

Balance equations for states $n \geq 2$ are similar to that of state 1 and can be easily be generated by inspecting the associated transitions in Fig. 3. This collection of balance

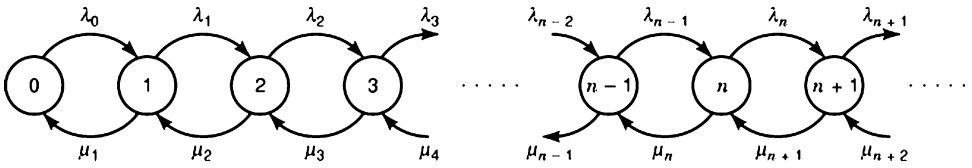

Figure 3. State transition diagram for a birth-and-death process.

equations along with the auxiliary equation

$$\sum_{n=0}^{\infty} P_n = 1 \tag{21}$$

can be solved for P_n, $n = 0, 1, 2, 3, \ldots$, resulting in the following set of steady-state probabilities for the number of customers in the system:

$$P_0 = \frac{1}{1 + \sum_{n=1}^{\infty} C_n} \tag{22}$$

$$P_n = C_n P_0 \quad \text{for } n = 1, 2, 3, \ldots \tag{23}$$

where

$$C_n = \frac{\lambda_{n-1}\lambda_{n-2}\cdots\lambda_0}{\mu_n \mu_{n-1}\cdots\mu_1} \quad \text{for } n = 1, 2, 3, \ldots \tag{24}$$

Given these expressions for the steady-state probability of number of customers in the system, we can derive the average number of customers in the system by

$$L = \sum_{n=0}^{\infty} n P_n \tag{25}$$

These steady-state results have been derived under the assumption that the λ_n and μ_n parameters are such that the process actually can reach a steady-state condition. This assumption always holds if $\lambda_n = 0$ for some value of n, so that only a finite number of states (those less than n) are possible. It also always holds when $\lambda_n = \lambda$ and $\mu_n = \mu$ for all n and when $\rho = \lambda/\mu < 1$.

M/*M*/1 Queue

Consider the simplest model of a nontrivial queueing model. This model assumes a Poisson arrival process (i.e., exponentially distributed interarrival times), an exponentially distributed service time, a single server, infinite queue capacity, infinite population of customers, and FCFS discipline. If the state of the system at time t, for $t \geq 0$, is given by the random variable $N(t)$, defined as the number of customers in the system at time t, it represents a birth-and-death process with rates

$$\lambda_n = \lambda \quad \text{and} \quad \mu_n = \mu \quad \text{for } n = 0, 1, 2, 3, \ldots \tag{26}$$

Therefore, by using Eqs. (22)–(24), we get

$$P_n = (1 - \rho)\rho^n \quad \text{for } n \geq 0, \rho < 1 \tag{27}$$

where $\rho = \lambda/\mu$. The mean number of customers in the system can now be computed as

$$L = E[N] = \sum_{k=0}^{\infty} k p_k = \frac{\rho}{1 - \rho} \tag{28}$$

Having found the mean number of customers in the system, we can now use Little's formula to determine the average total waiting time, W, as follows:

$$W = \frac{L}{\lambda} = \frac{\rho/\lambda}{1 - \rho} = \frac{1/\mu}{1 - \rho} = \frac{1}{\mu - \lambda} \tag{29}$$

Behavior of the average number of customers in the system (i.e., L) and the normalized average waiting time (i.e., $W\mu$) for the *M*/*M*/1 queue as a function of traffic intensity, ρ, has been graphically shown in Fig. 4. Note that the average waiting time and the queue length explode as traffic intensity approaches 1. Therefore, the *M*/*M*/1 queue is stable only if $0 \leq \rho < 1$.

Other Elementary Queueing Systems

There are a number of other single-queue models whose steady-state behavior can be determined via birth-and-death process techniques. We briefly mention the most important ones and refer the reader to standard texts on queueing theory (1–6) for detailed analysis and the associated mathematical expressions. Lack of space prevents us from listing all the associated results and formulas in these areas. The reader is referred to Ref. 3 (pp. 400–409) for a tabular listing of all the key formulas related to important queueing models.

***M*/*M*/1/*K*.** The *M*/*M*/1 model is somewhat unrealistic in the sense that, for example, no communication link can have an unlimited number of buffers. The *M*/*M*/1/*K* system is a more accurate model of this type of system in which a limit of K customers is allowed in the system. When the system contains K customers, arriving customers are turned away. This model can easily be analyzed by truncating the birth-and-death state diagram of the *M*/*M*/1 queue to only K states. This results in a birth-and-death process with coefficients

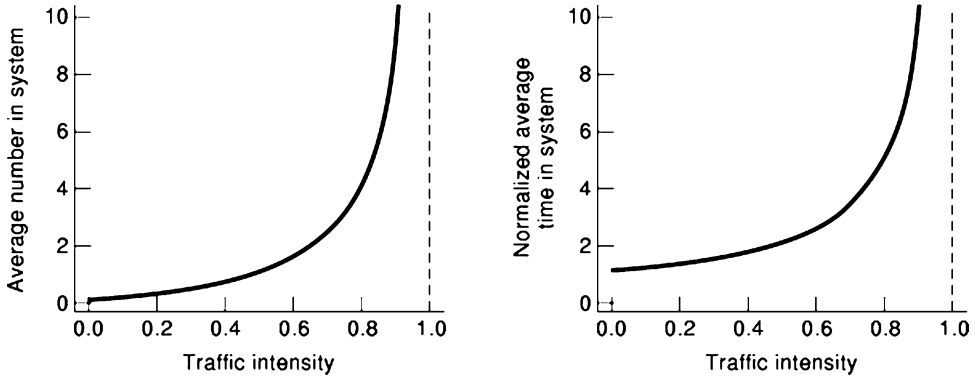

Figure 4. Performance characteristics of $M/M/1$ queue.

$$\lambda_n = \begin{cases} \lambda & \text{for } n = 0, 1, 2, 3, \ldots, K-1 \\ 0 & \text{for } n \geq K \end{cases} \tag{30}$$

and

$$\mu_n = \begin{cases} \mu & \text{for } n = 1, 2, 3, \ldots, K \\ 0 & \text{for } n \geq K \end{cases} \tag{31}$$

M/M/c. For this model we assume exponential interarrival times, exponential service times, and c identical servers. This system can be modeled as a birth-and-death process with the coefficients

$$\lambda_n = \lambda \quad \text{for } n = 0, 1, 2, \ldots \tag{32}$$

and

$$\mu_n = \begin{cases} n\mu & \text{for } n = 1, 2, 3, \ldots, c \\ c\mu & \text{for } n \geq c \end{cases} \tag{33}$$

Note that Eq. (33) agrees with Eq. (26) when $c = 1$; that is, for the $M/M/1$ queueing system, as it should. Historically, the expression of the probability that an arriving customer must wait is known as "Erlang's C Formula" or "Erlang's Delay Formula" (3, p. 404). Tables of values of Erlang's C Formula are often given in standard queueing texts; see, for example, Ref. 1 (pp. 320–323).

M/M/c/c. This system is sometimes called the "$M/M/c$ loss system" because customers who arrive when all the servers are busy are not allowed to wait for service and are lost. Each newly arriving customer is given his private server; however, if a customer arrives when all servers are occupied, that customer is lost; when modeling telephone calls, it is said that this is a system where blocked calls are cleared. The birth-and-death coefficients for this model are

$$\lambda_n = \begin{cases} \lambda & \text{for } n < c \\ 0 & \text{for } n \geq c \end{cases}$$

and

$$\mu_n = n\mu \quad \text{for } n = 1, 2, 3, \ldots, c$$

Historically, the expression for the probability that all servers are busy in an $M/M/c/c$ queueing system is referred to as "Erlang's B Formula" or "Erlang's Loss Formula" (3, p. 404). Tables of values of Erlang's B Formula are often given in standard queueing texts; see, for example, Ref. 1 (pp. 316–319).

***M/M/∞* Queueing System.** Mathematically speaking, an $M/M/\infty$ queueing system has an infinite number of servers which cannot be physically realized. $M/M/\infty$ queueing systems are used to model situations where a server is always immediately provided for each arriving customer. The coefficients of the associated birth-and-death process are given by

$$\lambda_n = \lambda \quad \text{for } n = 0, 1, 2, 3, \ldots \tag{34}$$

and

$$\mu_n = n\mu \quad \text{for } n = 1, 2, 3, \ldots \tag{35}$$

Solving the birth-and-death equations for the steady-state probability of number of customers in the queue results in

$$P_n = \frac{(\lambda/\mu)^n}{n!} e^{-\lambda/\mu} \quad \text{for } n = 0, 1, 2, 3, \ldots$$

Therefore, the number of customers in an $M/M/\infty$ queue is distributed according to a Poisson distribution with parameter λ/μ. The average number of customers in the system is simply $L = \lambda/\mu$ and the average waiting time is $W = 1/\mu$. This answer is obvious since if we provide each arriving customer his own server, then his time in the system is equal to his service time. $M/M/\infty$ models can be used to estimate the number of lines in use in large communications networks or as an estimate of values of $M/M/c$ or $M/M/c/c$ systems for large values of c.

***M/M/1/K/K* and *M/M/c/K/K* Queueing Systems.** These queueing systems, with a limited source model in which there are only K customers, is usually referred to as the

"machine repair model" or "machine interference model." One way to interpret these models is to assume that there is a collection of K machines, each of which has an up time which is exponentially distributed. The operating machines are outside of the system and enter the system only when they break down and thus need repair. The one repairman (or c repairmen) repairs the machines at an exponential rate. The coefficients of the associated birth-and-death process are

$$\lambda_n = \lambda(K - n) \quad \text{for } n = 0, 1, 2, \ldots, K - 1 \tag{36}$$

and

$$\mu_n = \mu \quad \text{for } n = 1, 2, \ldots, K \tag{37}$$

REFERENCES TO MORE ADVANCED TOPICS

The discussion of previous sections has been limited to some of the more elementary, but important, queueing models. However, the queueing theory literature currently contains a vast amount of results dealing with much more advanced and sophisticated queueing systems whose discussions are outside of the scope of this introductory article. The purpose of this section is to inform the reader of the existence of such advanced and complex models and to refer the interested reader to appropriate sources for further investigation.

Imbedded Markov Chain Queueing Models

Our discussion of queueing models in the previous section was limited to those whose probabilistic characterization could be captured by birth-and-death processes. When one ventures beyond the birth-and-death models into the more general Markov processes, then the type of solution methods used previously no longer apply. In the preceding sections we dealt mainly with queues with Poisson arrivals and exponential service times. These assumptions imply that the future evolution of the system will depend only on the present state of the system and not on the past history. In these systems, the state of the system was always defined as the number of customers in the system.

Consider the situation in which we like to study a queueing system for which the knowledge of the number of customers in the system is not sufficient to fully characterize its behavior. For example, consider a $D/M/1$ queue in which the service times are exponentially distributed, but the customer interarrival times are a constant. Then the future evolution of the system from some time t would depend not only on the number of customers in the system at time t, but also on the elapsed time since the last customer arrival. This is so because the arrival epoch of the next customer in a $D/M/1$ queue is fully determined by the arrival time of the last customer. A different and powerful method for the analysis of certain queueing models, such as the one mentioned above, is referred to as the "imbedded Markov chain" which was introduced by Kendall (17). The reader is referred to Refs. 1–6 for detailed discussion of imbedded Markov chain techniques and its application for analyzing such queueing systems as $M/G/1$, $GI/M/c$, $M/D/c$, $E_k/M/c$.

Queueing Systems with Priority

Queueing models with priority are those where the queue discipline is based on a priority mechanism where the order in which the waiting customers are selected for service is dependent on their assigned priorities. Many real queueing systems fit these priority-discipline models. Rush jobs are taken ahead of other jobs, important customers may be given precedence over others, and data units containing voice and video signals may be given higher priority over data units containing no real-time information in a packet switched computer communication network. Therefore, the use of queueing models with priority often provides much needed insight into such situations. The inclusion of priority makes the mathematical analysis of models much more complicated. There are many ways in which notions of priority can be integrated into queueing models. The most popular ones were defined earlier in this article under queue disciplines. They include such priority disciplines as non-preemptive priority, preemptive resume priority, and preemptive repeat priority (21).

Networks of Queues

Many queueing systems encountered in practice are queueing networks consisting of a collection of service facilities where customers are routed from one service center to another, and they receive service at some or all of these service facilities. In such systems, it is necessary to study the entire network in order to obtain information about the performance of a particular queue in the network. Such models have become very important because of their applicability to modeling computer communication networks. This is a current area of great research and application interest with many difficult problems. Networks of queues can be described as a group of nodes (say n of them) where each node represents a service center each with c_i servers, where $i = 1, 2, \ldots, n$. In the most general case, customers may arrive from outside the system to any node and may depart the system from any node. The customers entering the system traverse the network by being routed from node to node after being served at each node they visit. Not all customers enter and leave from the same nodes, or take the same path through the network. Customers may return to nodes previously visited, skip some nodes, or choose to remain in the system forever. Analytical results on queueing networks have been limited because of the difficulty of the problem. Most of the work has been confined to cases with a Poisson input and exponential service times and probabilistic routing between the nodes. The reader is referred to Ref. 22 for a complete treatment of network of queues.

Simulation of Queueing Systems

Very often, analytical solutions to many practical queueing models are not possible. This is often due to many factors such as the complexity of the system architecture, the nature of the queue discipline, and the stochastic characteristics of the input arrival streams and service times. For example, it would be impractical to develop analytical solutions to a multinode multiserver system where the customers are allowed to recycle through the system, the ser-

vice times are distributed according to truncated Gaussian distribution, and each node has its own complex queueing discipline. For analytically intractable models, it may be necessary to resort to analysis by simulation. Another area that simulation could be used for is those models in which analytical results are only available for steady state and one needs to study the transient behavior of the system.

Generally speaking, simulation refers to the process of using computers to imitate the operation of various kinds of real-world systems or processes. While simulation may offer a mechanism for studying the performance of many analytically intractable models, it is not without its disadvantages. For example, since simulation can be considered analysis by experimentation, one has all the usual problems associated with running experiments in order to make inferences concerning the real world, and one must be concerned with such things as run length, number of replications, and statistical significance. Although simulation can be a powerful tool, it is neither cheap nor easy to apply correctly and efficiently. In practice, there seems to be a strong tendency to resort to simulation from the outset. The basic concept is easy to understand, it is relatively easy to justify to management, and many powerful simulation tools are readily available. However, an inexperienced analyst will usually seriously underestimate the cost of many resources required for an accurate and efficient simulation study.

Viewing it from a high level, a simulation model program consists of three phases. The data generation phase involves the production of representative interarrival times and service times where needed throughout the queueing system. This is normally achieved by employing one of the many random number generation schemes. The so-called bookkeeping phase of a simulation program deals with (a) keeping track of and updating the state of the system whenever a new event (such as arrival or departure) occurs and (b) monitoring and recording quantities of interest such as various performance measures. The final phase of a simulation study is normally the analysis of the output of the simulation run via appropriate statistical methods. The reader is referred to Refs. 23 and 24 for a comprehensive look at simulation techniques.

Cyclic Queueing Models

This area deals with situations where a collection of queues is served by a single server. The server visits each queue according to some predetermined (or random) order and serves each queue visited for a certain amount of time (or certain number of customers) before traversing to the next queue. Other terms used to refer to this area of queueing theory are "round-robin queueing" or "queueing with vacations." As an example, a time-shared computer system where the users access the central processing unit through terminals can be modeled as a cyclic queue. The reader is referred to Ref. 25 and Section 5.13 of Ref. 1 for detailed discussion of cyclic queues.

Control of Queues

This area of queueing theory deals with optimization techniques used to control the stochastic behavior and to optimize certain performance measures of a queueing systems. Examples of practical questions that deal with this area of queueing theory include the following (22, Chap. 8): When confronted with the choice of joining one waiting line among many (such a supermarket checkout counter or highway tool booths), how does one choose the "best" queue? Should a bank direct the customers to form a single waiting line, or should each bank teller have his or her own queue? Should a congested link in a communication network be replaced with another link twice as fast, or should it be augmented with a second identical link working in parallel with the first one?

BIBLIOGRAPHY

1. R. B. Cooper *Introduction to Queueing Theory*, 2nd ed., New York: Elsevier/North-Holland, 1981.
2. D. Gross C. H. Harris *Fundamentals of Queueing Theory*, New York: Wiley, 1985.
3. L. Kleinrock *Queueing Systems, Volume I: Theory*, New York: Wiley-Interscience, 1975.
4. L. Kleinrock *Queueing Systems, Volume II: Computer Applications*, New York: Wiley-Interscience, 1976.
5. E. Gelenbe G. Pujolle *Introduction to Queueing Networks*, Paris: Wiley, 1987.
6. T. L. Saaty *Elements of Queueing Theory*, New York: McGraw-Hill, 1961.
7. R. B. Cooper Queueing theory. In D. P. Heyman and M. J. Sobel (eds.), *Stochastic Models*, Handbooks of Operations Research and Management Science, Vol. 2, New York: North-Holland, 1990.
8. N. U. Prabhu A bibliography of books and survey papers on queueing systems: theory and applications, *Queueing Systems*, **1**: 1–4, 1987.
9. M. A. Leibowitz Queues, *Sci. Am.*, **219** (2): 96–103, 1968.
10. A. K. Erlang The theory of probabilities and telephone conversations, *Nyt Tidsskrift Matematik*, Series B, **20**: 33–39, 1909.
11. F. S. Hillier G. J. Lieberman *Introduction to Operations Research*, 4th ed., Oakland, CA: Holden-Day, 1986.
12. H. A. Taha *Operations Research: An Introduction*, New York: Macmillan, 1971.
13. E. Gelenbe I. Mitrani *Analysis and Synthesis of Computer Systems*, New York: Academic Press, 1980.
14. J. F. Hayes *Modeling and Analysis of Computer Communications Networks*, New York: Plenum Press, 1984.
15. C. H. Sauer K. M. Chandy *Computer Systems Performance Modeling*, Englewood Cliffs, NJ: Prentice-Hall, 1981.
16. J. N. Daigle *Queueing Theory for Telecommunications*, Reading, MA: Addison-Wesley, 1992.
17. D. G. Kendall Stochastic processes occurring in the theory of queues and their analysis by the method of imbedded markov chains, *Ann. Math. Stat.*, **24**: 338–354, 1953.
18. A. M. Lee *Applied Queueing Theory*, London: Macmillan, 1966.
19. A. Leon-Garcia *Probability and Random Processes for Electrical Engineering*, Reading, MA: Addison-Wesley, 1989.
20. J. D. C. Little A proof for the queueing formula $L = \lambda$ W, *Oper. Res.*, **9**: 383–387, 1961.
21. N. K. Jaiswell *Priority Queues*, New York: Academic Press, 1968.

22. J. Walrand *An Introduction to Queueing Networks*, Englewood Cliffs, NJ: Prentice-Hall, 1988.
23. P. Bratley B. L. Fox L. E. Schrage *Guide to Simulation*, New York: Springer-Verlag, 1983.
24. A. M. Law W. D. Kelton *Simulation Modeling and Analysis*, 2nd ed., New York: McGraw-Hill, 1991.
25. H. Takagi *Analysis of Polling Systems*, Cambridge, MA: MIT Press, 1986.
26. "Queueing Theory," http://en.wikipedia.org/wiki/Queuing_theory.
27. Myron Hlynka,"What is the proper spelling — queueing or queuing?," http://www2.uwindsor.ca/ hlynka/qfaq.html.
28. Jeff Miller,"A Collection of Word Oddities and Trivia," http://members.aol.com/gulfhigh2/words6.html.
29. Myron Klynka,"Myron Hlynka's Queueing Theory Page," http://www2.uwindsor.ca/ hlynka/queue.html.
30. John N. Daigle," Queueing Theory with Application to Packet Telecommunications," Springer, 2004.
31. N. U. Prabhu," Foundations of Queueing Theory," Springer, 1997.

NADER MEHRAVARI
Lockheed Martin, Owego, NY