D.P. Heyman, M.J. Sobel, Eds., *Handbooks in OR & MS, Vol. 2*
© Elsevier Science Publishers B.V. (North-Holland) 1990

Chapter 10

Queueing Theory

Robert B. Cooper

Department of Computer Science, Florida Atlantic University, P.O. Box 3091, Boca Raton, FL 33431-0991, U.S.A.

1. Introduction

Queueing theory concerns the construction and analysis of mathematical models of systems that provide service to customers whose arrival times and service requirements are random. The basic queueing model, from which more complicated models can be constructed, consists of three components: (1) the input process, (2) the service mechanism, and (3) the queue discipline.

The *input process* describes the statistical properties of the time instants (*epochs*) at which the customers arrive. Typically, this is expressed in terms of the distribution of the *interarrival times* (the time intervals separating successive arrival epochs), but other descriptions may be preferable, depending on context. Similarly, the *service mechanism* specifies the number of servers and describes the statistical properties of the *service times* (the lengths of time that the customers hold the servers). Usually, the service times are assumed to be independent of the arrival process and each other, and to be identically distributed, without regard for which server, if any, provides the service. And the *queue discipline* describes the behavior of the customers who are *blocked*, that is, who find all servers busy when they arrive. Two typical assumptions for the queue discipline are (1) all blocked customers leave immediately (*blocked customers cleared* or *lost*), or (2) all blocked customers wait in a queue (hence the name *Queueing Theory*) until they are served (*blocked customers delayed*); in the latter case it may be necessary to specify the order in which waiting customers are selected from the queue, such as first-in first-out (FIFO) or its reverse, last-in first-out (LIFO). Other, more complicated situations can be modeled, but the simple structure described above provides a sufficiently rich basis to construct a useful mathematical theory of the behavior of many service systems subject to random demands.

Queueing theory was originally developed to facilitate the analysis and design of telephone systems. (In telephony, the 'customers' might, for example, be the telephone calls, and the 'servers' the telephone trunks that carry them.) The first major results were obtained by A.K. Erlang, a scientist and mathematician with the Copenhagen Telephone Company who, in 1917, published the important formulas that today bear his name. (See Brockmeyer, Halstrøm and Jensen [1948], *The Life and Works of A.K. Erlang*.) With the codification of the subject of Operations Research in the early 1950s, Queueing Theory was adopted as one of its methodologies and its range of application enlarged, but its major applications remained primarily in telecommunications (under the names teletraffic theory, telephone traffic engineering, and the like) until the early 1970s. (Significantly, Syski's 1960 encyclopedic work, *Introduction to Congestion Theory in Telephone Systems*, has recently been reissued, with some revisions, Syski [1986]; Section 2.6 of Chapter 1 is a short historical survey of teletraffic theory, including many historically interesting references.) With the explosive growth of computer science and engineering, a major new area of application was born in the early 1960s, and new models and theory were developed to meet the needs of this new technology. (Here, the 'customers' might, for example, be transactions, and the 'server' a CPU.) And with the increasing convergence between telecommunications and computer technology in the 1980s, as well as other interesting developments such as flexible manufacturing systems, technology again is providing the models and applications for a new era of accelerating growth in queueing theory. (As evidence, 1986 saw the birth of a new international journal, *Queueing Systems: Theory and Applications*, edited by N.U. Prabhu, unique in its dedication solely to the 'study of queueing systems occurring in science and engineering'.)

In the 1950s, D.G. Kendall introduced the shorthand notation $a/b/c$ to describe a queueing model, where a refers to the input process, b refers to the service-time distribution, and c is the number of servers, and where it is assumed that the queue discipline is blocked customers delayed. This notation is convenient for describing certain standard models, but it becomes messy and ambiguous when it is adapted to include the many variations of the standard models. Therefore, we will use this notation to describe only these standard models, which we will specify when we come to them; otherwise, we shall describe the models verbally.

It may be surprising for the nonspecialist reader to learn that such elementary models would lead to, literally, thousands of papers and books. (According to an estimate reported in Disney and Kiessler [1987], there are at least 5000.) In fact, queueing theory has spawned subfields that are becoming disciplines in themselves. A good example is the subject of *Queueing Networks* (in which the classical models are interconnected in networks, with complicated routing patterns, including feedback), to which Chapter 11 in this Handbook has been dedicated. And, of course, queueing theory is both a consumer and producer of more fundamental mathematical theory, such as that of *Point Processes* (which can be used to represent arrival epochs and service times), and to which Chapter 1 of this Handbook has been dedicated.

In this chapter, we will restrict ourselves largely to what are now called the 'classical' queueing models, as exemplified by the models and notation of Kendall, and some of their variants. Our aim is to give a basic understanding of the models that have proven useful and, together with some of the other chapters in this Handbook, to provide the tools and insight for the continuing development of the theory and its application to increasingly complicated systems. Our method will be to survey the main concepts and results, omitting algebra and details of proofs; to direct the reader to more detailed sources; and to provide interpretations and commentary that unify and explain, where possible. The selection of topics and order of presentation follow, somewhat, Cooper [1981] (naturally); this chapter updates that text, and concludes with a list of English-language (and a few foreign-language) books on queueing theory published since 1981.

2. Some general theorems

In this section we will present some general theorems and concepts that we shall specialize and apply later when we discuss specific queueing models.

Let $N(t)$ be the number of customers present at time t, and define the *state probability* $P_j(t) = P\{N(t) = j\}$ $(j = 0, 1, 2, \ldots)$, the probability that the system is in state j at time t. Of particular interest in queueing theory and its applications is the *statistical-equilibrium* (or *steady-*) *state distribution* P_j $(j = 0, 1, 2, \ldots)$, defined by

$$P_j = \lim_{t\to\infty} P_j(t) . \tag{2.1}$$

P_j is interpreted as follows: Suppose the 'system' has been in operation for a time sufficiently long so that the initial conditions are irrelevant, that is, the system is in *statistical equilibrium*. Then P_j is the probability that, at any arbitrary point in time, the number of customers present is j; equivalently, P_j is the fraction of statistically identical systems that contain j customers at any arbitrary point in time. (The equivalence of these interpretations states that the system is *ergodic*.) The statistical-equilibrium distribution is *stationary*; that is, if the statistical-equilibrium distribution describes the system at time t_0, it will describe it at every epoch $t > t_0$, as long as there is no other information relating to the state of the system between t_0 and t. It is important to emphasize that the notion of statistical equilibrium relates to our *knowledge* of the system, not essentially to the length of time it has been in operation (because an observation made at any time, including an epoch at which the statistical-equilibrium probabilities apply, provides information equivalent to new initial conditions, and thus the equilibrium probabilities implied by stationarity are henceforth inapplicable until an (essentially) infinite additional time has elapsed).

For the most part, queueing theory is concerned with the calculation of the probabilities $\{P_j\}$ and their use in calculating performance measures such as probability of blocking, server utilization, and mean waiting time. Deeper mathematical questions, such as those pertaining to existence of limits and ergodicity of distributions, are more properly in the domain of the theory of stochastic processes. As a practical matter, theoretical questions of this type do not usually intrude on queueing-theory studies; except in unusual cases, a deep theoretical understanding of the theory of stochastic processes is not necessary for the understanding, development, or application of queueing theory. (I must report that this assertion generated some disagreement from readers of an early version of this chapter. Perhaps it depends on how deep is deep.) Furthermore, queueing theory tends to be 'error-detecting': results derived under false assumptions are often (but not always) obviously meaningless or incorrect.

The distribution $\{P_j\}$ is said to represent the viewpoint of the *outside observer*, because it describes the distribution of states that an outside observer would see if he were to observe a system over all time, or to sample many statistically identical systems at arbitrary points in time. The critical characteristic of the outside observer is that he only observes the system, but does not interact with it; and furthermore, his observation epochs are arbitrary, that is, ordinary, not special, in the statistical sense.

In contrast, consider the viewpoint of the *arriving customer*, who, by definition, observes the system at his arrival epoch, that is, when he is making a request for service. The arriving customer's observation epoch is clearly special, since any point at which the number of customers present increases is necessarily an arrival epoch (although an arrival that departs immediately because of blocking, say, will not directly affect the state of the system), and hence the arriving customer, in general, interacts with the system when he makes his 'observation'. If we let $C(t, t+h)$ be the event that a customer arrives in the interval $(t, t+h)$, then the probabilities (for $j = 0, 1, 2, \ldots$)

$$\Pi_j(t) = \lim_{h \to 0} P\{N(t) = j \,|\, C(t, t+h)\} \tag{2.2}$$

can be interpreted as the arriving customer's distribution. That is, $\Pi_j(t)$ is the conditional probability that the system is in state j at time t, *given* that a customer arrives just after time t; in other words, $\{\Pi_j(t)\}$ is the distribution that describes the state of the system as seen by the customers when they arrive.

It is a remarkable theorem that, if the customer arrival epochs follow a *Poisson process* (see Chapter 1), then

$$\Pi_j(t) = P_j(t)\,. \tag{2.3}$$

Note that (2.3) remains valid whether or not the system is in statistical equilibrium, and whether or not the arriving customer actually joins the system (and thereby causes a state transition) or merely observes the system and departs immediately (and hence does not cause a state transition).

To see the subtlety of this result, consider the following example: A single customer repeatedly requests service from a server, causing the system state to alternate between 'server idle' ($j = 0$) and 'server busy' ($j = 1$) according to some arbitrary stochastic process (in which the time from any service completion to the customer's next arrival epoch is assumed to be nonzero). Then, clearly, $\Pi_1(t) = 0$ because the server can never be occupied at (just prior to) an arrival epoch; on the other hand, $P_1(t)$ might have any value between 0 and 1. (In this example, the interarrival times have 'memory' and therefore the arrival epochs cannot constitute a Poisson process.)

Although we have defined the arriving customer's distribution $\{\Pi_j(t)\}$ for finite t, as a practical matter we will be concerned almost exclusively with the limiting (i.e., statistical-equilibrium) distribution,

$$\Pi_j = \lim_{t\to\infty} \Pi_j(t) . \tag{2.4}$$

Π_j is the probability that a customer, who arrives when the system is in statistical equilibrium, finds j other customers present; equivalently, Π_j is the fraction of arrivals (to a system in statistical equilibrium) who find the system in state j. Then, for systems with Poisson input in statistical equilibrium, (2.3) becomes

$$\Pi_j = P_j . \tag{2.5}$$

The Poisson process is characterized by the property of *memorylessness* (to be discussed shortly), which, in the absence of evidence to the contrary, makes it the least presumptive (and therefore, arguably, the most realistic) description of a customer arrival process. It is also the simplest mathematically, in the sense that it yields the most tractable models. (These facts are probably not unrelated.) The important theorem expressed by (2.5) is proved in Wolff [1982], who coined the acronym PASTA (*P*oisson *A*rrivals *S*ee *T*ime *A*verages) to describe this fundamental property. (See König and Schmidt [1980] and Niu [1984] for a discussion of models for which the stochastic inequality $\{\Pi_j\} \geqslant \{P_j\}$ obtains, and vice versa.) A discrete analogue of PASTA is given in Halfin [1983] and Whitt [1983], and extended in Makowski, Melamed, and Whitt [1989] (see also the observation following our equation (4.9)). Although the assumption of Poisson arrivals is sufficient for (2.5), it is not necessary; this and other aspects of (P)ASTA are discussed in recent papers by Brémaud [1989, 1990], König, Schmidt, and van Doorn [1989], Melamed and Whitt [1990a,b], Serfozo [1989a, b], and Stidham and El Taha [1989].

So far, we have considered two viewpoints, that of the outside observer and that of the arriving customer. Another natural viewpoint is that of the *departing customer*: Let Π_j^* be the statistical-equilibrium probability that just after a departure epoch there are j customers remaining in the system; equivalently, Π_j^* is the probability that a departing customer (from a system in statistical equilibrium) leaves behind him j other customers. Then, we have the following useful theorem:

$$\Pi_j = \Pi_j^* . \tag{2.6}$$

There are only two, rather weak, conditions required for (2.6): (1) the system is in statistical equilibrium, and (2) the system is *skip-free* (that is, state transitions occur only in single up-or-down steps or, in other words, arrivals and departures occur one-at-a-time).

This theorem becomes intuitively obvious when one observes that to every customer whose arrival causes the state transition $j \rightarrow j+1$, there corresponds another whose departure causes the transition $j+1 \rightarrow j$ (if the system is in statistical equilibrium). A formal statement and proof (by P.J. Burke) is given in Cooper [1981], pp. 185–188. A generalization to include the case when arrivals occur in batches of random size, or departures occur in batches of fixed size, is given in Hebuterne [1988] (see also Papaconstantinou and Bertsimas [1990]).

The usefulness of (2.5) and (2.6) follows from the fact that in many cases, as we shall see, adoption of one of the three viewpoints (outside observer, arriving customer, departing customer) greatly simplifies the calculations; which viewpoint to adopt depends on the specific model under consideration.

Another extremely useful result, remarkable in its generality, is *Little's theorem*, usually written

$$L = \lambda W . \tag{2.7}$$

Here L is the expected value ('time-average') of the number of customers present in the 'system' at an arbitrary point in time, W is the expected value ('customer-average') of the time spent in the 'system' by a customer, and λ is the rate at which customers enter the 'system'; (2.7) holds for any 'system' in statistical equilibrium. We have put *system* in quotes to indicate that its definition is very broad, not necessarily restricted to queueing systems. Even in the context of queueing models, the definition is broad. For example, the 'system' might be the queue of customers waiting for service, excluding those being served; then L is the mean (expected value) queue length, W is the mean waiting time, and λ is the rate at which customers join the queue. Or the 'system' might be the servers alone; then L is the mean number of busy servers, W is the mean service time, and λ is the rate at which customers enter the servers. The only requirements for (2.7) to hold are a 'system' in statistical equilibrium for which the expected values L, λ, and W exist, and some technical assumptions. The usefulness of (2.7) follows from its great generality and the fact that, in many cases, it is much easier to calculate L (which is an expected value of a discrete random variable) than to calculate W (which is an expected value of a (usually) continuous random variable), or vice versa.

Although the notation of (2.7) is traditional, the use of the symbol λ to represent the *entry* rate is somewhat confusing, because the same symbol λ is traditionally used also to represent the *arrival* rate, which includes those customers who arrive (request service) but never get served (because of impatience, lack of waiting space, or any other reason), as well as those arrivals that do receive service. In what follows, we will reserve the symbol λ to represent the arrival rate.

As an example of the generality and utility of Little's theorem (2.7), consider any queueing system in statistical equilibrium; and let λ denote the arrival rate, let B denote the fraction of arrivals who don't get served, and let τ denote the mean service time for those customers who do get served. Then, if a' is defined to be the mean number of busy servers, it follows from (2.7) (more precisely, from '$H = \lambda G$', a generalization of (2.7); see Exercise 11-15 of Heyman and Sobel [1982]) that $a' = \lambda(1 - B)\tau$, or

$$a' = a(1 - B) , \tag{2.8}$$

where $a \equiv \lambda\tau$. Furthermore, in the case of a single-server system, then $a' = P_1 + P_2 + \cdots = 1 - P_0$, where P_j is the probability that j customers are present (where the first equality follows from the definition of a' for a single-server system, and the second equality follows from the normalization requirement); hence, from (2.8) we have

$$P_0 = 1 - a(1 - B) . \tag{2.9}$$

We shall return to these results later.

Theorem (2.7) was first proved by Little [1961]. Stidham [1974] gives an elementary and intuitively appealing proof; and Heyman and Sobel [1982] reproduce Stidham's proof and give a detailed discussion and some generalizations. Ramalhoto, Amaral, and Cochito [1983] give a comprehensive survey, including discussions of historical aspects, different proofs, and generalizations. In Glynn and Whitt [1986] the authors establish a central-limit-theorem version of (2.7), and in Glynn and Whitt [1989] they apply their results to investigate the asymptotic efficiency of different statistical estimators of L and W; also, they provide many references. Halfin and Whitt [1989] give an insightful and useful 'ordinal version' of (2.7), in which time is measured solely in terms of the number of arrivals that occur. In a wide-ranging paper on *sample-path analysis* (drawing probabilistic conclusions through analysis of individual realizations of a stochastic process) in queueing theory, Stidham [1981] discusses Little's theorem and its generalizations, relations between the stationary distribution of a process and an *imbedded* (defined in Section 7) process, the phenomenon of *insensitivity* (invariance of some characteristic with respect to the form of the distribution function of an underlying random variable; see, e.g., Schassberger [1986], Whittle [1986]), and *operational analysis* (a heuristic method of analysis popular in performance evaluation of computer systems, not discussed further here).

A final useful result is the relationship between the Laplace–Stieltjes transform ϕ of the distribution function of the length of an interval and the probability-generating function g of the number of Poisson arrivals that occur during the interval: If the length of the interval has distribution function F, with Laplace–Stieltjes transform ϕ,

$$\phi(s) = \int_{0-}^{\infty} \mathrm{e}^{-st}\, \mathrm{d}F(t) , \tag{2.10}$$

and if

$$P_j = \int_0^\infty \frac{(\lambda t)^j}{j!} \mathrm{e}^{-\lambda t} \, \mathrm{d}F(t) , \tag{2.11}$$

with probability-generating function g,

$$g(z) = \sum_{j=0}^{\infty} P_j z^j , \tag{2.12}$$

then substitution of (2.11) into (2.12) yields

$$g(z) = \phi(\lambda - \lambda z) . \tag{2.13}$$

Keilson and Servi [1988] enlarge on the 'distributional form of Little's law' expressed by (2.13) in the case when g and ϕ describe number in system and time in system, respectively.

3. Performance measures

For a given queueing model, a typical objective is to find a formula that expresses the relationship between the demand on a system and its performance. Demand is often expressed in terms of the *offered load* a, which is defined as the mean number of arrivals per unit time, with the time unit taken to be the mean service time. If we let λ denote the customer arrival rate and τ denote the mean service time, then (consistent with the definition following (2.8))

$$a = \lambda \tau . \tag{3.1}$$

The unit of this dimensionless quantity is the *erlang*. In a system with an infinite number of servers (in which every arrival immediately enters service, and hence $B = 0$), it follows from (2.8) that the offered load equals the mean number of customers simultaneously in service; that is, the offered load is the mean number of servers that the customer population 'wants' to be able to hold simultaneously. A given offered load a, which is a measure of what the customers want, will result in a *carried load* a', defined as the actual mean number of busy servers, which is a measure of what the system provides. Then the *throughput*, defined as the rate at which customers depart after having been served, is given by $a'/\tau = \lambda(1 - B)$, which is the rate at which customers are accepted for service.

In terms of the equilibrium state probabilities $\{P_j\}$, the carried load for an s-server system is given by

$$a' = \sum_{j=1}^{s-1} jP_j + s \sum_{j=s}^{\infty} P_j \,. \tag{3.2}$$

Note that $a' \leqslant s$. In the case of one server, $a' = 1 - P_0$: the carried load equals the proportion of time the server is busy. Clearly, when $s = \infty$, then $a = a'$. We define the *lost load* or *overflow* α as the difference between the offered load and the carried load:

$$\alpha = a - a' \,. \tag{3.3}$$

To interpret α, note that if the blocked customers are immediately routed to an *overflow group* consisting of an infinite number of servers with the same mean service time, then α equals the mean number of simultaneously busy servers in the overflow group. Thus, α equals the load that is lost to (overflows from) the primary group of s servers; if all blocked customers wait until they are served on the primary group, then $\alpha = 0$ and $a = a'$. (Systems with facilities to handle overflow traffic are of particular interest in telephony; a list of references is provided in Sudo [1987].)

The server *utilization* (or *occupancy*) ρ is defined as the load carried per server:

$$\rho = \frac{a'}{s} \,. \tag{3.4}$$

Since carried loads are mean values, we can interpret the load carried by a group of s servers as the sum of the loads carried by each of the s servers individually; and the load carried by each server (viewed in isolation as a one-server group) equals the proportion of time (or probability, from the outside observer's viewpoint) that the server is busy. Hence, the utilization ρ can be interpreted as the 'average' probability that a server is busy, or the 'average' load carried by a server. For a single-server system with Poisson input, the server utilization ρ equals (by PASTA) the probability that an arriving customer will find the server busy.

An essential characteristic of (real) queueing systems is that, for a given configuration, when utilization increases, so does the probability of blocking; that is, more efficient utilization of equipment is gained only at the cost of poorer service to the customers. Hence, one of the main virtues of queueing theory is that it permits one to examine the tradeoffs between the cost of providing service and the quality of the service provided. Along with blocking (resulting in loss or delay of a customer), other typical measures of performance (which also deteriorate as utilization increases) from the customer's viewpoint are his *waiting time* (time from arrival until commencement of service) and *response time* or *sojourn time* (waiting time plus service time).

4. Length-biased sampling and the role of the exponential distribution

The distribution most commonly used to describe the service times and the interarrival times is the exponential; that is, if X is a random variable that is exponentially distributed, with parameter $\mu > 0$, then its distribution function $F_X(t) \equiv P\{X \leq t\}$ is given by $F_X(t) = 0$ when $t < 0$, and when $t \geq 0$,

$$F_X(t) = 1 - e^{-\mu t} . \tag{4.1}$$

The parameter μ, called the *rate*, is the reciprocal of the mean value,

$$E(X) = \mu^{-1} ; \tag{4.2}$$

also, the variance is

$$V(X) = \mu^{-2} . \tag{4.3}$$

The essential property of an exponentially distributed random variable is the *Markov* or *memoryless* property, one expression of which is

$$P\{X > y + t | X > y\} = P\{X > t\} . \tag{4.4}$$

This identity, which is easily verified by calculating each side of (4.4) from (4.1), can be interpreted to say that knowledge of the present 'age' of an exponentially distributed random variable provides no information about its remaining lifetime. The Markov property, therefore, simplifies mathematical analysis, because it permits one to neglect information that otherwise would be relevant. And as a model of 'pure randomness', it seems to be the least presumptive in the absence of explicit information about how the past affects the future; furthermore, it has been found to be consistent with observation in many cases. Therefore, the assumption (4.1) is ubiquitous in queueing theory, sometimes for the sake of mathematical convenience, sometimes because it is the best model of randomness.

In particular, if interarrival times are independent, identically distributed exponential random variables, then the number of arrivals occurring in any fixed interval has the Poisson distribution, and conversely; such an input process is called a' *Poisson process*. Thus, queueing models with Poisson input are both less presumptive (in a sense) and more tractable than their brethren; this is a consequence of the Markov property and its corollary, PASTA.

The use of the exponential distribution to describe service times is more difficult to justify than its use to describe interarrival times. Whereas the memoryless property seems reasonable for interarrival intervals that are generated by a large population of customers who behave independently of each other (see the *superposition theorem* in Chapter 1), it seems counterintuitive that a service time of a single customer would have the (memoryless) property that the future service requirement is independent of the amount of service received so far. Nevertheless, data often show that the assumption of exponential service times is sufficiently accurate for many applications and, furthermore, it turns out in some significant cases that the performance measures are insensitive, or nearly so, to the form of the service-time distribution function. We will discuss this phenomenon as it arises in specific cases.

Whenever a customer arrives in a single-server queueing system, either he finds the server idle, in which case he enters service immediately, or he finds the server busy, in which case his arrival epoch has 'selected' a service time (the one in progress at his arrival epoch) from the service-time population. It is important to realize that, in general, a service time so selected is statistically 'longer' than an arbitrary service time (that is, one selected according to an unbiased sampling procedure) because, all other things being equal, an arrival epoch is more likely to occur during a long service time than during a short one.

This phenomenon of length-biased sampling gives rise to the well known *waiting-time paradox*: How long does a blocked arrival have to wait until the completion of the (exponential) service time of the customer in service at his arrival epoch? If the mean service time is τ, then 'common sense' leads most people to guess $\frac{1}{2}\tau$. But the Markov property implies that whenever an exponential variable is interrupted, its remaining duration is statistically the same as it was when it began, namely, exponentially distributed with mean τ. The latter answer is correct; the paradox is resolved by the realization that the selection process is length-biasing and the bias is, in the case of the exponential variable, exactly a factor of two.

In fact, this phenomenon is general: If we let X denote the length of an arbitrary (i.e., ordinary, typical) interval and let I denote the length of an interval 'selected' by an arbitrary Poisson arrival (the *test customer*), with distribution functions F_X and F_I, respectively, then

$$\mathrm{d}F_I(t) = \frac{1}{E(X)}\, t\, \mathrm{d}F_X(t)\,. \tag{4.5}$$

(According to (4.5), the frequency $\mathrm{d}F_I(t)$ with which the length-biased intervals have length (duration) t is proportional (by $1/E(X)$) to the product of the length t and the frequency $\mathrm{d}F_X\,(t)$ at which intervals of length t appear in the general population.) Similarly, the *forward recurrence time* R, defined as the elapsed time from the sampling epoch until the end of the selected interval, is described by

$$F_R(t) = \frac{1}{E(X)} \int_0^t [1 - F_X(y)]\, \mathrm{d}y \tag{4.6}$$

and

$$E(R) = \frac{E(X)}{2}\left[1 + \frac{V(X)}{E^2(X)}\right] = \frac{E(X^2)}{2E(X)} \tag{4.7}$$

for *any* random variable X. (In particular, if X is exponential, then $F_R(t) = F_X(t)$ and $E(R) = E(X)$, as is easily verified.)

Also, if ϕ_R is the Laplace–Stieltjes transform of F_R, and ϕ_X is the transform of F_X, then it follows from (4.6) that

$$\phi_R(s) = \frac{1}{E(X)} \frac{1 - \phi_X(s)}{s} , \tag{4.8}$$

a result that will prove useful later.

These topics are usually discussed in the context of *renewal theory* (see Chapter 1). As we shall see, relations (4.5) and (4.6) often are imbedded in queueing models in ways that are not obvious to the naive observer. Equation (4.6), in particular, appears in many guises. A discrete analogue of (4.5) also is important in queueing theory: Suppose that customers are grouped into *batches*, and consider the batch that contains a typical customer, selected 'at random' (the test customer). If M_I denotes the total number of customers in the test customer's batch (including himself), then (Burke [1975])

$$P\{M_I = j\} = \frac{1}{E(M_X)} jP\{M_X = j\} , \tag{4.9}$$

where M_X is the number who belong to an arbitrary batch. An interesting observation: It follows easily from (4.9) that if (and only if) the batch size has a Poisson distribution, then the number of other customers in the test customer's batch has the same Poisson distribution (Cooper [1981], Ex. 20, p. 59); similarly, if (and only if) the batch size has a geometric distribution, then (assuming that the test customer is equally likely to be in each of the positions in the batch) the number of customers in front of him has the same geometric distribution (Halfin [1983], Whitt [1983]).

5. One-dimensional birth-and-death models

Let $\{N(t); t \geq 0\}$ represent the size of a 'population' at time t, and assume that 'births' and 'deaths' occur according to the *transition probabilities*

$$P\{N(t+h) = j+1 | N(t) = j\} = \lambda_j h + o(h) \quad (j = 0, 1, 2, \ldots) , \tag{5.1}$$

$$P\{N(t+h) = j-1 | N(t) = j\} = \mu_j h + o(h) \quad (j = 1, 2, \ldots) , \tag{5.2}$$

and

$$P\{N(t+h) = k | N(t) = j\} = o(h) \quad (|j - k| \geq 2) , \tag{5.3}$$

where $o(h)$ is any function with the property that $o(h)/h \to 0$ as $h \to 0$. It is easy to derive from these postulates the set of differential equations (for $j = 0, 1, 2, \ldots$, with all undefined terms taken to be zero)

$$\frac{\mathrm{d}}{\mathrm{d}t} P_j(t) = \lambda_{j-1} P_{j-1}(t) + \mu_{j+1} P_{j+1}(t) - (\lambda_j + \mu_j) P_j(t) . \tag{5.4}$$

The stochastic process $\{N(t); t \geqslant 0\}$, called a *birth-and-death process*, can be used to construct queueing models, in which the customers (waiting or in service) correspond to the 'population', arrivals are 'births', and departures are 'deaths'. Various queueing models can be obtained by judicious choice of the *birth coefficients* λ_j and the *death coefficients* μ_j; then, for each $t > 0$ the state probabilities $P_j(t)$ can, in principle, be determined, subject to specification of the initial conditions $P_j(0)$ $(j = 0, 1, 2, \ldots)$. Unfortunately, this *time-dependent* (or *transient*) solution (i.e., for finite t) is ordinarily very difficult to obtain in closed form (a brief discussion is given in Section 9); however, numerical solutions are not nearly as difficult to obtain (see Chapter 5).

Fortunately, one need not solve (5.4) in order to obtain the equilibrium probabilities (2.1). It can be shown that if the limits defined by (2.1) exist, they can be calculated by taking limits directly in (5.4); the derivative on the left-hand side goes to zero (as one would expect) and the resulting difference equations reduce to

$$\lambda_j P_j = \mu_{j+1} P_{j+1} \quad (j = 0, 1, 2, \ldots) \tag{5.5}$$

which, together with the normalization condition,

$$\sum_{j=0}^{\infty} P_j = 1 , \tag{5.6}$$

can easily be solved successively in any particular case to yield the statistical-equilibrium state distribution (as seen by the outside observer). This solution is

$$P_j = \frac{\lambda_0 \lambda_1 \cdots \lambda_{j-1}}{\mu_1 \mu_2 \cdots \mu_j} P_0 \quad (j = 1, 2, \ldots) \tag{5.7}$$

and

$$P_0 = \left(1 + \sum_{k=1}^{\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{k-1}}{\mu_1 \mu_2 \cdots \mu_k}\right)^{-1} . \tag{5.8}$$

Mathematical questions concerning existence and uniqueness of solutions are standard fare in textbooks on stochastic process; see, e.g., Chapter 2 of this Handbook.

It can be shown that for any skip-free queueing model (customers arrive and depart one-at-a-time), the birth-and-death postulates (5.1)–(5.3) will be valid if the length of time from any epoch t until the next event (arrival or departure) depends only on the population size at time t, and not on its history prior to time t (i.e., $\{N(t); t>0\}$ is a *Markov process*—see Chapter 2). In particular, queueing models with Poisson or quasirandom (to be defined shortly) input and exponential service times are described by (5.5); however, in some important cases (to be discussed shortly) the equilibrium state probabilities satisfy (5.5) for *any* service-time distribution (the phenomenon of *insensitivity*).

The recurrence (5.5) has the important interpretation: "rate up = rate down". That is, the long-run rate, in transitions per unit time, at which the system moves up from state j to state $j+1$ equals the rate at which the system moves down from state $j+1$ to state j. Thus, the recurrence relations (5.5) are *conservation-of-flow* equations; in Section 6 we will generalize them to describe multidimensional models, with the interpretation "rate out = rate in". We note in passing that these results can be derived from *system point theory*, originated by Brill [1975] (see also Brill and Posner [1977, 1981] and Cohen [1977]). Shanthikumar and Chandra [1982] have specialized the theory to discrete state-space models, giving some interesting applications (e.g., see Shanthikumar [1987]).

The probabilities $\{P_j\}$ given by (5.7) and (5.8) represent the viewpoint of the outside observer. To obtain the arriving customer's distribution $\{\Pi_j\}$, we apply the definition (2.2), with the right-hand side in a form usually referred to as Bayes' formula:

$$\Pi_j = \lim_{h\to 0} \frac{P_j P\{C(t, t+h)|N(t)=j\}}{\sum_{k=0}^{\infty} P_k P\{C(t, t+h)|N(t)=k\}}\,, \tag{5.9}$$

where t in (5.9) is an arbitrary epoch for a system in statistical equilibrium. Note that (5.9) describes the distribution of states as seen by an arbitrary arriving customer, whether he joins the system or departs immediately without causing a state transition. In particular, if the arrivals follow a Poisson process with rate λ, then

$$P\{C(t, t+h)|N(t)=j\} = \lambda h + o(h) \quad (j=0,1,2,\ldots) \tag{5.10}$$

and (5.9) reduces to (2.5). Next, we consider some important examples of birth-and-death models.

Erlang loss system

In the Erlang loss model, we assume that (i) customers arrive according to a Poisson process (with rate λ), and (ii) customers who arrive when all s servers are busy leave the system immediately and have no effect on it (i.e., *blocked customers cleared*). Then the birth coefficients (5.1) are

$$\lambda_j = \begin{cases} \lambda & (j=0,1,2,\ldots,s-1)\,, \\ 0 & (j=s)\,. \end{cases} \tag{5.11}$$

The form of the birth coefficients $\{\lambda_j\}$ reflects, through the statement $\lambda_s = 0$, the assumption that blocked customers are cleared: although arrivals can occur when all s servers are busy, such an arrival does not cause a state transition. If the service times were assumed to be exponentially distributed (with rate μ), then the aggregate service completion rate (5.2) for the system when in state j would be

$$\mu_j = j\mu \quad (j = 1, 2, \ldots, s)\,, \tag{5.12}$$

where $\mu^{-1} = \tau$, the mean service time. Then the equilibrium state probabilities, given by (5.7) and (5.8), are

$$P_j = \frac{a^j/j!}{\sum_{k=0}^{s} a^k/k!} \quad (j = 0, 1, 2, \ldots, s) \tag{5.13}$$

and $P_j = 0$ $(j > s)$, where $a = \lambda/\mu$ is the offered load. In particular, when $j = s$ in (5.13) we have the *Erlang loss formula*,

$$B(s, a) = \frac{a^s/s!}{\sum_{k=0}^{s} a^k/k!}\,. \tag{5.14}$$

Formula (5.14) is also called the *Erlang B formula* and *Erlang's first formula*, the latter denoted by $E_{1,s}(a)$; it can be interpreted (by virtue of PASTA (2.5)) as both the fraction of time all servers are busy and the fraction of arrivals who find all servers busy (and thus are *lost*). The distribution (5.13) is called the *Erlang loss* distribution or the *truncated Poisson* distribution (if the range of the index k were extended to $s = \infty$, the denominator in (5.13) would be e^a, yielding the Poisson distribution).

It is most important to note that (5.13) (and hence (5.14)) remains valid even when the service times are not exponentially distributed, a fact conjectured by Erlang himself (a proof is outlined in Section 6); this is an example of the phenomenon of insensitivity. This insensitivity of the Erlang loss distribution (5.13) to the form of the service-time distribution has been proved and generalized by a succession of authors (see, e.g., Takács [1969] and Section 10 of Disney and König [1985]).

It can be shown (using insensitivity and PASTA) that (5.13) remains valid when there are $n \geqslant 2$ independent Poisson streams of arrivals, with different arrival rates λ_i and different mean service times τ_i; that is, if $a_i = \lambda_i\tau_i$, then (5.13) applies with $a = a_1 + \cdots + a_n$. In particular, each stream experiences the same probability of blocking, given by (5.14) with $a = a_1 + \cdots + a_n$.

Substitution of (5.13) into (3.2) gives (see also (2.8)) for the carried load a',

$$a' = a(1 - B(s, a)) \tag{5.15}$$

(which can be interpreted to say that the carried load is the portion of the offered load that is not lost), and hence the overflow α is given by

$$\alpha = aB(s, a)\,. \tag{5.16}$$

If the traffic is distributed equally among the s servers, then each server carries $\rho = a'/s$ erlangs. An interesting and important case is that of *ordered hunt* or *ordered entry*: the servers are numbered $1, 2, \ldots$ and each arrival is carried by the lowest-numbered idle server. Let $\tilde{p}_j$ denote the probability that server j is busy, i.e., the utilization of (or load carried by) the jth server; then, for $a > 0$ and $B(0, a) = 1$,

$$\tilde{p}_j = a(B(j-1, a) - B(j, a)) \quad (j = 1, 2, \ldots)\,. \tag{5.17}$$

This can be interpreted to say that the load carried by the jth server is the difference between the overflow from server $j-1$ and the overflow from server j. Formula (5.17) is useful in calculating economic tradeoffs between flat-rate and measured-rate trunks in telecommunications systems. For a generalization to the case of *heterogeneous* servers (server j has mean service time τ_j) see Cooper [1976, 1987]. Direct numerical calculation of formulas (5.14)–(5.17) is difficult for large values of a and s. A fast and accurate computational scheme, easy to program, is based on the recursion

$$B(n, a) = \frac{aB(n-1, a)}{n + aB(n-1, a)} \quad (n = 1, 2, \ldots; B(0, a) = 1)\,. \tag{5.18}$$

Graphs of $B(s, a)$ as a function of a, for different values of the parameter s, are given on pp. 316, 318 of Cooper [1981]. Mathematical properties of the Erlang loss function $B(z, \alpha)$, with z and α complex, are developed in Jagerman [1974].

Erlang delay system

In the Erlang delay model, we assume that (i) customers arrive according to a Poisson process (with rate λ), (ii) service times are independent, identical, exponential random variables (with rate μ), and (iii) customers who arrive when all s servers are busy join the queue and wait as long as necessary for service to begin (i.e., *blocked customers delayed*), and the queue discipline is *nonbiased* (i.e., the selection of a customer from the queue to begin service is made without regard to the customer's service time). In the notation of Kendall, this model is called M/M/s (*M*arkov (Poisson) input, *M*arkov (exponential) service times, s servers).

The birth coefficients in (5.7) and (5.8) are

$$\lambda_j = \lambda \quad (j = 0, 1, 2, \ldots)\,, \tag{5.19}$$

and the death coefficients are

$$\mu_j = \begin{cases} j\mu & (j = 0, 1, 2, \ldots, s)\,, \\ s\mu & (j = s+1, s+2, \ldots)\,, \end{cases} \tag{5.20}$$

where $\mu^{-1} = \tau$, the mean service time. Substitution of (5.19) and (5.20) into (5.7) gives

$$P_j = \frac{a^j}{j!} P_0 \quad (j = 1, 2, \ldots, s-1) \tag{5.21}$$

and

$$P_j = \frac{a^j}{s!s^{j-s}} P_0 \quad (j = s, s+1, \ldots)\,, \tag{5.22}$$

where $a = \lambda/\mu$ is the offered load. If $a < s$ the infinite sum in (5.8) converges, and

$$P_0 = \left(\sum_{k=0}^{s-1} \frac{a^k}{k!} + \frac{a^s}{s!(1 - a/s)} \right)^{-1} . \tag{5.23}$$

If $a \geq s$, then the infinite sum diverges to infinity, and we can take $P_0 = 0$, which implies, from (5.7), that $P_j = 0$ for all finite j. When all s servers are busy, the service completion rate is $s\mu$. Thus, we can interpret the condition $a < s$ (i.e., $\lambda < s\mu$) to say that a proper state distribution will exist only if the arrival rate is less than the maximum service-completion rate; otherwise, the queue length will grow to infinity. (Of course, this is only a mathematical idealization; no real queue can be infinite.) Note that the queue length will grow to infinity not only when $a > s$, but even when $a = s$; this means that the potential work-time lost when the server is idle cannot be recovered later.

In particular, let $C(s, a)$ denote the probability that all s servers are occupied:

$$C(s, a) = \sum_{j=s}^{\infty} P_j\,. \tag{5.24}$$

Then, we have the *Erlang delay formula*,

$$C(s, a) = \frac{\dfrac{a^s}{s!(1 - a/s)}}{\displaystyle\sum_{k=0}^{s-1} \frac{a^k}{k!} + \frac{a^s}{s!(1 - a/s)}} \quad (a < s)\,. \tag{5.25}$$

Formula (5.25) is also called the *Erlang C formula* and *Erlang's second formula*, the latter denoted by $E_{2,s}(a)$. It can be interpreted (by virtue of PASTA (2.5)) as both the fraction of time all servers are busy and the fraction of arriving customers who find all servers busy (and thus are *delayed*, i.e., must wait in the queue).

In particular, when $s = 1$ (i.e., for M/M/1) the distribution defined by (5.22) and (5.23) becomes (for $a < 1$) the *geometric*,

$$P_j = (1 - a)a^j \quad (j = 0, 1, 2, \ldots)\,; \tag{5.26}$$

also,

$$C(1, a) = a\,. \tag{5.27}$$

(Observe that (5.27) follows also from (2.9), because here $B = 0$.)

Note that, unlike the Erlang B distribution, the Erlang C distribution is valid only for exponential service times. As a practical matter, the state probabilities become less sensitive to the form of the service-time distribution as the number of servers increases or the offered load decreases, all other things being equal. This is because as the probability of queueing decreases, the system looks more like an (insensitive) Erlang loss system, in which queueing never occurs.

Substitution of (5.21) and (5.22) into (3.2) gives

$$a' = \begin{cases} a & (a < s)\,, \\ s & (a \geqslant s)\,. \end{cases} \tag{5.28}$$

(See also (2.8).) Hence, the server utilization (see (3.4)) is

$$\rho = \begin{cases} a/s & (a < s)\,, \\ 1 & (a \geqslant s)\,. \end{cases} \tag{5.29}$$

In the case of ordered hunt, if we let p_j be the load carried by the jth server, then it can be shown (Cooper [1981], pp. 153–157) that

$$p_j = \tilde{p}_j(1 - \rho C(s, a)) + \rho C(s, a)\,, \tag{5.30}$$

where $\tilde{p}_j$ is the corresponding value for the Erlang B model, given by (5.17). Again, for a generalization to the case of heterogeneous servers, see Cooper [1976, 1987].

As with the Erlang B formula, direct numerical calculation of the Erlang C formula is difficult for large values of a and s. However, it is easy to verify that, for every integer $n > a$.

$$C(n, a) = \frac{nB(n, a)}{n - a(1 - B(n, a))}\,; \tag{5.31}$$

formula (5.31), in conjunction with the recurrence relation (5.18), provides a basis for a fast, accurate computational algorithm. Also, in light of (5.15), (5.31) shows easily that, for $a > 0$, $C(s, a) > B(s, a)$. Graphs of $C(s, a)$ as a function of a, for different values of the parameter s, are given on pp. 320, 322 of Cooper [1981].

Next, we consider the distribution of waiting times when the queue discipline is FIFO, i.e., service in order of arrival. Let W be the waiting time of an arbitrary customer. Then we can write

$$P\{W>t\}=P\{W>0\}P\{W>t|W>0\}\,. \tag{5.32}$$

Now, clearly, for any nonbiased queue discipline,

$$P\{W>0\}=C(s,a) \tag{5.33}$$

and, it can be shown, for FIFO,

$$P\{W>t|W>0\}=\mathrm{e}^{-(1-\rho)s\mu t} \tag{5.34}$$

Observe that, according to (5.34), the waiting-time distribution for those customers who wait is (again) exponential. Thus, the conditional mean waiting time (for those who are not served immediately) is

$$E(W|W>0)=\frac{1}{(1-\rho)s\mu} \tag{5.35}$$

and, therefore, the mean waiting time for all customers is

$$E(W)=\frac{C(s,a)}{(1-\rho)s\mu}\,; \tag{5.36}$$

also, by Little's theorem (2.7), if Q is the queue length, then $E(Q)=\lambda E(W)$. It is important to note that formulas (5.35) and (5.36) remain valid for *any* nonbiased queue discipline (an easy consequence of Little's theorem and the fact that queue lengths are stochastically invariant with respect to nonbiased queue disciplines).

Quasirandom input: Blocked customers cleared, delayed

With Poisson input, which is characterized by (5.10), the stream of arriving customers is external to the system in the sense that future arrival epochs are not affected by the current state of the system. In particular, the Poisson-input model ignores the fact that when the arrivals are generated by a finite number of potential customers, or *sources*, the instantaneous arrival rate might be affected by the number of sources that are ineligible to generate requests for service because they are currently waiting in the queue or being served. *Quasirandom input* accounts for this 'finite-source effect' by assuming a finite customer population of size n, in which each customer generates requests with rate γ when idle and rate 0 when waiting or in service; then, in contrast with (5.10),

$$P\{C(t, t+h)|N(t)=j\} = (n-j)\gamma h + \mathrm{o}(h) \quad (j=0,1,2,\ldots,n). \tag{5.37}$$

One consequence of quasirandom input is that, in contrast with Poisson input, the arrival rate λ now must be *calculated* from the state probabilities. When (5.37) is inserted into (5.9), we have

$$\Pi_j[n] = \frac{(n-j)P_j[n]}{\sum_k (n-k)P_k[n]}, \tag{5.38}$$

where we have written $P_j = P_j[n]$ and $\Pi_j = \Pi_j[n]$ to emphasize the dependence on the number n of sources. After inserting (5.7) into (5.38) and simplifying, we get the remarkable result:

$$\Pi_j[n] = P_j[n-1]. \tag{5.39}$$

(For a more general statement and a more rigorous proof (by P.J. Burke), see Exercise 1, p. 188 of Cooper [1981].) Equation (5.39) can be interpreted to say that, for systems with quasirandom input, the arriving customer's viewpoint is the same as that of the outside observer of the corresponding system with one less source; that is, the arriving customer sees what he would see if at his arrival epochs he were only to observe the system (containing only the *other* customers), but never join it. This theorem has a long history; according to Wilkinson [1955], it was used as early as 1907, in a memorandum by E.C. Molina of the American Telephone and Telegraph Company. Recently, Kelly [1979, Theorem 3.12], Lavenberg and Reiser [1980], and Sevcik and Mitrani [1981] have generalized it (as the *arrival theorem*) in the context of queueing networks (see also Melamed [1982], Disney and König [1985], Disney and Kiessler [1987], Melamed and Whitt [1990a], and Chapter 11).

When blocked customers are cleared and $n > s$, we retain (5.12) and replace (5.11) with

$$\lambda_j = \begin{cases} (n-j)\gamma & (j=0,1,2,\ldots,s-1), \\ 0 & (j=s). \end{cases} \tag{5.40}$$

Then (5.7) and (5.8) yield

$$P_j[n] = \binom{n}{j}\hat{a}^j \bigg/ \sum_{k=0}^{s} \binom{n}{k}\hat{a}^k \quad (j=0,1,2,\ldots,s) \tag{5.41}$$

and, from (5.39),

$$\Pi_j[n] = \binom{n-1}{j}\hat{a}^j \Big/ \sum_{k=0}^{s} \binom{n-1}{k}\hat{a}^k \quad (j = 0, 1, 2, \ldots, s), \tag{5.42}$$

where $\hat{a} = \gamma/\mu$ is the load offered *per idle source*. Formula (5.42) with $j = s$, which is the analogue of the Erlang B formula, is often called the *Engset formula* (after Engset [1918], who derived (5.44), below).

Observe that if we make the substitution

$$\hat{a} = \frac{p}{1-p} \tag{5.43}$$

in (5.41), we obtain the *truncated binomial* distribution:

$$P_j[n] = \frac{\binom{n}{j} p^j (1-p)^{n-j}}{\sum_{k=0}^{s} \binom{n}{k} p^k (1-p)^{n-k}} \quad (j = 0, 1, 2, \ldots, s). \tag{5.44}$$

From (5.43), we have

$$p = \frac{\hat{a}}{1+\hat{a}}, \tag{5.45}$$

which equals the probability that an arbitrary source would be busy if $n \leqslant s$, in which case there would be no interaction among the sources; this provides a remarkable interpretation of (5.44): The loss system with quasirandom input behaves as if the sources become busy and idle independently of each other, when, in fact, they do interact. From (5.44), $a' = np$ when $s = n$. Hence, np can be interpreted as the load that would be offered if there were enough servers so that blocking never occurs; that is, np is the load the sources 'want' to offer, which we call the *intended offered load* a^*. From (5.45),

$$a^* = n\,\frac{\hat{a}}{1+\hat{a}}. \tag{5.46}$$

When blocked customers are cleared (and $n > s$), then $a > a^*$; the actual (measured) offered load increases as the number of servers decreases (all other things being equal), because customers who are blocked are returned immediately to the calling population. (When $s = 0$, then $a = n\hat{a}$, because all sources are always idle.) This fact embodies the essential reason why quasirandom-input models are more complicated than their Poisson-input counterparts.

Finally, we observe that, like its Poisson-input counterpart, the quasirandom-input loss system is insensitive to the form of the service-time distribution (for proofs and generalizations see Kosten [1949], Cohen [1957], and König [1965]).

The model with quasirandom-input, exponential service times, and blocked customers delayed (sometimes called the *machine interference* or *repairman* model) is the finite-source analogue of the Erlang C model. The state probabilities are given by (5.7) and (5.8), with

$$\lambda_j = (n-j)\gamma \quad (j = 0, 1, 2, \ldots, n) \tag{5.47}$$

and

$$\mu_j = \begin{cases} j\mu & (j = 1, 2, \ldots, s), \\ s\mu & (j = s+1, s+2, \ldots, n). \end{cases} \tag{5.48}$$

Unfortunately, unlike its Poisson-input counterpart, this model does not yield simple expressions for its performance measures or the state probabilities, which are most easily calculated numerically directly from the recurrence relations (5.5). However, this model is theoretically more elementary than its Poisson-input counterpart in that it raises no questions about convergence of infinite series; it is, in effect, self-regulating, shutting off the arrival stream when the queue length reaches $n-s$. Note that $a < a^*$ (when $n > s$), the opposite effect from that observed when blocked customers are cleared, and explained similarly. (When $s = 0$, then $a = 0$, because all sources are always waiting in the queue.) Numerical results for this model are given in Descloux [1962].

A useful relationship, which follows easily from Little's theorem, is

$$n = \lambda(\gamma^{-1} + E(W) + \mu^{-1}), \tag{5.49}$$

but either the arrival rate λ (which, in this case, equals the throughput) or the mean waiting time $E(W)$ must be calculated from the state probabilities; (5.49) then determines the other.

In summary, the quasirandom-input models are a finite-calling-population generalization of their Poisson-input counterparts (which can be obtained from the former by taking appropriately the limit as the size of the calling population approaches infinity). Unfortunately, they are more difficult both in concept and calculation than their Poisson-input counterparts.

6. Multidimensional birth-and-death models

The multidimensional birth-and-death model is the generalization of the one-dimensional birth-and-death model to the case where more than one variable is required to describe the system. Instead of the statistical-equilibrium "rate up = rate down" equations (5.5), we now have conservation-of-flow equations that can be interpreted to say "rate out = rate in"; that is, for each 'state' (appropriately defined, possibly a *macrostate*, i.e., a collection of states), the rate at which the system leaves that state (because of arrivals or departures) is equated to the rate at which the system enters the state. This is best explained through an example.

We consider a model of a simple circuit-switched telecommunications network: Suppose city A is connected to city B by s_1 telecommunications channels (trunks), and city B is connected to city C by s_2 trunks. Suppose calls between A and B occur according to a Poisson process with rate λ_1, and each such call, if it finds an idle trunk among the s_1 that connect A and B, holds the trunk for an exponentially distributed time with mean value μ_1^{-1}. Similarly, calls between B and C arrive at rate λ_2 and have mean holding time μ_2^{-1}. And calls between A and C (which are routed through B) arrive at rate λ_3, have mean holding time μ_3^{-1}, and require simultaneously two trunks, one connecting A and B and one connecting B and C. Any arriving call that cannot commence immediately is cleared from the system.

Let $P(j_1, j_2, j_3)$ be the statistical-equilibrium probability that there are j_1 calls between A and B, j_2 calls between B and C, and j_3 calls between A and C. Then the "rate out = rate in" equations are, when $j_1 + j_3 < s_1$ and $j_2 + j_3 < s_2$,

$$\begin{aligned}(\lambda_1 + \lambda_2 + \lambda_3 + j_1\mu_1 + j_2\mu_2 + j_3\mu_3)&P(j_1, j_2, j_3)\\ = \lambda_1 P(j_1 - 1, j_2, j_3) &+ \lambda_2 P(j_1, j_2 - 1, j_3) + \lambda_3 P(j_1, j_2, j_3 - 1)\\ &+ (j_1 + 1)\mu_1 P(j_1 + 1, j_2, j_3) + (j_2 + 1)\mu_2 P(j_1, j_2 + 1, j_3)\\ &+ (j_3 + 1)\mu_3 P(j_1, j_2, j_3 + 1)\,. \end{aligned} \tag{6.1}$$

Now consider the boundary conditions $j_1 + j_3 = s_1$, $j_2 + j_3 < s_2$; then

$$\begin{aligned}(\lambda_2 + j_1\mu_1 + j_2\mu_2 + j_3\mu_3)&P(j_1, j_2, j_3)\\ = \lambda_1 P(j_1 - 1, j_2, j_3) &+ \lambda_2 P(j_1, j_2 - 1, j_3) + \lambda_3 P(j_1, j_2, j_3 - 1)\\ &+ (j_2 + 1)\mu_2 P(j_1, j_2 + 1, j_3)\,. \end{aligned} \tag{6.2}$$

Note that (6.2) can be obtained from (6.1) by deleting terms that correspond to transitions prohibited by the boundary conditions. Similar equations hold for the boundary conditions $j_1 + j_3 < s_1$, $j_2 + j_3 = s_2$, and the boundary conditions $j_1 + j_3 = s_1$, $j_2 + j_3 = s_2$.

Clearly, equations of this type cannot, in general, be solved by recurrence, in contrast with the one-dimensional flow equations (5.5). There are many solution strategies, including generating functions, numerical analysis (see Chapter 5), and ad hoc serendipity but, as it turns out, the method of separation of variables, leading to a *product-form* solution, works in a surprising number of cases. For example, it is easy to verify that the following product-form solution satisfies (6.1) and all the boundary equations:

$$P(j_1, j_2, j_3) = \frac{(\lambda_1/\mu_1)^{j_1}}{j_1!}\,\frac{(\lambda_2/\mu_2)^{j_2}}{j_2!}\,\frac{(\lambda_3/\mu_3)^{j_3}}{j_3!}\,c\,, \tag{6.3}$$

where c is the normalization constant. One can now calculate network performance measures, such as the blocking probabilities for the three types of traffic.

As a second example, we consider a model of a simple store-and-forward telecommunications network: Suppose that two sets of servers are arranged in tandem, so that the output (customers completing service) from the first set of servers is the input to the second set. Assume that the arrival process at the first stage of this tandem queueing system is Poisson with rate λ, the service times in the first stage are exponential with mean μ_1^{-1}, and the queue discipline is blocked customers delayed. The customers completing service in the first stage enter the second stage, where the service times are assumed to be exponential with mean μ_2^{-1} and (unrealistically for this application) independent of their values in the previous stage. Customers leaving the first stage who find all servers occupied in the second stage wait in a queue in the second stage until they are served. The number of servers in stage i is s_i.

Let $P(j_1, j_2)$ be the statistical-equilibrium probability that there are j_1 customers in stage 1 and j_2 customers in stage 2. To save rewriting the state equations for each set of boundary conditions, let

$$\mu_i(j) = \begin{cases} j\mu_i & (j = 0, 1, \ldots, s_i)\,, \\ s_i\mu_i & (j = s_i + 1, s_i + 2, \ldots) \end{cases} \quad (i = 1, 2,)\,. \tag{6.4}$$

Then the statistical-equilibrium state equations, obtained by equating the rate the system leaves each state to the rate it enters that state, are

$$\begin{aligned} &[\lambda + \mu_1(j_1) + \mu_2(j_2)]P(j_1, j_2) \\ &\quad = \lambda P(j_1 - 1, j_2) + \mu_1(j_1 + 1)P(j_1 + 1, j_2 - 1) \\ &\qquad + \mu_2(j_2 + 1)P(j_1, j_2 + 1)\,. \end{aligned} \tag{6.5}$$

The term $\mu_1(j_1 + 1)P(j_1 + 1, j_2 - 1)$ reflects the fact that a departure from stage 1 constitutes an arrival at stage 2.

Again, it is easy to verify that the following product-form solution satisfies (6.5):

$$P(j_1, j_2) = P_1(j_1)P_2(j_2) \tag{6.6}$$

where $P_i(j)$ is given by (5.21)–(5.23) with $a = \lambda/\mu_i$ and $s = s_i$ $(i = 1, 2)$. The product solution (6.6) shows that, remarkably, the number of customers in each stage is independent of the number in the other; and furthermore, the second stage has the same state distribution it would have if the first stage weren't there.

These remarkable results, which were first obtained by R.R.P. Jackson [1954, 1956], suggest that the output process might, in fact, be the same as the input process to the first stage, that is, a Poisson process. The truth of this conjecture was proved by Burke [1956]; *Burke's theorem* (the *output theorem*) states that the sequence of departures from an Erlang delay system in equilibrium follows a Poisson process and, further, (as a consequence of *reversibility*—see, e.g., Kelly [1979]) the state of this Erlang delay system at any arbitrary time t_0 is independent of the departure process previous to t_0. J.R. Jackson [1957, 1963] first considered queueing networks with feedback; incredibly, the introduction of feedback preserves the product form of the solution even though it destroys the 'Poisson-ness' of the internal flows (see, e.g., Burke [1972] and Disney and Kiessler [1987].) Today, networks of queues that yield to product-form solutions are called *Jackson networks* (presumably, J.R.) or, in the computer science literature, *BCMP networks* (after Baskett, Chandy, Muntz, and Palacios [1975]). Driven by these kinds of theoretical results and many important applications in computer science and industrial engineering, the subspecialty of queueing networks has generated a huge literature of its own—see, e.g., Disney [1985], Disney and König [1985], Kelly [1979], and Chapter 11.

These two examples (circuit-switched and store-and-forward telecommunications networks) were chosen to illustrate both the simplicity of the product-form solution and the applicability of queueing models to the analysis and design of telecommunications networks. Some recent papers relating to circuit-switched networks are Whitt [1985a], Kelly [1986, 1988], and Heyman [1987], who discuss exact solutions, computational issues, insensitivity (e.g., the assumption of exponential service times is not necessary for the validity of (6.3)), and approximations (such as the important *Erlang fixed point* or *reduced load* iteration procedure, for calculation of point-to-point blocking probabilities without completely neglecting the dependencies among the network links that form a communications path). Models for store-and-forward networks were studied early on by Kleinrock [1964] who, in order to obtain a product-form solution, made (as in our example) the reasonable (but false) assumption that a message that traverses several links has its service times chosen independently at each link. The mathematical difficulties of removing this independence assumption are explored by Boxma [1979].

As a third example, we consider the *method of phases* (or *stages*), which is an important procedure according to which a random variable with an arbitrary distribution is replaced by either a sum, or a mixture, or a combined sum and mixture of independent (but not necessarily identical) exponential random variables (each being a *phase* or *stage* of the lifetime of the original random variable). This technique allows transformation of the original model into a multidimensional birth-and-death model, thus making it more amenable to analysis. In what follows, we first give an example in which the method of phases is used for an *approximation*; we then use this example to indicate how the method of phases can be used as a *theoretical* tool, especially for the investigation of insensitivity.

Consider, for example, the s-server Erlang loss system; we assume that blocked customers are cleared, arrivals follow a Poisson process with rate λ, and service times are independent, identical random variables with an arbitrary distribution. In this example of the application of the method of phases, we assume that the service time X can be approximated by a sum of n independent, but not necessarily identical, exponential random variables,

$$X = X_1 + \cdots + X_n \; ; \tag{6.7}$$

that is, we imagine that the service time X is composed of n independent phases of service, the ith phase being exponentially distributed with distribution function $F_i(t) = P\{X_i \leq t\} = 1 - \mathrm{e}^{-\mu_i t}$. Then $E(X) = \Sigma_{i=1}^{n} \mu_i^{-1}$ and $V(X) = \Sigma_{i=1}^{n} \mu_i^{-2}$. Since it is true that $(\Sigma_{i=1}^{n} \mu_i^{-1})^2 > \Sigma_{i=1}^{n} (\mu_i^{-1})^2$, it follows that any service time described by a random variable X, where $E(X) > \sqrt{V(X)}$, can be approximated as a sum of independent, exponential phases, as in (6.7), with the given mean and variance. Furthermore, by judicious choice of the values n and μ_i $(i = 1, 2, \ldots, n)$, other moments might also be fitted to better approximate the given service-time distribution. Of course, the phases $X_1, \ldots, X_n$ do not necessarily correspond to any actual phases of service, but are only artifices introduced for the purpose of approximating the original process by a birth-and-death process.

Now suppose the service time X has greater variability than the exponential distribution prescribes; that is, assume $E(X) < \sqrt{V(X)}$. In this case, we can model the random variable X as a parallel arrangement of exponential phases; that is, the realization of X is obtained by choosing, with probability p_i, the realization of the exponential random variable X_i. Thus, the distribution function of X is $F_X(t) = \Sigma_{i=1}^{n} p_i F_i(t)$, where, as before, $F_i(t) = 1 - \mathrm{e}^{-\mu_i t}$; then

$$E(X) = \sum_{i=1}^{n} p_i \mu_i^{-1}$$

and

$$V(X) = 2 \sum_{i=1}^{n} p_i \mu_i^{-2} - \Big(\sum_{i=1}^{n} p_i \mu_i^{-1} \Big)^2 .$$

In this case, X is said to be a *mixture* of exponentials, and $F_X(t)$ is called the *hyperexponential* distribution function.

To continue with our example of the method of phases applied to the s-server Erlang loss system, suppose that a representation of the form (6.7) has been fitted to the original data or hypothesized service-time distributon. For ease of exposition let us assume $n = 2$. Now, if we let $P(j_1, j_2)$ be the equilibrium probability that simultaneously there are j_1 customers in phase 1 of service and j_2 customers in phase 2, the corresponding conservation-of-flow equations are exactly the same as (6.5) with $\mu_i(j_i) = j_i \mu_i$; hence the solution is

$$P(j_1, j_2) = \frac{(\lambda/\mu_1)^{j_1}}{j_1!} \, \frac{(\lambda/\mu_2)^{j_2}}{j_2!} \, c \quad (j_1 + j_2 \leq s) \, . \tag{6.8}$$

The distribution $\{P_j\}$ of the total number of customers present is given by

$$P_j = \sum_{j_1+j_2=j} P(j_1, j_2) \,; \tag{6.9}$$

insertion of (6.8) into (6.9) yields, with the help of the binomial theorem, the Erlang loss distribution (5.13), where $a = \lambda(\mu_1^{-1} + \mu_2^{-1})$.

This conclusion should not be surprising in light of the asserted insensitivity of the distribution $\{P_j\}$ for the Erlang loss system to the service-time distribution (with a given mean). Clearly, the restriction to $n = 2$ phases is irrelevant. This 'approximate' analysis indicates that the method of phases might be useful as a theoretical tool in the investigation of insensitivity (because it shows explicitly that the value of the parameter n is irrelevant).

Regarding the use of the method of phases as an approximation, Schassberger [1973, pp. 32–33] proves a theorem that states, roughly speaking, that any nonnegative random variable X can be represented as accurately as desired by a compound sum of independent, identical, exponential variables; this, together with certain continuity results for stochastic models (see, e.g., Whitt [1980] and the references therein), provides the theoretical justification for its use in approximations.

The method of phases was generalized with the introduction of the class of *PH-distributions* (*PH*ase-type distributions) by Neuts [1975] and Takahashi and Takami [1976]: A nonnegative random variable X has a PH-distribution if X can be viewed as the time until absorption in a Markov chain with a finite number m of transient states and a single absorbing state $m + 1$. PH-distributions enjoy certain closure properties, with the consequence that they can be used as the basis for computational (i.e., numerical) algorithms for a wide variety of queueing models. These *matrix-analytic methods* have been developed systematically by Neuts [1981, 1989] (see Neuts [1984] for a nontechnical review) and his coworkers (see, e.g., Ramaswami and Latouche [1989], Ramaswami and Lucantoni [to appear], Neuts [1988]); see also Chapter 5.

Botta, Harris, and Marchal [1987] examine a similar class of distribution functions (the *G*eneralized *H*yperexponential (GH) family), and discuss the relationships among these and other similar classes.

Finally, we remark that a particularly useful technique for the numerical analysis of multidimensional birth-and-death queueing models is the *Gauss-Seidel* iteration method and its variants—see Kaufman [1983] and Mitra and Tsoucas [1988] for a theoretical discussion and references, and Chapter 5 for a discussion in the context of general numerical methods.

7. The M/G/1 queue

The archetypical classical queueing model is, in Kendall's notation, the M/G/1 queue: *M*arkov (Poisson) input (with rate λ), *G*eneral service times (with distribution function H, with mean $\tau = \mu^{-1}$ and variance σ^2), and 1 server; blocked customers wait in the queue as long as necessary for service to commence and the queue discipline is nonbiased.

The difficulty of this model, relative to M/M/1, is that when the service times are not assumed to be exponential, the transition rates at any time t depend on the amount of service attained by the customer (if any) who is being served at time t; hence, the "rate up = rate down" equations (5.5) are no longer applicable. One of the most fruitful approaches to this problem (and many similar problems) is to 'imbed' a Markov chain (see Chapter 2) at the service-completion epochs (Kendall [1951, 1953]): Let N_k^* be the number of customers left behind by the kth departing customer; then, from the theorem of total probability, we can write

$$P\{N_{k+1}^* = j\} = \sum_{i=0}^{\infty} P\{N_{k+1}^* = j | N_k^* = i\} P\{N_k^* = i\} . \tag{7.1}$$

Now define

$$\Pi_j^* = \lim_{k\to\infty} P\{N_k^* = j\} \quad (j = 0, 1, 2, \ldots) . \tag{7.2}$$

It can be shown (see Cohen [1982]) that (7.2) defines a unique proper distribution, independent of the initial conditions, if and only if $a = \lambda\tau < 1$ (in which case $a = \rho$, the server utilization). (The interpretation of this result is the same as that given earlier for the Erlang delay model—see the discussion between (5.22) and (5.29)). Then, the equilibrium *imbedded-Markov-chain* equations, which can be obtained by taking limits formally in (7.1), are

$$\Pi_j^* = p_j \Pi_0^* + \sum_{i=1}^{j+1} p_{j-i+1} \Pi_i^* \quad (j = 0, 1, 2, \ldots) , \tag{7.3}$$

where

$$p_k = \int_0^{\infty} \frac{(\lambda t)^k}{k!} \mathrm{e}^{-\lambda t} \, \mathrm{d}H(t) . \tag{7.4}$$

(Note that p_k is the probability of k arrivals during a service time.) Then, from (7.3) and (7.4) the probability-generating function $g(z) = \Sigma \, \Pi_j^* z^j$ of the distribution (7.2), which satisfies the normalization condition $g(1) = 1$, is

$$g(z) = \frac{(z-1)\eta(\lambda - \lambda z)}{z - \eta(\lambda - \lambda z)} (1 - \rho) , \tag{7.5}$$

where η is the Laplace–Stieltjes transform of H. Note that from (2.5) and (2.6),

$$\Pi_j^* = \Pi_j = P_j \tag{7.6}$$

(but Π_j and P_j are more difficult to calculate directly). Differentiation of (7.5) (and two applications of L'Hospital's rule) yields

$$g'(1) = \rho + \frac{\rho^2}{2(1-\rho)}\left(1 + \frac{\sigma^2}{\tau^2}\right). \tag{7.7}$$

If N denotes the total number of customers in the system and Q the number in the queue, then $E(N) = g'(1)$ and $E(Q) = E(N) - \rho$; hence,

$$E(Q) = \frac{\rho^2}{2(1-\rho)}\left(1 + \frac{\sigma^2}{\tau^2}\right) \tag{7.8}$$

and, by Little's theorem, the waiting time W has mean value

$$E(W) = \frac{\rho\tau}{2(1-\rho)}\left(1 + \frac{\sigma^2}{\tau^2}\right). \tag{7.9}$$

These important formulas exhibit two characteristic terms: the term $1-\rho$ in the denominator, and the term $1 + \sigma^2/\tau^2$, which is a manifestation of the length-biased sampling by which a customer 'selects' the service time during which he arrives. (Note that (7.9), for example, can be written $E(W) = \rho(1-\rho)^{-1}E(R)$, where $E(R)$ is given by (4.7) with X being the service time.)

Comparison of (7.7) for M/M/1 ($\sigma^2 = \tau^2$) and M/D/1 (*D*eterministic, i.e., constant service times, $\sigma^2 = 0$) shows that the mean waiting time in the case of exponential service times is exactly twice that in the case of constant service times (all other things being equal); but, in both cases, the probability of blocking is the same:

$$P\{W > 0\} = 1 - \Pi_0 = 1 - \Pi_0^* = 1 - g(0) = \rho\ , \tag{7.10}$$

where the last equality in (7.10) follows from (7.5). (Note that $\Pi_0 = P_0$, by PASTA, and hence (7.10) is consistent with (2.9) (with $B = 0$).) Thus, remarkably, the probability of waiting is insensitive to the form of the service-time distribution, but the length of wait is not.

Now, suppose the queue discipline is FIFO, and let ω be the Laplace–Stieltjes transform of the waiting-time distribution function. If ϕ is the corresponding transform for the sojourn time (waiting time plus service time) then, since the customer's service time is independent of his waiting time,

$$\phi(s) = \omega(s)\eta(s)\ . \tag{7.11}$$

Furthermore, since every departing customer leaves behind him exactly those customers who arrived during his sojourn time, (2.13) applies, with F interpreted as the distribution function of sojourn times. Combination of (2.13) with (7.5) and (7.11) yields the celebrated *Pollaczek–Khintchine formula* (Pollaczek [1930], Khintchine [1932]):

$$\omega(s)=\frac{s(1-\rho)}{s-\lambda(1-\eta(s))}\,. \tag{7.12}$$

(Equation (7.9), which of course can be obtained from (7.12) according to $E(W)=-\omega'(0)$, is also often called the Pollaczek–Khintchine formula.)

To invert (formally) the Pollaczek–Khintchine transform, note that the right-hand side of (7.12) can be expanded in a geometric series, and thus

$$\omega(s)=\sum_{j=0}^{\infty}(1-\rho)\rho^{j}\left(\frac{1}{\tau}\,\frac{1-\eta(s)}{s}\right)^{j}. \tag{7.13}$$

Comparison of (7.13) with (4.8) shows that term-by-term inversion gives

$$P\{W\leqslant t\}=\sum_{j=0}^{\infty}(1-\rho)\rho^{j}\tilde{H}^{*j}(t)\,, \tag{7.14}$$

where $\tilde{H}$ is the distribution function of the forward recurrence time of a service time,

$$\tilde{H}(t)=\frac{1}{\tau}\int_{0}^{t}[1-H(y)]\,\mathrm{d}y\,, \tag{7.15}$$

(see (4.6)) and $\tilde{H}^{*j}$ is its j-fold self-convolution.

Formula (7.14), found by Beneš [1957], seemed quite mysterious until recently: It says that the waiting time in M/G/1-FIFO is a geometric convolution of partial service times. This is clearly true for M/M/1, where (i) the state distribution is geometric (see (5.26)) and (ii) $\tilde{H}=H$ (see the discussion following (4.6)) but, in general, neither (i) nor (ii) is true (but, of course, (7.14) is).

One explanation of this mystery lies in the realization that waiting times in M/G/1-FIFO are the same as *remaining work* (the time until the system would become empty if no new customers arrived) in M/G/1-LIFO *preemptive-resume* (a new arrival preempts the customer in service, pushing him back to the head of the queue, and service resumes from where it left off when a preempted customer reenters service); in the latter case, the state distribution is indeed given by (5.26) (and is thus insensitive to the form of H), and the remaining service time for a preempted customer is, not surprisingly, described by (7.15). (For an intuitive derivation and references, see Cooper and Niu [1986].) In some sense, then, one can say that the queue discipline LIFO preemptive-resume is simpler, more natural, than FIFO.

We have discussed these arguments for M/G/1 in some detail because variations of them are common in the analysis of more complicated M/G/1-type queueing models. We now summarize some results that relate to other aspects of the M/G/1 queue.

Consider the queue from the viewpoint of the arrivals: Let N be the number of customers present when a typical customer (the test customer) arrives, and let R be the remaining service time of the customer (if any) in service when the test customer arrives. Define $\Pi_j(x) = P\{R \leqslant x, N = j\}$, with probability-generating function $\Pi(z, x) = \Sigma_{j=1}^{\infty} \Pi_j(x) z^j$. Then, Wishart [1961] (see also Takács [1963]) has shown (after much calculation) that

$$\Pi(z, x) = \frac{(1-\rho)\lambda z(1-z)}{\eta(\lambda - \lambda z) - z} \int_0^{\infty} \mathrm{e}^{-\lambda(1-z)\xi}[H(\xi + x) - H(\xi)]\,\mathrm{d}\xi\,, \tag{7.16}$$

from which it follows easily that, as one might expect,

$$P\{R \leqslant x | N \geqslant 1\} = \tilde{H}(x)\,, \tag{7.17}$$

where $\tilde{H}$ is given by (7.15).

The *busy period* B is defined as the length of time from the instant a customer enters a previously empty system until the next instant at which the system is completely empty (i.e., the continuous busy-time of the server). If β denotes the Lapace–Stieltjes transform of the distribution function $B(t) = P\{B \leqslant t\}$, then it can be shown (Takács [1962]) that β is the solution with smallest absolute value of the functional equation

$$\beta(s) = \eta(s + \lambda - \lambda\beta(s))\,; \tag{7.18}$$

furthermore, when $\lambda\tau \geqslant 1$ the mean busy period b is infinite, and when $\lambda\tau < 1$ b can be calculated from (7.18) according to $b = -\beta'(0)$, which yields (with $\rho = \lambda\tau < 1$)

$$b = \frac{\tau}{1-\rho} \tag{7.19}$$

(which is insensitive to the distribution of service times). (Observe that, characteristically, (7.19) 'tells' you that b is finite only when $\lambda\tau < 1$, because otherwise the right-hand side is infinite or meaningless.)

Define a *j-busy period* B_j as a busy period that begins with j customers present ($B_1 = B$). A simple argument shows that

$$B_j(t) = B^{*j}(t)\,, \tag{7.20}$$

where $B_j(t) = P\{B_j \leqslant t\}$: For imagine that the server serves all the *descendants* of the first of the j original customers (those who arrive during the first service time, plus those who arrive during the service times of those who arrived during the first service time, and so on) before beginning to serve the second of the original j, and so on; that is, each of the original j customers generates, through his descendants, his own 1-busy period. Hence, the j-busy period is the sum of these j (independent) 1-busy periods.

Equations (7.18) and (7.20) turn out to be quite useful in the analysis of some of the more complicated variants of M/G/1, such as *vacation models* (to be discussed shortly) and *priority queues* (see Takagi [1987]), and queues served in cyclic order by a server that moves from queue to queue (*cyclic queues*, or *polling models*—see Takagi [1986, 1990]). Interestingly, the right-hand side of (7.19) also equals the mean sojourn time in the M/G/1 queue with *processor-sharing* (in which each of the n customers present receives $(1/n)$th of the server's work; this model, which, like M/G/1-LIFO preemptive-resume, has its state distribution given by (5.26), insensitive to the form of H, is surveyed in Yashkov [1987, 1989]).

Now let K_j be the number of customers served during a j-busy period. Then, the joint distribution of K_j and B_j is

$$P\{K_j = n, B_j \leqslant t\} = \frac{j}{n} \int_0^t \frac{(\lambda y)^{n-j}}{(n-j)!} \mathrm{e}^{-\lambda y} \, \mathrm{d}H^{*n}(y) \quad (n \geqslant j) . \qquad (7.21)$$

Equation (7.21) suggests that the factor j/n equals the probability that the $n-j$ arrivals occur in such a way that the n service times form a busy period (that is, the server remains continuously busy until the completion of the nth service time). This interpretation is, in fact, correct, and has led to the realization that many such results can be obtained by combinatorial methods (in particular, from generalizations of the classical *ballot theorem*), and that there are deeper reasons, not immediately obvious, why apparently different questions often have strikingly similar answers. These ideas have been developed systematically by Takács [1967]; for some new results of this type, see Niu and Cooper [1989].

Another example of a result that is composed of familiar-looking terms, begging for interpretation, is the distribution function of the waiting times in nonpreemptive LIFO M/G/1:

$$P\{W \leqslant t\} = 1 - \rho + \sum_{j=1}^{\infty} \frac{(\lambda t)^j}{j!} \mathrm{e}^{-\lambda t} \frac{1}{t} \int_0^t [1 - H^{*j}(y)] \, \mathrm{d}y . \qquad (7.22)$$

To illustrate how the 'pieces' of the classical M/G/1 model fit together in the analysis of a variant of M/G/1-FIFO, and to convey the flavor of the classical transform-type arguments, as examplified by Takács [1962], we will outline a derivation of (7.22).

Let T_c be the arrival epoch of an arbitrary customer (the test customer, whose viewpoint we adopt). Assume first that he arrives when the server is busy, and let T_1 be the next subsequent service-completion epoch. Then the joint probablity $\tilde{P}_j(x)$ that j other customers arrive in (T_c, T_1) and that $T_1 - T_c \leqslant x$ is given by

$$\tilde{P}_j(x) = \int_0^x \frac{(\lambda \xi)^j}{j!} \mathrm{e}^{-\lambda \xi} \, \mathrm{d}\tilde{H}(\xi) , \qquad (7.23)$$

where $\tilde{H}$ is, according to (7.17), given by (7.15). Because the queue discipline is nonpreemptive LIFO, the test customer's waiting time W is the j-busy period generated by the j customers who arrived in (T_c, T_1). Therefore,

$$P\{W \leqslant t \mid W > 0\} = \sum_{j=0}^{\infty} \int_0^t \tilde{P}_j(t-x)\,\mathrm{d}B_j(x)\,, \tag{7.24}$$

where, by (7.20), B_j is the j-fold self-convolution of the distribution function of the 1-busy period, whose Laplace–Stieltjes transform is defined by (7.18). Therefore, in light of (7.10), the unconditional distribution of W is

$$P\{W \leqslant t\} = 1 - \rho + \rho \sum_{j=0}^{\infty} \int_0^t \tilde{P}_j(t-x)\,\mathrm{d}B^{*j}(x)\,. \tag{7.25}$$

Inserting (7.23) into (7.25) and taking Laplace–Stieltjes transforms (and using the fact that, if β_j is the Laplace–Stieltjes transform of B_j, then, from (7.20), $\beta_j = \beta^j$), we get, after some calculation,

$$\omega(s) = 1 - \rho + \lambda\,\frac{1 - \eta(s + \lambda - \lambda\beta(s))}{s + \lambda(1 - \beta(s))}\,, \tag{7.26}$$

which, with the help of (7.18), reduces to

$$\omega(s) = 1 - \rho + \frac{\lambda(1 - \beta(s))}{s + \lambda(1 - \beta(s))}\,. \tag{7.27}$$

Finally, inversion of (7.27) yields (7.22). (See Riordan [1961], Wishart [1960], and, especially, Takács [1963].)

This example illustrates not only the characteristic interlocking of the pieces of the M/G/1 jigsaw puzzle, but also the virtuosity required to obtain and invert Laplace–Stieltjes transforms. These techniques, while powerful in the hands of those at home in the complex plane, are felt by some to hide understanding behind the 'Laplacian curtain' (Kendall). This has motivated some to search for probabilistic or combinatorial derivations, which are technically more elementary, but may be conceptually more complicated. In truth, each approach complements the other. We shall return to this point shortly, when we indicate how (7.22) can be obtained by a combinatorial argument coupled with the concept of *duality* (between the M/G/1 queue and its dual GI/M/1 (*G*eneral *I*ndependent interarrival times) obtained by interchanging the interarrival-time and service-time distributions).

An interesting and important variant of M/G/1 is the case when the number n of waiting positions is finite. Then Π_{n+1} is the probability that an arriving customer will find all n waiting positions occupied; that is, Π_{n+1} is the fraction of customers who overflow and are lost. As a practical matter, a system designer might want to find the smallest value of n to satisfy a given criterion for probability of loss. The analysis is complicated by the fact that although the departing customer's distribution $\{\Pi_0^*, \Pi_1^*, \ldots, \Pi_n^*\}$ can be determined (at least numerically) from (7.3), one cannot calculate Π_{n+1} directly from Π_{n+1}^* (because, clearly, $\Pi_{n+1}^* = 0$). Discussions are given in Cooper [1981] and Takagi [1985], who provide complementary references.

Another interesting variant is the M/G/1 queue with *batch arrivals*: Each arrival epoch now corresponds to the arrival of a batch of customers, where the batch sizes are independent, identically distributed random variables (with mean m and variance $\hat{\sigma}^2$) and customers are served one-at-a-time, with service-time distribution H as before, and $\lambda m \tau < 1$. Let W be the equilibrium waiting time of an arbitrary (test) customer. Then $W = W_1 + W_2$, where W_1 is the time from arrival to start of service of the test customer's batch, and W_2 is the remaining time until start of service of the test customer; hence $E(W) = E(W_1) + E(W_2)$, where

$$E(W_1) = \frac{\lambda m^2 \tau^2}{2(1 - \lambda m \tau)} \left(1 + \frac{m\sigma^2 + \tau^2 \hat{\sigma}^2}{m^2 \tau^2}\right) \tag{7.28}$$

and

$$E(W_2) = \frac{(m-1)\tau}{2} + \frac{\hat{\sigma}^2 \tau}{2m} . \tag{7.29}$$

Formula (7.28) follows directly from (7.9) if we view the batch as if it were a single customer whose service time is the *compound sum* of the service times of the customers in a batch (with mean $m\tau$ and variance $m\sigma^2 + \tau^2\hat{\sigma}^2$). Formula (7.29) follows from the recognition that the batch that contains the test customer is 'larger' than an arbitrary batch; this is the discrete analogue of the length-biasing effect alluded to in the discussion surrounding (4.9). Interestingly, several treatments of this model in the literature have overlooked this effect, with the consequence that the second term in (7.29) is omitted. (This is an example where the 'error-detecting' property of queueing theory failed.) A more detailed discussion is given in Burke [1975]; see also Halfin [1983] and Whitt [1983], who study further the relationship between batch delays and customer delays.

There are many more related topics and variants of the M/G/1 model; our selection is based on a combination of criteria including importance, simplicity, and illumination of concepts that are characteristic of queueing theory. We close this section with a short discussion of vacation models, which have received much attention of late, both because of their interesting theoretical properties and because of their use as components in the analysis of the more complicated cyclic-queueing (or polling) models, which have important applications in telecommunications and computer network design (see, e.g., Bertsekas and Gallager [1987]).

In a *vacation model* the server intermittently goes on a 'vacation', during which time it is not available to serve the main stream of customers. The vacations can model server breakdowns, scheduled maintenance, or time devoted to other work, such as serving customers of a different priority class.

The idea of allowing a removable server dates at least from Gaver [1962] and Keilson [1962], but the general vacation model was introduced in Skinner [1967] and Cooper [1970]. Their motivation in defining the general vacation model was to permit calculation of waiting times in an N-queue cyclic-queueing model in the case $N = 2$ (Skinner [1967]) and arbitrary N (Cooper and Murray [1969]), in which a set of queues is served in cyclic order by a single server that travels from queue to queue. Then, from the viewpoint of any particular queue, the server is on vacation when it is not serving this queue. In their cyclic-queueing model, Cooper and Murray [1969] considered two vacation disciplines, *exhaustive service* (the server takes a vacation whenever the particular queue is empty) and *gated service* (the server takes a vacation when it finishes serving the customers who were waiting in the particular queue at the end of the last vacation); Cooper [1970] then used the M/G/1 vacation model to obtain exact waiting-time information for the (more complicated) N-queue, cyclic-queueing model with these two vacation disciplines. In this application, it is natural to assume that if the server finds no customers waiting when it returns from a vacation, it immediately takes another vacation, which is now called the *multiple-vacation* model (as opposed to the *single-vacation* model, introduced by Levy and Yechiali [1975], in which the server remains available for service if there are no customers waiting at the end of its vacation).

The interesting fact about vacation models is that, in certain cases, the equilibrium distribution of the number of customers present *decomposes* into a convolution of distributions, one of which relates to the corresponding model without vacations, and the other to the vacations alone. For example, let N_1 and N_2 be, respectively, the number of customers present at (1) an arbitrary (or arrival or departure) point in time and (2) an arbitrary (or arrival) point in time when (given that) the server is on vacation. If ψ and χ are the probability-generating functions of the equilibrium distributions of N_1 and N_2, then, as shown by Fuhrmann and Cooper [1985b] under quite general conditions,

$$\psi(z) = \chi(z) g(z) , \tag{7.30}$$

where g is given by (7.5) and, for the multiple-vacation model,

$$\chi(z) = \zeta(z) \frac{1 - \alpha(z)}{\alpha'(1)(1 - z)} , \tag{7.31}$$

where ζ an α are, respectively, the probability-generating functions of the number of customers present at the start of a vacation and the (presumed independent) number of customers who arrive during the ensuing vacation. Note that the vacations are not required to be mutually independent.

The *decomposition theorem* (7.30) in the particular case of exhaustive service was first obtained by Skinner [1967] and Cooper [1970]; in that case, $\zeta(z) = 1$ and, as subsequently observed by Levy and Yechiali [1975], Scholl and Kleinrock [1983], and Fuhrmann [1984], the right-hand side of (7.31) is then the probability-generating function of the number of (Poisson) arrivals that occur during a time interval that is distributed as the forward (or backward) recurrence time of a vacation interval (if the rules that govern when the server begins and ends the vacations do not anticipate future jumps of the Poisson arrival process). In certain cases, the queue-length decomposition (7.30) translates directly into a waiting-time decomposition; for example, if the vacations are not anticipative of the arrival process, and the queue discipline is FIFO, then application of (2.13) leads directly to an analogous decomposition of the Laplace–Stieltjes transform of the waiting-time distribution (see, e.g., Exercise 12, p. 222 of Cooper [1981]).

Two particular multiple-vacation exhaustive-service vacation disciplines of interest are the N-policy and T-policy. In the *N-policy*, the server is turned off (sent on vacation) when there are no customers present, and turned on again as soon as there are n customers present; then $\alpha(z) = z^n$. In the *T-policy*, the server is turned off at the end of a busy period and then 'looks' periodically at intervals (i.e., after vacations) of length t, terminating its vacation at a 'look' only when there is a customer waiting; then $\alpha(z) = \mathrm{e}^{-(1-z)\lambda t}$. These models have been analyzed from the viewpoint of *optimal control* (balancing costs of waiting against costs of operating the server) by Yadin and Naor [1963] and Heyman [1968] (N-policy) and Heyman [1977] (T-policy). (See Heyman and Sobel [1984] for a comprehensive discussion of optimization and control of queues.)

Recently, decomposition theorems of the form (7.30) (and similar forms—e.g., Ott [1984]) have been proved to hold for a wide variety of vacation disciplines in M/G/1 (see Fuhrmann and Cooper [1985b], and also Baba [1986], Boxma [1989], Doshi [1990a], Harris and Marchal [1988], Keilson and Servi [1990], Kella and Yechiali [1988], Lee [1988], Shanthikumar [1988], and Wolff [1989]) and even, in some cases, for G/G/1 (see, e.g., Doshi [1985, 1986, 1990c], Fricker [1987], Gelenbe and Iasnogorodski [1980], Keilson and Servi [1986], Kella and Whitt [1989], Lucantoni, Meier-Hellstern, and Neuts [1989], Shanthikumar and Sumita [1989], and Takagi [1989]).

Finally, we note (to come full circle) that decomposition results of the type (7.30) have been applied recently to the analysis of waiting times in cyclic-queueing models—see Fuhrmann [1985], Fuhrmann and Cooper [1985a], Levy and Kleinrock [1986], Servi [1986], Boxma and Groenendijk [1987, 1988], Boxma, Groenendijk, and Weststrate [1988], and Servi and Yao [1989]. The most recent comprehensive surveys of vacation models and polling models are Doshi [1990b] and Takagi [1986, 1987, 1990]. (As the reader can see from these citation dates, papers on this topic are now appearing so frequently that the present discussion will be out of date soon after its appearance.)

8. The GI/M/s queue

We now turn to the classical GI/M/s model: Interarrival times (the times between successive arrival epochs) are *G*eneral *I*ndependent random variables (with distribution function G and mean λ^{-1}), service times are exponentially distributed (with rate μ), there are s servers, blocked customers wait in the queue as long as necessary for service to commence, and the queue discipline is nonbiased. Let $\{\Pi_j\}$ be the arriving customer's equilibrium distribution, as defined by (2.2) and (2.4) (and note that, in general, (2.5) does not apply, because the input process is no longer assumed to be Poisson). As with M/G/1, we can write the imbedded Markov-chain equations for the distribution $\{\Pi_j\}$, where now the imbedding occurs at the arrival (instead of departure) epochs (but note that (2.6) still applies).

It can be shown that, for $\lambda/s\mu < 1$,

$$\Pi_j = Ar^{j-s} \quad (j \geqslant s-1)\,, \tag{8.1}$$

where the constant A will be given shortly, and r is the unique root in $(0, 1)$ of the equation

$$r = \gamma((1-r)s\mu)\,, \tag{8.2}$$

where γ is the Laplace–Stieltjes transform of G. Equation (8.1) shows, significantly, that the distribution $\{\Pi_j\}$ is essentially geometric, with r playing the same role for GI/M/s as that played by ρ $(= \lambda/s\mu = a/s)$ in M/M/s (see (5.22)). Therefore, by the same arguments that led to (5.34) and (5.35) we have, for FIFO,

$$P\{W > t | W > 0\} = \mathrm{e}^{-(1-r)s\mu t} \tag{8.3}$$

and, for any nonbiased queue discipline,

$$E(W|W>0) = \frac{1}{(1-r)s\mu}\,. \tag{8.4}$$

Note that although r replaces ρ in (5.22), (5.34), and (5.35), the server utilization a'/s is still given by $\rho = \lambda/s\mu$; that is, (5.28) and (5.29) remain valid for GI/M/s (of course, when $G(t) = 1 - \mathrm{e}^{-\lambda t}$, then $r = \rho$).

The analogues of (5.33) and (5.34) are

$$P\{W > 0\} = \frac{A}{1-r} \tag{8.5}$$

and

$$E(W) = \frac{A}{(1-r)^2 s\mu}\,. \tag{8.6}$$

The constant A is given by

$$A = \left\{\frac{1}{1-r} + \sum_{j=1}^{s} \frac{1}{C_j(1-\gamma_j)} \binom{s}{j} \frac{s(1-\gamma_j)-j}{s(1-r)-j}\right\}^{-1}, \tag{8.7}$$

where

$$\gamma_j = \gamma(j\mu) \quad (j = 0, 1, \ldots, s) \tag{8.8}$$

and

$$C_j = \prod_{i=1}^{j} \left(\frac{\gamma_i}{1 - \gamma_i} \right) \quad (j = 1, 2, \ldots, s) . \tag{8.9}$$

A derivation of (8.7), as well as the probabilities $\Pi_0, \Pi_1, \ldots, \Pi_{s-2}$, is given in Takács [1962]. The outside observer's distribution $\{P_j\}$ can be calculated from the arriving customer's distribution $\{\Pi_j\}$ and the "rate up = rate down" equations

$$\lambda \Pi_{j-1} = \mu_j P_j \quad (j = 1, 2, \ldots) , \tag{8.10}$$

where μ_j is defined in (5.20) (see Heyman and Stidham [1980]).

In the particular case of GI/M/1, things simplify somewhat and there are similarities with M/G/1. Indeed, it makes sense to define for any M/G/1 queue its *dual* (or *inverse*), which is the GI/M/1 queue that is obtained from its M/G/1 counterpart by interchanging the service-time distribution and the interarrival-time distribution. Then, any busy period can be translated into a busy period in its dual by replacing each arrival epoch by a departure epoch and vice versa, and letting time run in the reverse direction. This observation has been exploited recently by Niu and Cooper [1989] (1) to provide some new derivations of old results and (2) to derive some new results for GI/M/1. (Similar arguments were used in Bhat [1968].)

As an example of (1), consider the duration B of the busy period in GI/M/1. It is well known that

$$P\{B \leqslant t\} = \sum_{j=1}^{\infty} \frac{(\mu t)^j}{j!} \, \mathrm{e}^{-\mu t} \, \frac{1}{t} \int_0^t [1 - G^{*j}(y)] \, \mathrm{d}y , \tag{8.11}$$

which looks remarkably like (7.22). Although (8.11) can be derived from first principles (Takács [1962]), the derivation via its M/G/1 dual 'explains' the striking similarity between (8.11) and (7.22). (The point here is that it makes sense to look for connections between results that look similar but do not, at first glance, yield to term-by-term interpretation or an obvious mapping of one into the other.)

As an example of (2), we give the joint distribution of the number K served during a busy period, its duration B, and the duration I of the idle period that follows:

$$P\{K=n, B \leqslant t, I \leqslant z\}$$
$$= \iint\limits_{\substack{x+y \leqslant t \\ x,y \geqslant 0}} \frac{[\mu(x+y)]^{n-1}}{(n-1)!} \mathrm{e}^{-\mu(x+y)} \left(\frac{y}{x+y}\right)$$
$$\cdot [G(y+z) - G(y)] \mu \, \mathrm{d}y \, \mathrm{d}G^{*(n-1)}(x) \,. \tag{8.12}$$

There are many other facts about the classical single-server models (for example, Cohen's [1982] *The Single Server Queue* runs about 700 pages) but the results given here include the most important and useful, and reflect the flavor of the subject.

9. Other topics

As noted earlier, there are at least 5000 papers on queueing theory (Disney and Kiessler [1987]). Clearly, there are many topics we have not yet discussed. We shall mention a few:

The GI/G/1 *queue*

One of the first studies of the FIFO GI/G/1 queue is Lindley [1952], who began with a relation that (clearly) expresses the waiting time of a customer in terms of the waiting time and service time of the previous customer and the time between their arrival epochs:

$$W_{n+1} = (W_n + X_n - I_{n+1})^+ \,, \tag{9.1}$$

where W_n is the waiting time of the nth arrival, X_n is his service time, I_{n+1} is the interarrival time between the arrival epochs of customer $n+1$ and customer n, and $Z^+ = \max(Z, 0)$. Relation (9.1) can be written as an integral equation,

$$W_{n+1}(t) = \int_{-\infty}^{t} W_n(t-u) \, \mathrm{d}P\{X_n - I_{n+1} \leqslant u\} \quad (t \geqslant 0) \,, \tag{9.2}$$

where $W_n(t) \equiv P\{W_n \leqslant t\}$. Lindley showed that a unique equilibrium distribution $W(t) = \lim_{n\to\infty} W_n(t)$ exists if and only if $\rho \equiv E(X_n)/E(I_n) < 1$ (in which case ρ equals the server utilization); and furthermore, the distribution function $W(t)$ is the solution of the integral equation (which is of the type called *Wiener–Hopf*) that is obtained formally by taking limits in (9.2). Wiener–Hopf equations can be solved by the methods of complex analysis; typically, this requires the use of such theorems from analytic-function theory as *Liouville's theorem* and *Rouché's theorem*. (See for example, Prabhu [1974] and Cohen [1975].) These results are of primarily mathematical interest, and do not usually lead to simple formulas or computational algorithms (but see Ott [1987]). Other methods for the study of GI/G/1 queues include *combinatorial analysis*, *fluctuation theory*, and *random walks* (see e.g., Chapter 3). A survey and synthesis is given in Kingman [1966]. As with M/G/1, there has been for GI/G/1 recent interest in the properties of the LIFO queue discipline (e.g., Fakinos [1987], Niu [1988], Shanthikumar and Sumita [1986], and Yamazaki [1984]) and in vacation models (e.g., Doshi [1986]).

The GI/G/*s queue*

This model, of course, includes GI/G/1 as a special case, but it is much more complicated (see Kingman [1966]). Kiefer and Wolfowitz [1955] (see also Wolfson [1986]) proved the existence of a unique equilibrium waiting-time distribution function for FIFO GI/G/*s* if and only if $\rho \equiv E(X_n)/sE(I_n) < 1$ (again, ρ equals the server utilization). GI/G/*s* was studied in a series of papers by Pollaczek, beginning in the 1930s and spanning over three decades. A discussion of the Pollaczek method is given in Pollaczek [1965] and the appended discussion by Syski, Takács, Kingman, and others. Pollaczek's method has been extended and simplified (somewhat) by de Smit [1973].

Because of the generality and complexity of the GI/G/*s* model, Pollaczek's method does not yield simple formulas or computational algorithms. However, tables of performance measures for GI/G/*s* queues (including the case of finite waiting-room), calculated using the method of phases, are given in Seelen, Tijms, and Van Hoorn [1985].

A formula that approximates the mean waiting time (for any nonbiased queue discipline) in GI/G/*s* is

$$E(W_{\mathrm{GI/G}/s}) = \frac{c_{\mathrm{a}}^2 + c_{\mathrm{s}}^2}{2}\, E(W_{\mathrm{M/M}/s}), \tag{9.3}$$

where c_{a} and c_{s} are, respectively, the coefficients of variation of the interarrival-time and service-time distributions, and $E(W_{\mathrm{M/M}/s})$ is given by the right-hand side of (5.36). Formula (9.3) works well for M/G/*s*, and holds asymptotically as $\rho \to 1$ for GI/G/*s* (the latter statement being an example of a *heavy-traffic* theorem). Whitt [1985b] discusses this and other approximations for GI/G/1 and GI/G/*s*, including *diffusion* approximations (see Chapter 4), and gives tables and many references. Similarly, Tijms [1986], in a long chapter, "Algorithms and approximations for queueing models", devotes a section to GI/G/*s* and gives many references.

Statistical analysis

Typically (as in this chapter), queueing theory is approached from the viewpoint of the probabilist, according to which models are constructed and analyzed under the assumption that the underlying distributions and their parameters are known or can be obtained easily. Therefore, application of queueing-theory formulas often requires that these quantities be measured or estimated, and hence the viewpoint of the statistician becomes relevant (see Chapter 6). A survey on the statistical analysis of queueing systems is given in Bhat and Rao [1987].

Transient analysis

Finally, we remark that this chapter has been focused primarily on queues in equilibrium. *Transient* (or *time-dependent*) analysis, for systems assumed to have been in existence for only a finite length of time, is much more difficult than equilibrium analysis. To illustrate this complexity, we give one form of the transient solution of (5.4) for M/M/1 with initial condition $P_i(0) = 1$:

$$P_j(t) = a^{(j-1)/2}\,\mathrm{e}^{-(1+a)t}\Big[I_{j-i}(2a^{1/2}t) + a^{-1/2}I_{j+i+1}(2a^{1/2}t) + (1-a)\sum_{k=2}^{\infty} a^{-k/2}\,I_{j+i+k}(2a^{1/2}t)\Big]\,, \qquad (9.4)$$

where the functions I are modified Bessel functions. (For related results see Pegden and Rosenshine [1982], Boxma [1984], Towsley [1987], Syski [1988], Baccelli and Massey [1989], and Parthasarathy and Sharafali [1989].) Numerical methods for transient analysis are discussed in Abate and Whitt [1989] and Chapter 5; approximate transient descriptions are discussed in Abate and Whitt [1987, 1988].

10. Concluding remarks

As I have stressed repeatedly throughout this chapter, Queueing Theory has a large literature that seems to be growing exponentially. Therefore, instead of trying to cover as many different models as possible, I have restricted myself to a few basic models, and emphasized what I see as the flavor of the subject, with a proper balance between classical and recent developments. My choice of topics reflects those I feel are most important (and those on which I have worked—one writes what one knows).

I have tried to choose the references with care; the papers and books cited reflect a compromise among several criteria. The cited works are (in my opinion) either (1) historically important, (2) technically seminal, (3) wide-ranging and definitive, with many references, (4) recently published, and therefore not cited in other references, or (5) some combination of these criteria. Unfortunately, many important papers and books are not cited here but, I think, almost all remaining important works are cited in the papers and books cited here; the reader has been provided with a list of pointers.

This chapter is keyed somewhat to my textbook Cooper [1981], which contains an annotated bibliography of almost all English-language hardcover books on Queueing Theory published prior to 1981. Also, detailed solutions to the Exercises in Cooper [1981] are given in Tilt [1981]. Many of the details omitted in the present survey can be supplied by consulting these books. Two other textbooks at about the same mathematical level are Gross and Harris [1985] and Kleinrock [1975, 1976] (with solutions manuals by Kleinrock and Gail [1982, 1986]). Serfozo [1986] gives a survey of recent developments; Syski [1985] and Pakes [1986] provide short surveys in the form of encyclopedia articles; and Prabhu [1987] gives a bibliography of books and survey papers. Heyman and Sobel [1982] cover the ground midway between the queueing-theory texts and those that give a more traditional treatment of the theory of stochastic processes. Franken, König, Arndt and Schmidt [1981], using the theory of stationary random marked point processes, give a unified treatment of many theoretical issues (such as stationarity, ergodicity, insensitivity, and Little's theorem). Whittle [1986] gives the latest on insensitivity. The most important classical books (defined as theoretical works that predate 1970 and remain useful today) include Cohen [1982, first edition 1969], Prabhu [1965], Riordan [1962], Syski [1986, first edition 1960], and Takács [1962, 1967].

For completeness, and to illustrate further the vitality of the field, the following list of recent books, not cited in the annotated bibliography of Cooper [1981] and not previously cited in this chapter, is provided: Agrawal [1985], Akimaru and Cooper [1985], Asmussen [1987], Aven, Coffman, and Kogan [1987], Baccelli and Brémaud [1987], Borovkov [1984], Brémaud [1981], Bruell and Balbo [1980], Bunday [1986], Carmichael [1987], Chaudhry and Templeton [1983], Cohen and Boxma [1983], Conway and Georganas [1989], Daigle [to appear], Fujiki and Gambe [1980], Gelenbe and Mitrani [1980], Gelenbe and Pujolle [1987], Gnedenko and König [1983, 1984], Gnedenko and Kovalenko [1989], Hillier and Yu [1981], Kashyap and Chaudhry [1988], Lavenberg [1983], Medhi [1984], Newell [1984], Prabhu [1980], Shedler [1987], Srivastava and Kashyap [1982], Stoyan [1983], Trivedi [1982], Van Doorn [1981], and Walrand [1988]. Finally, a comprehensive bibliography of books, including works in Russian and Japanese, is given in Takagi and Boguslavsky [1989].

Acknowledgments

I am indebted to my friends and colleagues who were kind enough to read and comment on early versions of this chapter: Kehao Chen, Benjamin C. Cooper, Ralph L. Disney, Bharat Doshi, Winfried K. Grassman, Carl M. Harris, Daniel P. Heyman, Heinrich Niederhausen, Shun-Chen Niu, Linn Sennott, Leslie D. Servi, J. George Shanthikumar, Matthew J. Sobel, Kaoru Sudo, Ryszard Syski, Hideaki Takagi, Jean Walrand, Ward Whitt, and Eric Wolman.

References

Abate, J. and W. Whitt [1987]. Transient behavior of the M/M/1 queue: Starting at the origin. *Queueing Systems* **2** (1), 41–65.

Abate, J. and W. Whitt [1988]. Transient behavior of the M/M/1 queue via Laplace transforms. *Adv. in Appl. Probab.* **20** (1), 145–178.

Abate, J. and W. Whitt [1989]. Calculating time-dependent performance measures for the M/M/1 queue. *IEEE Trans. Comm.* **37** (10), 1102–1104.

Agrawal, S.C. [1985]. *Metamodeling: A Study of Approximations in Queueing Models*. MIT Press, Cambridge, MA.

Akimaru, H. and R.B. Cooper [1985]. *Teletraffic Engineering* (In Japanese). Ohm-sha, Tokyo.

Asmussen, S. [1987]. *Applied Probability and Queues*. Wiley, New York.

Aven, O.I., E.G. Coffman, Jr., and Y.A. Kogan [1987]. *Stochastic Analysis of Computer Storage*. Reidel, Dordrecht–Boston.

Baba, Y. [1986]. On the $M^X/G/1$ queue with vacation time. *Oper. Res. Lett.* **5** (2), 93–98.

Baccelli, F. and P. Brémaud [1987]. *Palm Probabilities and Stationary Queues*. Springer-Verlag, Berlin–New York.

Baccelli, F. and W.A. Massey [1989]. A sample path analysis of the M/M/1 queue. *J. Appl. Probab.* **26** (2), 418–422.

Baskett, F., K.M. Chandy, R.R. Muntz, and F.G. Palacios [1975]. Open, closed and mixed networks of queues with different classes of customers. *J. Assoc. Comput. Machinery* **22** (2), 248–260.

Beneš, V.E. [1957]. On queues with Poisson arrivals. *Ann. Math. Statist.* **28,** 670–677.

Bertsekas, D. and R. Gallager [1987]. *Data Networks*. Prentice-Hall, Englewood Cliffs, NJ.

Bhat, U.N. [1968]. *A Study of the Queuing Systems* M/G/1 *and* GI/M/1. Springer–Verlag, Berlin–New York.

Bhat, U.N. and S.S. Rao [1987]. Statistical analysis of queueing systems. *Queuing Systems* **1** (3), 217–247.

Borovkov, A.A. [1984]. *Asymptotic Methods in Queuing Theory*. Wiley, New York.

Botta, R.F., C.M. Harris, and W.G. Marchal [1987]. Characterizations of generalized hyperexponential distribution functions. *Stochastic Models* **3** (1), 115–148.

Boxma, O.J. [1979]. On a tandem queueing model with identical service times at both counters, I, II. *Adv. in Appl. Probab.* **11** (3), 616–643, 644–659.

Boxma, O.J. [1984]. The joint arrival and departure process for the M/M/1 queue. *Statistica Neerlandica* **38,** 199–208.

Boxma, O.J. [1989]. Workloads and waiting times in single-server systems with multiple customer classes. *Queueing Systems* **5** (1–3), 185–214.

Boxma, O.J. and W.P. Groenendijk [1987]. Pseudo-conservation laws in cyclic-service systems. *J. Appl. Probab.* **24** (4), 949–964.

Boxma, O.J. and W.P. Groenendijk [1988]. Waiting times in discrete-time cyclic-service systems. *IEEE Trans. Comm.* **36** (2), 164–170.

Boxma, O.J., W.P. Groenendijk and J.A. Weststrate [1988]. A pseudoconservation law for service systems with a polling table. Report, Centre for Mathematics and Computer Science, P.O. Box 4079, 1009 AB Amsterdam, The Netherlands.

Brémaud, P. [1981]. *Point Processes and Queues: Martingale Dynamics*. Springer-Verlag, Berlin–New York.

Brémaud, P. [1989]. Characteristics of queueing systems observed at events and the connection between stochastic intensity and Palm probability. *Queueing Systems* **5** (1–3), 99–112.

Brémaud, P. [1990]. Necessary and sufficient condition for the equality of event averages and time averages. To appear in *J. Appl. Probab.* **27**.

Brill, P.H. [1975]. System Point Theory in Exponential Queues. Ph.D. Dissertation, Dept. of Industrial Engineering, University of Toronto.

Brill, P.H. and M.J.M. Posner [1977]. Level crossings in point processes applied to queues: Single-server case. *Oper. Res.* **25** (4), 662–674.

Brill, P.H. and M.J.M. Posner [1981]. The system point method in exponential queues: A level crossing approach. *Math. Oper. Res.* **6** (1), 31–49.

Brockmeyer, E., H.A. Halstrøm, and A. Jensen [1948]. *The Life and Works of A.K. Erlang*. Trans 2. Danish Academy of Technical Sciences, Copenhagen, 1–277.

Bruell, S.C. and G. Balbo [1980]. *Computational Algorithms for Closed Queueing Networks*. North-Holland (Elsevier), Amsterdam.

Bunday, B.D. [1986]. *Basic Queueing Theory*. Edward Arnold, London.

Burke, P.J. [1956]. The output of a queuing system. *Oper. Res.* **4** (6), 699–704.

Burke, P.J. [1972]. Output processes and tandem queues. *Proceedings of the Symposium on Computer-Communications Networks and Teletraffic*. Polytechnic Institute of Brooklyn Press, 419–428.

Burke, P.J. [1975]. Delays in single-server queues with batch input. *Oper. Res.* **23** (4), 830–833.

Carmichael, D.G. [1987]. *Engineering Queues in Construction and Mining*. Ellis Horwood, Chichester.

Chaudhry, M.L. and J.G.C. Templeton [1983]. *A First Course in Bulk Queues*. Wiley, New York.

Cohen, J.W. [1957]. The generalized Engset formulas. *Philips Telecomm. Rev.* **18** (4), 158–170.

Cohen, J.W. [1975]. The Wiener–Hopf technique in applied probability. In: J. Gani (Ed.), *Perspectives in Probability and Statistics* (Papers in honour of M.S. Bartlett). Applied Probability Trust (distributed by Academic Press), 145–156.

Cohen, J.W. [1977]. On up- and downcrossings. *J. Appl. Probab.* **14** (2), 405–410.

Cohen, J.W. [1982]. *The Single Server Queue*, Second Edition. North-Holland (Elsevier), Amsterdam.

Cohen, J.W. and O.J. Boxma [1983]. *Boundary Value Problems in Queueing System Analysis*. North-Holland (Elsevier), Amsterdam.

Conway, A.E. and N.D. Georganas [1989]. *Queueing Networks—Exact Computational Algorithms*. MIT Press, Cambridge, MA.

Cooper, R.B. [1970]. Queues served in cyclic order: Waiting times. *Bell System Tech. J.* **49** (3), 399–413.

Cooper, R.B. [1976]. Queues with ordered servers that work at different rates. *Opsearch* **13** (2), 69–78.

Cooper, R.B. [1981]. *Introduction to Queueing Theory*, Second Edition. North-Holland (Elsevier), Amsterdam. First Edition, 1972, Macmillan, New York. Republished, 1990, by CEEPress, The George Washington University, Washington, DC.

Cooper, R.B. [1987]. Queues with ordered servers that work at different rates: An exact analysis of a model solved approximately by others. *Performance Evaluation* **7** (2), 145–149.

Cooper, R.B. and G. Murray [1969]. Queues served in cyclic order. *Bell System Tech. J.* **48** (3), 675–689.

Cooper, R.B. and S.-C. Niu [1986]. Beneš's formula for M/G/1-FIFO 'explained' by preemptive-resume LIFO. *J. Appl. Probab.* **23** (2), 550–554.

Daigle, J.N. [to appear]. *Queueing Theory for Computer Communications*. Addison-Wesley, Reading, MA.

De Smit, J.H.A. [1973]. Some general results for many server queues. *Adv. in Appl. Probab.* **5** (1), 153–169.

Descloux, A. [1962]. *Delay Tables for Finite- and Infinite-Source Systems*. McGraw-Hill, New York.

Disney, R.L. [1985]. Networks of queues. In: S. Kotz and N.L. Johnson (Eds.), *Encyclopedia of Statistical Sciences*, Vol. 6. Wiley, New York, 191–198.

Disney, R.L. and P.C. Kiessler [1987]. *Traffic Processes in Queueing Networks: A Markov Renewal Approach*. Johns Hopkins University Press, Baltimore, MD.

Disney, R.L. and D. König [1985]. Queueing networks: A survey of their random processes. *SIAM Rev.* **27** (3), 335–403.

Doshi, B.T. [1985]. A note on stochastic decomposition in a GI/G/1 queue with vacations or set-up times. *J. Appl. Probab.* **22** (2), 419–428.

Doshi, B.T. [1986]. Queueing systems with vacations—A survey. *Queueing Systems* **1** (1), 29–66.

Doshi, B.T. [1990a]. Conditional and unconditional distributions for M/G/1 type queues with server vacations. *Queueing Systems*.

Doshi, B.T. [1990b]. Single-server queues with vacations. In: H. Takagi (Ed.), *Stochastic Analysis of Computer and Communication Systems*. North-Holland (Elsevier), Amsterdam.

Doshi, B.T. [1990c]. Generalizations of the stochastic decompostion results for single server queues with vacations. *Stochastic Models* **6** (2).

Engset, T. [1918]. Die Wahrscheinlichkeitsrechnung zur Bestimmung der Wählerzahl in automatischen Fernsprechämtern. *Elektrotech. Z.*, 304–306.

Fakinos, D. [1987]. The single-server queue with service depending on queue size and with the preemptive-resume last-come-first-served queue discipline. *J. Appl. Probab.* **24** (3), 758–767.

Franken, P., D. König, U. Arndt and V. Schmidt [1981]. *Queues and Point Processes*. Akademie-Verlag, Berlin.

Fricker, C. [1987]. Note sur un modèle de file GI/G/1 à service autonome (avec vacances du serveur). *Adv. in Appl. Probab.* **19** (1), 289–291.

Fuhrmann, S.W. [1984]. A note on the M/G/1 queue with server vacations. *Oper. Res.* **32** (6), 1368–1373.

Fuhrmann, S.W. [1985]. Symmetric queues served in cyclic order. *Oper. Res. Lett.* **4** (3), 139–144.

Fuhrmann, S.W. and R.B. Cooper [1985a]. Application of decomposition principle in M/G/1 vacation model to two continuum cyclic queueing models—Especially token-ring LANs. *AT&T Tech. J.* **64** (5), 1091–1099.

Fuhrmann, S.W. and R.B. Cooper [1985b]. Stochastic decompositions in the M/G/1 queue with generalized vacations. *Oper. Res.* **33** (5), 1117–1129.

Fujiki, M. and E. Gambe [1980]. *Teletraffic Theory* (In Japanese). Maruzen, Tokyo.

Gaver, D.P. Jr. [1962]. A waiting line with interrupted service, including priorities. *J. Roy. Statist. Soc. Ser. B* **24,** 73–90.

Gelenbe, E. and R. Iasnogorodski [1980]. A queue with server of walking type (Autonomous service). *Ann. Inst. Henri Poincaré*, Sect. B **16** (1), 63–73.

Gelenbe, E. and I. Mitrani [1980]. *Analysis and Synthesis of Computer Systems*. Academic Press, New York.

Gelenbe, E. and G. Pujolle [1987]. *Introduction to Queueing Networks*. Wiley, New York.

Glynn, P.W. and W. Whitt [1986]. A central-limit-theorem version of $L = \lambda W$. *Queueing Systems* **1** (2), 191–215.

Glynn, P.W. and W. Whitt [1989]. Indirect estimation via $L = \lambda W$. *Oper. Res.* **37** (1), 82–103.

Gnedenko, B.V. and D. König (Eds.) [1983]. *Handbuch der Bedienungstheorie*, Vol. I. Akademie-Verlag, Berlin.

Gnedenko, B.V. and D. König (Eds.) [1984]. *Handbuch der Bedienungstheorie*, Vol. II. Akademie-Verlag, Berlin.

Gnedenko, B.V. and I.N. Kovalenko [1989]. *Introduction to Queueing Theory*, Second Edition: Revised and Supplemented. Birkhäuser, Boston.

Gross, D. and C.M. Harris [1985]. *Fundamentals of Queueing Theory*, Second Edition. Wiley, New York.

Halfin, S. [1983]. Batch delays versus customer delays. *Bell System Tech. J.* **62** (7), Part 1, 2011–2015.

Halfin, S. and W. Whitt [1989]. An extremal property of the FIFO discipline via an ordinal version of $L = \lambda W$. *Stochastic Models* **5** (3), 515–529.

Harris, C.M. and W.G. Marchal [1988]. State dependence in M/G/1 server vacation models. *Oper. Res.* **36** (4), 560–565.

Hebuterne, G. [1988]. Relations between states observed by arriving and departing customers in bulk systems. *Stochastic Process. Appl.* **27,** 279–289.

Heyman, D.P. [1968]. Optimal operating policies for M/G/1 queuing systems. *Oper. Res.* **16** (2), 362–382.

Heyman, D.P. [1977]. The T-policy for the M/G/1 queue. *Management Sci.* **23** (7), 775–778.

Heyman, D.P. [1987]. Asymptotic marginal independence in large networks of loss systems. *Ann. Oper. Res.* **8,** 57–73.

Heyman, D.P. and M.J. Sobel [1982]. *Stochastic Models in Operations Research, Vol. I: Stochastic Processes and Operating Characteristics*. McGraw-Hill, New York.

Heyman, D.P. and M.J. Sobel [1984]. *Stochastic Models in Operations Research, Vol. II: Stochastic Optimization*. McGraw-Hill, New York.

Heyman, D.P. and S. Stidham, Jr. [1980]. The relation between customer and time averages in queues. *Oper. Res.* **28** (4), 983–994.

Hillier, F.S. and O.S. Yu [1981]. *Queueing Tables and Graphs.* North-Holland (Elsevier), Amsterdam.

Jackson, J.R. [1957]. Networks of waiting lines. *Oper. Res.* **5,** 518–521.

Jackson, J.R. [1963]. Jobshop-like queueing systems. *Management Sci.* **10,** 131–142.

Jackson, R.R.P. [1954]. Queueing systems with phase type service. *Oper. Res. Quart.* **5,** 109–120.

Jackson, R.R.P. [1956]. Queueing processes with phase-type service. *J. Roy. Statist. Soc. Ser. B* **18** (1), 129–132.

Jagerman, D.L. [1974]. Some properties of the Erlang loss function. *Bell System Tech. J.* **54** (3), 525–551.

Kashyap, B.R.K. and M.L. Chaudhry [1988]. *An Introduction to Queueing Theory.* Aarkay, 1/3B Southend Park, Calcutta 700029, India.

Kaufman, L. [1983]. Matrix methods for queuing problems. *SIAM J. Sci. Statist. Comput.* **4** (3), 525–552.

Keilson, J. [1962]. Queues subject to service interruption. *Ann. Math. Statist.* **33,** 1314–1322.

Keilson, J. and L.D. Servi [1986]. Oscillating random walk models for GI/G/1 vacation systems with Bernoulli schedules. *J. Appl. Probab.* **23** (3), 790–802.

Keilson, J. and L.D. Servi [1988]. A distributional form of Little's law. *Oper. Res. Lett.* **7** (5), 223–227.

Keilson, J. and L.D. Servi [1990]. The distributional form of Little's Law and Fuhrmann–Cooper Decomposition. *Oper. Res. Lett.* **9** (4).

Kella, O. and W. Whitt [1989]. Queues with vacations and jump-Lévy processes.

Kella, O. and U. Yechiali [1988]. Priorities in M/G/1 queue with server vacations. *Naval Res. Logist.* **35** (1), 23–34.

Kelly, F.P. [1979]. *Reversibility and Stochastic Networks.* Wiley, New York.

Kelly, F.P. [1986]. Blocking probabilities in large circuit-switched networks. *Adv. in Appl. Probab.* **18** (2), 473–505.

Kelly, F.P. [1988]. Routing in circuit-switched networks: Optimization, shadow prices and decentralization. *Adv. in Appl. Probab.* **20** (1), 112–144.

Kendall, D.G. [1951]. Some problems in the theory of queues. *J. Roy. Statist. Soc. Ser. B* **13,** 151–185.

Kendall, D.G. [1953]. Stochastic processes occurring in the theory of queues and their analysis by means of the imbedded Markov chain. *Ann. Math. Statist.* **24,** 338–354.

Khintchine, A.Y. [1932]. Mathematical theory of a stationary queue. (In Russian.) *Mat. Sbornik* **39** (4), 73–84.

Kiefer, J. and J. Wolfowitz [1955]. On the theory of queues with many servers. *Trans. Amer. Math. Soc.* **78,** 1–18.

Kingman, J.F.C. [1966]. *On the Algebra of Queues.* Methuen, London. Originally published in *J. Appl. Probab.* **3** (1966), 285–326.

Kleinrock, L. [1964]. *Communication Nets.* McGraw-Hill, New York. Reprinted by Dover, New York.

Kleinrock, L. [1975]. *Queueing Systems, Vol. I: Theory.* Wiley, New York.

Kleinrock, L. [1976]. *Queueing Systems, Vol. II: Computer Applications.* Wiley, New York.

Kleinrock, L. and R. Gail [1982]. *Solutions Manual for Queueing Systems, Volume I: Theory.* Technology Transfer Institute, Santa Monica, CA.

Kleinrock, L. and R. Gail [1986]. *Solutions Manual for Queueing Systems, Volume II: Computer Applications.* Technology Transfer Institute, Santa Monica, CA.

König, D. [1965]. Generalization of the Engset formulae. (In German.) *Math. Nachr.* **23,** 145–155.

König, D. and V. Schmidt [1980]. Stochastic inequalities between customer-stationary and time-stationary characteristics of queueing systems with point processes. *J. Appl. Probab.* **17** (3), 768–777.

König, D., V. Schmidt, and E.A. van Doorn [1989]. On the 'PASTA' property and a further relationship between customer and time averages in stationary queueing systems. *Stochastic Models* **5** (2), 261–272.

Kosten, L. [1949]. On the validity of the Erlang and Engset loss-formulae. *Het. P.T.T. Bedrijf* **2,** 42–45.

Lavenberg, S.S., Ed. [1983]. *Computer Performance Modeling Handbook*. Academic Press, New York.

Lavenberg, S.S. and M. Reiser [1980]. Stationary state probabilities at arrival instants for closed queueing networks with multiple types of customers. *J. Appl. Probab.* **17** (4), 1048–1061.

Lee, H.W. [1988]. M/G/1 batch arrival queue with variable vacations. APORS '88 (First Conference of the Association of Asian-Pacific Operational Research Societies within IFORS), August 24–26, Seoul, Korea.

Levy, H. and L. Kleinrock [1986]. A queue with starter and a queue with vacations: Delay analysis by decomposition. *Oper. Res.* **34** (3), 426–436.

Levy, Y., and U. Yechiali [1975]. Utilization of idle time in an M/G/1 queueing system. *Management Sci.* **22** (2), 202–211.

Lindley, D.V. [1952]. The theory of queues with a single server. *Proc. Cambridge Phil. Soc.* **48,** 277–289.

Little, J.D.C. [1961]. A proof for the queuing formula: $L = \lambda W$. *Oper. Res.* **9** (3), 383–387.

Lucantoni, D.M., K.S. Meier-Hellstern, and M.F. Neuts [1989]. A single server queue with server vacations and a class of non-renewal arrival processes, *Adv. in Appl. Probab.* **19**.

Makowski, A., B. Melamed and W. Whitt [1989]. On averages seen by arrivals in discrete time. *Proceedings 28th Conference on Decision and Control*. 1084–1086. Tampa, Florida.

Medhi, J. [1984]. *Recent Developments in Bulk Queueing Models*. Wiley, Eastern Ltd.

Melamed, B. [1982]. On Markov jump processes imbedded at jump epochs and their queueing-theoretic applications. *Math. Oper. Res.* **7** (1), 111–128.

Melamed, B. and W. Whitt [1990a]. On arrivals that see time averages. *Oper. Res.* **38** (1).

Melamed, B. and W. Whitt [1990b]. On arrivals that see time averages: A martingale approach. *J. Appl. Probab.* **27** (3).

Mitra, D. and P. Tsoucas [1988]. Relaxations for the numerical solutions of some stochastic problems. *Stochastic Models* **4**, 387–419.

Neuts, M.F. [1975]. Probability distributions of phase type. In: *Liber Amicorum Prof. Emeritus H. Florin*, Department of Mathematics, University of Louvain, Belgium, 173–206.

Neuts, M.F. [1981]. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. The Johns Hopkins University Press, Baltimore, MD.

Neuts, M.F. [1984]. Matrix-analytic methods in queuing theory. *European J. Oper. Res.* **15,** 2–12.

Neuts, M.F. [1988]. Phase-type distributions: A bibliography. Working Paper, Dept. of Systems and Industrial Engineering, University of Arizona, Tucson, AZ 85721.

Neuts, M.F. [1989]. *Structured Stochastic Matrices of* M/G/1 *Type and Their Applications*. Dekker, New York.

Newell, G.F. [1984]. *The* M/M/∞ *Service System with Ranked Servers in Heavy Traffic*. Springer-Verlag, Berlin–New York.

Niu, S.-C. [1984]. Inequalities between arrival averages and time averages in stochastic processes arising from queueing theory. *Oper. Res.* **32** (4), 785–795.

Niu, S.-C. [1988]. Representing workloads in GI/G/1 queues through the preemptive-resume LIFO queue discipline. *Queueing Systems* **3** (2), 157–178.

Niu, S.-C. and R.B. Cooper [1989]. Duality and other results for M/G/1 and GI/M/I queues, via a new ballot theorem. *Math. Oper. Res.* **14** (2), 281–293.

Ott, T.J. [1984]. On the M/G/1 queue with additional inputs. *J. Appl. Probab.* **21** (1), 129–142.

Ott, T.J. [1987]. On the stationary waiting-time distribution in the GI/G/1 queue, I: Transform methods and almost-phase-type distributions. *Adv. in Appl. Probab.* **19** (1), 240–265.

Pakes, A.G. [1986]. Queueing theory. In: S. Kotz and N.L. Johnson (Eds.), *Encyclopedia of Statistical Sciences*, Vol. 7. Wiley, New York, 483–489.

Papaconstantinou, X. and D. Bertsimas [1990]. Relations between the pre-arrival and post-departure state probabilities and the FCFS waiting time distribution in the $E_k/G/s$ queue. *Naval Res. Logis.* **37** (1), 135–149.

Parthasarathy, P.R. and M. Sharafali [1989]. Transient solution to the many-server Poisson queue: A simple approach. *J. Appl. Probab.* **26** (3), 584–594.

Pegden, C.D. and M. Rosenshine [1982]. Some new results for the M/M/1 queue. *Management Sci.* **28** (7), 821–828.

Pollaczek, F. [1930]. Uber eine Aufgabe der Wahrscheinlichkeitstheorie, I–II. *Math. Z.* **32,** 64–100, 729–750.

Pollaczek, F. [1965]. Concerning an analytic method for the treatment of queueing problems. In: W.L. Smith and W.E. Wilkinson (Eds.), *Proceedings of the Symposium on Congestion Theory*. University of North Carolina Press, Chapel Hill, NC, 1–42.

Prabhu, N.U. [1965]. *Queues and Inventories: A Study of Their Basic Stochastic Processes*. Wiley, New York.

Prabhu, N.U. [1974]. Wiener–Hopf techniques in queueing theory. In: A.B. Clarke (Ed.), *Mathematical Methods in Queueing Theory*, Lecture Notes in Economics and Mathematical Systems No. 98. Springer-Verlag, Berlin–New York. 81–90.

Prabhu, N.U. [1980]. *Stochastic Storage Processes: Queues, Insurance Risk, Dams*. Springer-Verlag, Berlin–New York.

Prabhu, N.U. [1987]. A bibliography of books and survey papers on queueing systems: Theory and applications. *Queueing Systems* **2** (4), 393–398.

Ramalhoto, M.F., J.A. Amaral, and M. Teresa Cochito [1983]. A survey of J. Little's formula. *Internat. Statist. Rev.* **51,** 255–278.

Ramaswami, V. and G. Latouche [1989]. An experimental evaluation of the matrix-geometric method for the PH/PH/1 queue. *Stochastic Models* **5** (4), 629–667.

Ramaswami, V. and D.M. Lucantoni [to appear]. *An Introduction to Algorithmic Methods in Queueing Theory*.

Riordan, J. [1961]. Delays for last-come first-served service and the busy period. *Bell System Tech. J.* **40** (3), 785–793.

Riordan, J. [1962]. *Stochastic Service Systems*. Wiley, New York.

Schassberger, R. [1973]. *Warteschlangen*. Springer-Verlag, Berlin–New York.

Schassberger, R. [1986]. Two remarks on insensitive stochastic models. *Adv. in Appl. Probab.* **18** (3), 791–814.

Scholl, M. and L. Kleinrock [1983]. On the M/G/1 queue with rest periods and certain service-independent queueing disciplines. *Oper. Res.* **31** (4), 705–719.

Seelen, L.P., H.C. Tijms and M.H. Van Hoorn [1985]. *Tables for Multiserver Queues*. North-Holland (Elsevier), Amsterdam.

Serfozo, R.F. [1986]. Recent developments in queueing theory and a review of *Asymptotic Methods in Queueing Theory* by A.A. Borovkov. *Appl. Probab. Newslett.* **10** (1), 1–4. See also *SIAM Rev.* **28** (2), 279–283.

Serfozo, R.F. [1989a]. Poisson functionals of Markov processes and queueing networks. *Adv. in Appl. Probab.* **21** (3), 595–611.

Serfozo, R.F. [1989b]. Markovian network processes: Congestion-dependent routing and processing. *Queueing Systems* **5** (1–3), 5–36.

Servi, L.D. [1986]. Average delay approximation of M/G/1 cyclic service queues with Bernoulli schedules. *IEEE J. Sel. Areas Comm.* **4** (6), 813–822.

Servi, L.D. and D.D. Yao [1989]. Stochastic bounds for queueing systems with limited service schedules. *Performance Evaluation* **9** (4), 247–261.

Sevcik, K.C. and I. Mitrani [1981]. The distribution of queueing network states at input and output instants. *J. Assoc. Comput. Mach.* **28** (2), 358–371.

Shanthikumar, J.G. [1988]. On stochastic decomposition in M/G/1 type queues with generalized server vacations, *Oper. Res.* **36** (4), 566–569.

Shanthikumar, J.G. and M.J. Chandra [1982]. Application of level crossing analysis to discrete state processes in queueing systems. *Naval Res. Logist. Quart.* **29** (4), 593–608.

Shanthikumar, J.G. and U. Sumita [1986]. On G/G/1 queues with LIFO-P service discipline. *J. Oper. Res. Soc. Japan* **29** (3), 220–230.

Shanthikumar, J.G. and U. Sumita [1989]. Modified Lindley process with replacement: Dynamic behavior, asymptotic decomposition and applications. *J. Appl. Probab.* **26** (3), 552–565.

Shedler, G.S. [1987]. *Regeneration and Networks of Queues*. Springer-Verlag, Berlin–New York.

Skinner, C.E. [1967]. A priority queuing system with server-walking time. *Oper. Res.* **15** (2), 278–285.

Srivastava, H.M. and B.R.K. Kashyap [1983]. *Special Functions in Queueing Theory and Related Stochastic Processes*. Academic Press, New York.

Stidham, S., Jr. [1974]. A last word on $L = \lambda W$. *Oper. Res.* **22** (2), 417–421.

Stidham, S., Jr. [1981]. Sample-path analysis of queues. In: R.L. Disney, T.J. Ott (Eds.), *Applied Probability-Computer Science: The Interface*, Vol. II. Birkhäuser, Boston–Basel–Stuttgart, 41–70.

Stidham, S., Jr. and M. El Taha [1989]. Sample-path analysis of processes with imbedded point processes. *Queueing Systems* **5** (1–3), 131–166.

Stoyan, D. and (edited with revisions by) D. J. Daley [1983]. *Comparison Methods for Queues and Other Stochastic Models*. Wiley, New York.

Sudo, K. [1987]. *Studies on Moments of Overflow Traffic from a Queue with a Finite Waiting Room*. Master's Thesis, Dept. of Electrical Engineering, Nihon University, Koriyama, Fukushima 963, Japan.

Syski, R. [1985]. Multiserver queues. In: S. Kotz and N.L. Johnson (Eds.), *Encyclopedia of Statistical Sciences*, Vol. 5. Wiley, New York, 727–732.

Syski, R. [1986]. *Introduction to Congestion Theory in Telephone Systems*, Second Edition. North-Holland (Elsevier), Amsterdam.

Syski, R. [1988]. Further comments on the solution of the M/M/1 queue. *Adv. in Appl. Probab.* **20** (3), 693.

Takács, L. [1962]. *Introduction to the Theory of Queues*. Oxford University Press, New York.

Takács, L. [1963]. Delay distributions for one line with Poisson input, general holding times, and various orders of service. *Bell System Tech. J.* **43** (2), 487–503.

Takács, L. [1967]. *Combinatorial Methods in the Theory of Stochastic Processes*. Wiley, New York.

Takács, L. [1969]. On Erlang's formula. *Ann. Math. Statist.* **40,** 71–78.

Takagi, H. [1985]. Analysis of a finite-capacity queue with a resume level. *Performance Evaluation* **5,** 197–203.

Takagi, H. [1986]. *Analysis of Polling Systems*. MIT Press, Cambridge, MA.

Takagi, H. [1987]. Queueing analysis of vacation models, Part I: M/G/1, and Part II: M/G/1 with Vacations, TRL Research Report TR87-0032. Part III: M/G/1 with Priorities, TR87-0038. Part IV: M/G/1//N, TR87-0043. IBM Tokyo Research Laboratory, 5-19 Sanbancho, Chiyoda-ku, Tokyo 102.

Takagi, H. [1989]. Time-dependent analysis of M/G/1 vacation models with exhaustive service. To appear in *Queueing Systems*.

Takagi, H. [1990]. Queueing analysis of polling models: An update. In: H. Takagi (Ed.), *Stochastic Analysis of Computer and Communication Systems*. North-Holland (Elsevier), Amsterdam.

Takagi, H. and L.B. Boguslavsky [1989]. *A Bibliography of Books on Queueing Analysis and Performance Evaluation*.

Takahashi, Y. and Y. Takami [1976]. A numerical method for the steady-state probabilities of a GI/G/*c* queueing system in a general class. *J. Oper. Res. Soc. Japan* **19,** 147–157.

Tijms, H.C. [1986]. *Stochastic Modelling and Analysis: A Computational Approach*. Wiley, New York.

Tilt, B. [1981]. *Solutions Manual for Robert B. Cooper's Introduction to Queueing Theory, Second Edition*. North-Holland (Elsevier), Amsterdam.

Towsley, D. [1987]. An application of the reflection principle to the transient analysis of the M/M/1 queue. *Naval Res. Logist.* **34** (3), 451–456.

Trivedi, K. [1982]. *Probability and Statistics with Reliability, Queuing, and Computer Science Applications*. Prentice-Hall, Englewood Cliffs, NJ.

Van Doorn, E. [1981]. *Stochastic Monotonicity and Queueing Applications of Birth-Death Processes*. Springer-Verlag, Berlin–New York.

Walrand, J. [1988]. *An Introduction to Queueing Networks*. Prentice-Hall, Englewood Cliffs, NJ.

Whitt, W. [1980]. Continuity of generalized semi-Markov processes. *Math. Oper. Res.* **5** (4), 494–501.

Whitt, W. [1983]. Comparing batch delays and customer delays. *Bell System Tech. J.* **62** (7), Part 1, 2001–2009.

Whitt, W. [1985a]. Blocking when service is required from several facilities simultaneously. *AT&T Tech. J.* **64** (8), 1807–1856.

Whitt, W. [1985b]. Approximations for the GI/G/*m* queue. Report, AT&T Bell Laboratories.

Whittle, P. [1986]. *Systems in Stochastic Equilibrium*. Wiley, New York.

Wilkinson, R.I. [1955]. The beginnings of switching theory in th United States. Paper presented at the *First International Teletraffic Congress*, Copenhagen, June 1955. Reprinted in *Electrical Engineering* (published by AIEE), Sept. 1956.

Wishart, D.M.G. [1960]. Queuing systems in which the discipline is 'last come, first-served'. *Oper. Res.* **8,** 591–599.

Wishart, D.M.G. [1961]. An application of Ergodic theorems in the theory of queues. In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 2. University of California Press, Berkeley, CA, 581–592.

Wolff, R.W. [1982]. Poisson arrivals see time averages. *Oper. Res.* **30** (2), 223–231.

Wolff, R.W. [1989]. *Stochastic Modeling and the Theory of Queues*. Prentice-Hall, Englewood Cliffs, NJ.

Wolfson, B. [1986]. The uniqueness of stationary distributions for the GI/G/*s* queue. *Math. Oper. Res.* **3** (11), 514–520.

Yadin, M. and P. Naor [1963]. Queueing systems with a removable service station. *Oper. Res. Quart.* **14** (4), 393–405.

Yamazaki, G. [1984]. Invariance relations of GI/G/1 queueing systems with preemptive-resume last-come–first-served queue discipline. *J. Oper. Res. Soc. Japan* **27** (4), 338–346.

Yashkov, S.F. [1987]. Processor-sharing queues: Some progress in analysis. *Queueing Systems* **2** (1), 1–17.

Yashkov, S.F. [1989]. *Queueing Analysis for Computers*. (In Russian.) Radio E Svyaz', Moscow.