I

A Queueing Theory Primer

In this chapter we summarize the important results to which one is exposed in a first course on queueing theory. This material is drawn from the companion volume [KLEI 75] in which will be found a list of results that key the reader to the location where each result is derived. Our purpose is to lay the foundation for the remainder of the book, which is devoted to the application of this theory in real-world situations; these applications require sound judgment and experience in formulating models as well as in developing operational formulas (exact or approximate) that may be used for analysis and design of systems. We give a rather complete review here (by stating—*not* deriving—results) so that this material will form a self-contained body of results, to be used in later chapters.

Consider any system that has a capacity C, the maximum rate at which it can perform work. Assume that R represents the average rate at which work is demanded from this system. One fundamental law of nature states that if $R < C$ then the system can "handle" the demands placed upon it, whereas if $R > C$ then the system capacity is insufficient and all the unpleasant and catastrophic effects of saturation will be experienced. However, even when $R < C$ we still experience a different set of unpleasantnesses that come about because of the *irregularity* of the demands. For example, consider the corner telephone booth, which on the average can handle the load demanded of it. Suppose now that two people approach that telephone booth almost simultaneously; it is clear that only one of the two can obtain service at a given time and the other must wait in a queue until that one is finished. Such queues arise from two sources: the first is the unscheduled arrival times of the customers; the second is the random demand (duration of service) that each customer requires of the system. The characterization of these two unpredictable quantities (the arrival times and the service times) and the evaluation of their effect on queueing phenomena form the essence of queueing theory. In the following section we introduce some of the usual notation for queueing systems and then we proceed to summarize the major results for various systems.

1.1. NOTATION

Here we introduce only that notation required for the statement of results in this chapter. A more complete listing is given in the glossary at the end of the book.

We let C_n denote the nth customer to arrive at a queueing facility. The important random variables to associate with C_n are

$$\tau_n = \text{arrival time for } C_n \tag{1.1}$$

$$t_n = \tau_n - \tau_{n-1} = \text{interarrival time between } C_n \text{ and } C_{n-1} \tag{1.2}$$

$$x_n = \text{service time for } C_n \tag{1.3}$$

It is the sequence of random variables $\{t_n\}$ and $\{x_n\}$ that really "drives" the queueing system. All these random variables are selected independently of each other, and so we define the two generic random variables

$$\tilde{t} = \text{interarrival time} \tag{1.4}$$

$$\tilde{x} = \text{service time} \tag{1.5}$$

Associated with each is a probability distribution function (PDF), that is,

$$A(t) = P[\tilde{t} \le t] \tag{1.6}$$

$$B(x) = P[\tilde{x} \le x] \tag{1.7}$$

and the related probability density function (pdf), namely,

$$a(t) = \frac{dA(t)}{dt} \tag{1.8}$$

$$b(x) = \frac{dB(x)}{dx} \tag{1.9}$$

In this last definition for the pdf we permit the use of impulse functions as discussed, for example, in Volume I of this text. The moments associated with these random variables are denoted by

$$E[\tilde{t}] = \bar{t} = \frac{1}{\lambda} \tag{1.10}$$

$$E[(\tilde{t})^k] = \overline{t^k} \tag{1.11}$$

$$E[\tilde{x}] = \bar{x} = \frac{1}{\mu} \tag{1.12}$$

$$E[(\tilde{x})^k] = \overline{x^k} \tag{1.13}$$

where the symbol μ is often reserved only for the case of exponentially distributed service times. Furthermore, we need the Laplace transform

associated with these pdf's, namely

$$E[e^{-s\tilde{t}}] = A^*(s) \tag{1.14}$$

$$E[e^{-s\tilde{x}}] = B^*(s) \tag{1.15}$$

The integral representation of this transform [say for $a(t)$] is simply

$$A^*(s) = \int_{0^-}^{\infty} a(t)e^{-st}\,dt \tag{1.16}$$

A key use of this transform is its moment generating property; for example, the moments $\overline{t^k}$ may be generated from $A^*(s)$ through the relationship

$$\left.\frac{d^k A^*(s)}{ds^k}\right|_{s=0} = (-1)^k \overline{t^k} \tag{1.17}$$

We often denote the kth derivative of a function $f(t)$ evaluated at $t = t_0$ by

$$\left.\frac{d^k f(t)}{dt^k}\right|_{t=t_0} = f^{(k)}(t_0) \tag{1.18}$$

Thus Eq. (1.17) may be written as $A^{*(k)}(0) = (-1)^k \overline{t^k}$.

Both $\tilde{t}$ and $\tilde{x}$ are the input random variables to the queueing system; now we must define some of the important *performance* variables, namely, the number of customers in the system, the waiting time per customer, and the total time that a customer spends in the system, that is,

$$N(t) = \text{number of customers in system at time } t \tag{1.19}$$

$$w_n = \text{waiting time (in queue) for } C_n \tag{1.20}$$

$$s_n = \text{system time (queue plus service) for } C_n \tag{1.21}$$

The corresponding limiting random variables (after the system has been in operation a long time) for a stable queue are N, $\tilde{w}$, and $\tilde{s}$. As with $\tilde{t}$ and $\tilde{x}$ we may define the PDF, the pdf, the first moment, and the appropriate transform for N, $\tilde{w}$, and $\tilde{s}$ as follows:

$P[N \le k]$	$W(y) = P[\tilde{w} \le y]$	$S(y) = P[\tilde{s} \le y]$
$P[N = k]$	$w(y) = \dfrac{dW(y)}{dy}$	$s(y) = \dfrac{dS(y)}{dy}$
$E[N] = \bar{N}$	$E[\tilde{w}] = W$	$E[\tilde{s}] = T$
$E[z^N] = Q(z)$	$E[e^{-s\tilde{w}}] = W^*(s)$	$E[e^{-s\tilde{s}}] = S^*(s)$

The study of queues naturally breaks into three cases: elementary queueing theory, intermediate queueing theory, and advanced queueing theory. What distinguishes these three cases are the assumptions regarding $a(t)$ and $b(x)$. In order to name the different kinds of systems we wish to discuss, a rather simple shorthand notation is used for describing queues. This involves a three-component description, A/B/m, which denotes an m-server queueing system where A and B "describe" the interarrival time distribution and service time distribution, respectively. A and B take on values from the following set of symbols, which are meant to remind the reader which distributions they refer to:

M = exponential (i.e., **M**arkovian)

E_r = r-stage **E**rlangian

H_R = R-stage **H**yperexponential

D = **D**eterministic

G = **G**eneral

Specifically, if one of these symbols were used in place of B then it would refer to the following pdf ($x \geq 0$):

$$\text{M:} \quad b(x) = \mu e^{-\mu x} \tag{1.22}$$

$$\text{E}_r\text{:} \quad b(x) = \frac{r\mu(r\mu x)^{r-1}e^{-r\mu x}}{(r-1)!} \tag{1.23}$$

$$\text{H}_R\text{:} \quad b(x) = \sum_{i=1}^{R} \alpha_i \mu_i e^{-\mu_i x} \quad \left(\sum_{i=1}^{R} \alpha_i = 1\right) \quad (\alpha_i \geq 0) \tag{1.24}$$

$$\text{D:} \quad b(x) = u_0\left(x - \frac{1}{\mu}\right) \tag{1.25}$$

G: $b(x)$ is arbitrary

where in the next to last expression $u_0(x - 1/\mu)$ refers to a unit impulse occurring at the position $x = 1/\mu$. Any distribution is permitted when G is assumed. Occasionally we add one or two more items to our three-component description in order to describe the system's storage capacity (denoted by K) or the size of the customer population (denoted by M), and these will be commented on appropriately when used (otherwise they are assumed to be infinite). The simplest interesting system we consider in this chapter is the M/M/1 queue in which we have exponential interarrival times, exponential service times, and a single server (see Section 1.4). The most complicated system we consider in this chapter is G/G/1 in which the exponential distributions are replaced by arbitrary distributions (see Sections 1.2 and 1.10). In this review the majority of our results apply only to the first-come–first-serve queueing discipline; in Chapter 3, we study the effect of other queueing disciplines. Let us now proceed with our summary of results.

1.2. GENERAL RESULTS

Perhaps the most important system parameter for G/G/1 is the *utilization factor* ρ, defined as the product of the average arrival rate of customers to the system times the average service time each requires, that is,

$$\rho = \lambda \bar{x} \tag{1.26}$$

This quantity gives the fraction of time that the single server is busy and is also equal to the ratio of the rate at which work arrives to the system divided by the capacity of the system to do work, that is, R/C as discussed earlier.* In the multiple-server system G/G/m the corresponding definition is

$$\rho = \frac{\lambda \bar{x}}{m} \tag{1.27}$$

which also is equal to R/C and may be interpreted as the expected fraction of busy servers when each server has the same distribution of service time; more generally, ρ is the expected fraction of the system's capacity that is in use. In all cases a stable system (one that yields finite average delays and queue lengths) is one for which

$$0 \leq \rho < 1 \tag{1.28}$$

and we note that the case $\rho = 1$ is not permitted (except in the very special situation of a D/D/m queue). As we shall see, the closer ρ approaches unity, the larger are the queues and the waiting times; it is this quantity that essentially reflects the way in which the system performance varies with the *average* system load.

The average time in system is simply related to the average service time and the average waiting time through the fundamental equation

$$T = \bar{x} + W \tag{1.29}$$

and it is the quantity W that reflects the price we must pay for sharing a given resource (server) with other customers. Whereas ρ is the most important system parameter, it is fair to say that one of the more famous

* On the average, λ customers arrive per second and each brings $\bar{x}$ sec of work for the system; thus $R = \lambda\bar{x}$. The (single-server) system can perform 1 sec of work per second of elapsed time, and so $C = 1$.

formulas from queueing is *Little's result*, which relates the average number in the system to the average arrival rate and the average time spent in that system, namely,

$$\bar{N} = \lambda T \tag{1.30}$$

This result enters most of the calculations we make in this book and is extremely general in its application. The corresponding result for number and time in *queue* is simply given by

$$\bar{N}_q = \lambda W \tag{1.31}$$

where $\bar{N}_q$ is merely the average queue size. Furthermore, it is true in G/G/m that these quantities are related by*

$$\bar{N}_q = \bar{N} - m\rho \tag{1.32}$$

We have already given one fundamental law that applies to queueing systems, namely that $R < C$ in order for the system to be stable. A second common and general law of nature also finds its way into our analyses; it relates the rate at which accumulation within a system occurs as a function of the input and output rates to and from that system. In particular, if we let E_k denote the system state in which k customers are present and if we let

$$P_k(t) = P[N(t) = k] \tag{1.33}$$

which is merely the probability that the system state at time t is E_k, then, loosely stated, we have

$$\frac{dP_k(t)}{dt} = [\text{flow rate of probability into } E_k \text{ at time } t] - [\text{flow rate of probability out of } E_k \text{ at time } t] \tag{1.34}$$

Equation (1.34) will allow us to write down time-dependent relationships among the system probabilities in a straightforward fashion. Now consider a stable system, for which the probability $P_k(t)$ has a limiting value (as $t \to \infty$) which we denote by p_k, (this represents the fraction of time that the system will contain k customers in the steady state). If the interarrival times are exponentially distributed (that is, they form a Poisson arrival process), then the equilibrium probability, r_k, that an arriving customer finds k in the system upon his arrival will in fact equal the long-run probability of there being k customers in the system, that is $p_k = r_k$. On the other hand, if we denote by d_k the equilibrium probability that a departure leaves behind k customers in the system, then $d_k = r_k$ if

* This follows from $T = \bar{x} + W$ and Little's result.

the system state $N(t)$ is permitted to change by at most one at any time. Thus, if we have unit state changes and Poisson arrivals, then we have the situation in which $p_k = r_k = d_k$.

1.3. MARKOV, BIRTH-DEATH, AND POISSON PROCESSES

Before we proceed to discuss the results for elementary queueing systems it is convenient to list some of the well-known results for some simple and important random processes that form the foundation for the queueing results we shall quote.

We begin with discrete-state discrete-time Markov processes such that X_n denotes the discrete value of the (random) process at its nth step. The defining condition for such a Markov chain is

$$P[X_n = j \mid X_{n-1} = i_{n-1}, \ldots, X_1 = i_1] = P[X_n = j \mid X_{n-1} = i_{n-1}] \tag{1.35}$$

This is merely an expression of the fact that the present state completely summarizes all of the pertinent past history so far as that history affects the future of the process. If we let

$$\pi_i^{(n)} = P[X_n = i] \tag{1.36}$$

and denote the vector of these probabilities by

$$\boldsymbol{\pi}^{(n)} = [\pi_0^{(n)}, \pi_1^{(n)}, \ldots] \tag{1.37}$$

and moreover if we denote the one-step transition probabilities for homogeneous Markov chains by

$$p_{ij} = P[X_n = j \mid X_{n-1} = i] \tag{1.38}$$

and collect these into a square matrix denoted by $\mathbf{P} = (p_{ij})$, then we have the basic results for the time-dependent probabilities of this Markov process, namely,

$$\boldsymbol{\pi}^{(n)} = \boldsymbol{\pi}^{(n-1)}\mathbf{P} \tag{1.39}$$

$$\boldsymbol{\pi}^{(n)} = \boldsymbol{\pi}^{(0)}\mathbf{P}^n \tag{1.40}$$

The sequence $\mathbf{P}^n$ $(n = 0, 1, 2, \ldots)$ is equal to the inverse z-transform of the matrix $[\mathbf{I} - z\mathbf{P}]^{-1}$, where $\mathbf{I}$ represents the identity matrix and -1 refers to the matrix inverse. The more useful steady-state behavior of these probabilities may be found by solving the equation

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P} \tag{1.41}$$

along with the condition that

$$\sum_{i=0}^{\infty} \pi_i = 1 \tag{1.42}$$

where we have used the notation $\pi_i = \lim_{n\to\infty} \pi_i^{(n)}$. Finally, we comment that the time the process spends in any state is geometrically distributed (an inherent property of all Markov processes); this distribution is, of course the only discrete memoryless distribution.

Let us now consider the case of a discrete-state continuous-time homogeneous Markov process $X(t)$; here we have a defining property much as we did in Eq. (1.35). The time the process spends in any state is exponentially distributed for all continuous-time Markov processes; this is the only (continuous) memoryless distribution, and it is this property that makes the analysis simple. We now define the transition probabilities as

$$p_{ij}(t) = P[X(s+t) = j \mid X(s) = i] \tag{1.43}$$

The matrix of these transition probabilities will be denoted by $\mathbf{H}(t)$, and in terms of this matrix we may express the Chapman–Kolmogorov equations as

$$\mathbf{H}(t) = \mathbf{H}(t-s)\mathbf{H}(s) \tag{1.44}$$

In a real sense $\mathbf{H}(t)$ corresponds to $\mathbf{P}^n$ and that which corresponds to $\mathbf{P}$ itself is $\mathbf{H}(\Delta t)$ (namely the transition probabilities over an infinitesimal interval). Of more use is the matrix $\mathbf{Q} = [q_{ij}]$, referred to as the infinitesimal generator of the process; it is defined by

$$\mathbf{Q} = \lim_{\Delta t\to 0} \frac{\mathbf{H}(\Delta t) - \mathbf{I}}{\Delta t} \tag{1.45}$$

In terms of this matrix we may then express the time-dependent behavior of our Markov process by the equation

$$\frac{d\mathbf{H}(t)}{dt} = \mathbf{H}(t)\mathbf{Q} \tag{1.46}$$

whose solution is

$$\mathbf{H}(t) = e^{\mathbf{Q}t} \tag{1.47}$$

The steady-state behavior of this process, namely, the stable probabilities $\boldsymbol{\pi}$, are given through the basic equation

$$\boldsymbol{\pi}\mathbf{Q} = 0 \tag{1.48}$$

along with the normalizing equation (1.42). We have occasion to discuss the discrete-state continuous-time and continuous-state continuous-time

processes in Chapter 2 below. (A more complete summary for Markov chains is given in tabular form in the summary of results in Volume I.)

Perhaps the most fundamental random process we encounter in queueing theory is the Poisson process that describes a collection of arrivals for which the interarrival times are independent and exponentially distributed with a mean interarrival time $\bar{t} = 1/\lambda$. In particular, the probability $P_k(t)$ of k arrivals in an interval whose duration is t sec is given by

$$P_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \tag{1.49}$$

The average number of arrivals during this interval is merely

$$\bar{N}(t) = \lambda t \tag{1.50}$$

and the variance is given by

$$\sigma^2_{N(t)} = \lambda t \tag{1.51}$$

We note that the mean and variance for this process are identical. The z-transform for this process is simply given by

$$E[z^{N(t)}] = e^{\lambda t(z-1)} \tag{1.52}$$

The assumption of an exponential interarrival time means, of course,

$$a(t) = \lambda e^{-\lambda t} \qquad t \geq 0 \tag{1.53}$$

which, we repeat, is the memoryless distribution. Here, the mean and variance are, respectively, $\bar{t} = 1/\lambda$ and $\sigma^2 = 1/\lambda^2$.

Among the class of continuous-time Markov processes there is the special case of birth–death processes in which the system state changes by at most one (up or down) in any infinitesimal interval. In such cases we talk about the birth rate λ_k, which is the average rate of births when the system contains k customers, and also of the death rate μ_k, which is the average rate at which deaths occur when the population is of size k. The time-dependent behavior for such a system is essentially given in Eq. (1.47). The equilibrium behavior as defined in Eq. (1.48) takes on an especially simple form for this class of birth–death processes whose solution is given as follows (here we use the more usual notation p_k rather than π_k to denote the probability of having k customers in the system):

$$p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \tag{1.54}$$

with the constant p_0 being evaluated through

$$p_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \lambda_i/\mu_{i+1}} \tag{1.55}$$

The application of this equilibrium solution leads us directly to the class of elementary queueing systems which we discuss in the next three sections.

1.4. THE M/M/1 QUEUE

The M/M/1 queue is the simplest interesting queueing system we present. It is the classic example and the analytical techniques required are rather elementary. Whereas these *techniques* do not carry over into more complex queueing systems, the *behavior* of M/M/1 is in many ways similar to that observed in the more complex cases.

Since this system has a Poisson input (with an average arrival rate λ) and makes unit step changes (single service and single arrivals), then $p_k = r_k = d_k$. (Recall that the average service time is $\bar{x} = 1/\mu$.) This distribution is given by

$$p_k = (1-\rho)\rho^k \tag{1.56}$$

and so we immediately find that the average number in the system is given by

$$\bar{N} = \frac{\rho}{1-\rho} \tag{1.57}$$

with variance

$$\sigma_N^2 = \frac{\rho}{(1-\rho)^2} \tag{1.58}$$

Using Little's result and Eq. (1.32), we may immediately write down the two basic performance expressions for average delays in M/M/1:

$$W = \frac{\rho/\mu}{1-\rho} \tag{1.59}$$

$$T = \frac{1/\mu}{1-\rho} \tag{1.60}$$

The terms $\bar{N}$, W, and T all demonstrate the same common behavior as regards the utilization factor ρ; namely, they all behave inversely with respect to the quantity $(1-\rho)$. This effect is dominant for M/M/1 as well as for most common queueing systems, and in Figure 1.1 we show the average time in system as a function of the utilization factor. Thus as ρ approaches unity from below, these average delays and queue sizes grow without bound! This is true of essentially every queueing system one will encounter and shows the extreme price that must be paid if one is interested in running the system close to its capacity ($\rho = 1$).

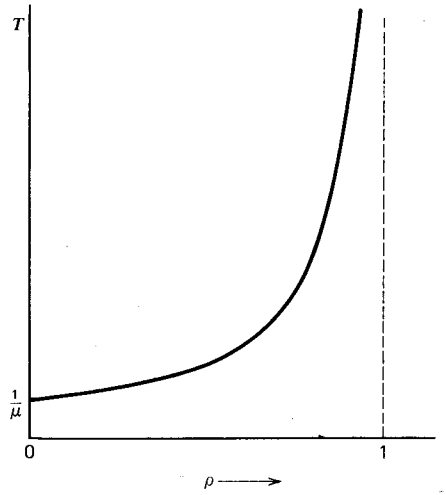

Figure 1.1 Average delay as a function of ρ for M/M/1.

As for the distributions, we have already seen that there is a geometrically distributed number of customers in the system and we now give the waiting time and system time pdf's along with the corresponding PDF's for the case of first-come–first-serve (FCFS):

$$w(y) = (1-\rho)u_0(y) + \lambda(1-\rho)e^{-\mu(1-\rho)y} \qquad y \geq 0 \tag{1.61}$$

[where $u_0(y)$ is the unit impulse (Dirac delta) function],

$$W(y) = 1 - \rho e^{-\mu(1-\rho)y} \qquad y \geq 0 \tag{1.62}$$

$$s(y) = \mu(1-\rho)e^{-\mu(1-\rho)y} \qquad y \geq 0 \tag{1.63}$$

$$S(y) = 1 - e^{-\mu(1-\rho)y} \qquad y \geq 0 \tag{1.64}$$

With the exception of the accumulation of probability at the origin for the waiting time, we note that these are all exponential in nature. The idle period I (the interval of time from the departure of a customer who leaves the system empty until the next arrival) and the interdeparture time D (the time between successive departures) are also both exponentially distributed with the parameter λ:

$$P[I \leq y] = P[D \leq y] = 1 - e^{-\lambda y} \qquad y \geq 0 \tag{1.65}$$

The busy period (the interval of time between successive idle periods) has a pdf denoted by $g(y)$ given in terms of the modified Bessel function of the first kind as

$$g(y) = \frac{1}{y\sqrt{\rho}} e^{-(\lambda+\mu)y} I_1(2y\sqrt{\lambda\mu}) \tag{1.66}$$

The probability f_n that n customers are served during a busy period is given by

$$f_n = \frac{1}{n}\binom{2n-2}{n-1}\rho^{n-1}(1+\rho)^{1-2n} \tag{1.67}$$

Two simple extensions for the M/M/1 system are easily described. First, there is the case of bulk arrivals where with probability g_k a group of k customers arrives at each arrival instant from the Poisson process; we then define the generating function for this distribution as usual by $G(z)=\sum_{k=0}^{\infty} g_k z^k$ with which we may then give the generating function for the number of customers in this bulk arrival M/M/1 system,* namely,

$$Q(z)=\frac{\mu(1-\rho)(1-z)}{\mu(1-z)-\lambda z[1-G(z)]} \tag{1.68}$$

The second generalization is a bulk service system in which a free server will take up to, but no more than, r customers and serve them collectively (as if they were a single customer) with an exponentially distributed service time. The probability of finding k customers in this system is given by

$$p_k = \left(1-\frac{1}{z_0}\right)\left(\frac{1}{z_0}\right)^k \qquad k=0,1,2,\ldots \tag{1.69}$$

where z_0 is that unique root lying outside the unit disk, that is, $|z_0|>1$, for the equation

$$r\rho z^{r+1}-(1+r\rho)z^r+1=0 \tag{1.70}$$

and where, as usual, $\rho=\lambda/r\mu$.

A final generalization, which we will use in Chapter 4, involves the case of an M/M/1 system with a finite number of customers, namely M, that behave in the following way. A customer is either in the system (waiting for or being served) or outside the system and arriving; the interval from the time he leaves the system until he returns once again is exponentially distributed with mean $1/\lambda$. This case gives the following expression for the probability for finding k customers in the system:

$$p_k = \frac{[M!/(M-k)!](\lambda/\mu)^k}{\sum_{i=0}^{M}[M!/(M-i)!](\lambda/\mu)^i} \tag{1.71}$$

* That is, recall $Q(z)=E[z^N]$, not to be confused with the infinitesimal generator $\mathbf{Q}$ defined in Eq. (1.45).

So much for the classic M/M/1 system. In the next section, we retain the Markovian assumptions but consider the case of multiple servers.

1.5. THE M/M/m QUEUEING SYSTEM

We now consider the generalization to the case of m servers. A single queue forms in front of this collection of m servers and the customer at the head of the queue will be handled by the first available server. As usual, λ is the arrival rate and $1/\mu$ is the average service time, with $\rho \triangleq \lambda/m\mu$. The equilibrium probability of finding k customers in the system is given by

$$p_k = \begin{cases} p_0 \dfrac{(m\rho)^k}{k!} & k \le m \\ p_0 \dfrac{(\rho)^k m^m}{m!} & k \ge m \end{cases} \tag{1.72}$$

where

$$p_0 = \left[\sum_{k=0}^{m-1}\frac{(m\rho)^k}{k!}+\frac{(m\rho)^m}{m!\,(1-\rho)}\right]^{-1} \tag{1.73}$$

A. K. Erlang, the father of queueing theory, considered this system as one model for the behavior of telephone systems early in this century [BROC 48]. Identified with his name is the Erlang-C formula, which gives the probability that an arriving customer must wait for a server; his expression is given by p_m from Eq. (1.72). Extensive tables of this quantity are available in the many books dealing with telephony [TELE 70].

Further results for M/M/m may be found in Section 1.9, which discusses G/M/m. Specifically W and $W(y)$ are given in Eqs. (1.113) and (1.114), respectively, where for M/M/m we have simply that $\sigma=\rho$.

Erlang considered a second model for telephone systems that is the same as M/M/m but permits no customers to wait; that is, it is a loss system with at most m customers present at any one time. In this case, the probability of finding k customers in the system is given by

$$p_k = \frac{(\lambda/\mu)^k/k!}{\sum_{i=0}^{m}(\lambda/\mu)^i/i!} \tag{1.74}$$

for the range $0 \le k \le m$. The important quantity of interest here is the probability that a customer upon arrival to the system will find no empty servers and will therefore be "lost;" this is referred to as the Erlang-B formula or as Erlang's Loss Formula (also commonly tabulated) and is given simply by p_m from Eq. (1.74).

1.6. MARKOVIAN QUEUEING NETWORKS

Before leaving the comfortable world of exponential distributions, we wish to discuss another class of results that applies to networks of queues in which customers move from one queueing facility to another in some random fashion until they depart from the system at various points. Specifically, we consider an N-node network in which the ith node consists of a single queue served by m_i servers, each of which has an exponentially distributed service time of mean $1/\mu_i$. The ith node receives from outside the network a sequence of arrivals from an independent Poisson source at an average rate of γ_i customers per second. When a customer completes service at the ith node he will proceed next to the jth node with probability r_{ij}; thus he becomes an "internal" arrival to the jth node. On the other hand, upon leaving the ith node a customer will depart from the entire network with probability $1-\sum_{j=1}^{N} r_{ij}$. We define the total arrival rate to the ith node to be, on the average, λ_i customers per second, and this consists both of external and internal arrivals. The set of defining equations for λ_i is given by

$$\lambda_i = \gamma_i + \sum_{j=1}^{N} \lambda_j r_{ji} \tag{1.75}$$

A large measure of independence exists among the nodes in such a network, as may be seen from the expression given below for the joint distribution of finding k_1 customers in the first node, k_2 customers in the second node, and so on:

$$p(k_1, k_2, \ldots, k_N) = p_1(k_1)p_2(k_2)\ldots p_N(k_N) \tag{1.76}$$

The factoring of this joint distribution exposes the independence. In Chapters 4 and 5 we are delighted to take advantage of this independence. In particular, each factor in this last expression, say $p_i(k_i)$, is merely the solution to an *isolated* M/M/m_i queueing facility operating by itself with an input rate λ_i; the solution for $p_i(k_i)$ is given in Eq. (1.72).

Another class of Markovian queueing networks consists of those networks in which customers are permitted neither to leave nor to enter. In particular, we assume that K customers are placed (trapped) within a network similar to the one described above and that they move around from node to node, but no departures from any node are permitted; that is, $1-\sum_{j=1}^{N} r_{ij} = 0$ for all i. These closed networks have the following solution for the joint distribution of finding customers in various nodes:

$$p(k_1, k_2, \ldots, k_N) = \frac{1}{G(K)} \prod_{i=1}^{N} \frac{x_i^{k_i}}{\beta_i(k_i)} \tag{1.77}$$

where the set of numbers $\{x_i\}$ must satisfy the following linear equations [similar to Eq. (1.75) with $\gamma_i = 0$]:

$$\mu_i x_i = \sum_{j=1}^{N} \mu_j x_j r_{ji} \qquad i = 1, 2, \ldots, N \tag{1.78}$$

and where

$$G(K) = \sum_{\mathbf{k}\in A} \prod_{i=1}^{N} \frac{x_i^{k_i}}{\beta_i(k_i)} \tag{1.79}$$

where $\mathbf{k} = (k_1, k_2, \ldots, k_N)$ and A is that set of vectors $\mathbf{k}$ for which $k_1 + k_2 + \cdots + k_N = K$ and where

$$\beta_i(k_i) = \begin{cases} k_i! & k_i \leq m_i \\ m_i!\, m_i^{k_i - m_i} & k_i \geq m_i \end{cases} \tag{1.80}$$

These open and closed networks will be developed further in Chapter 4.

1.7. THE M/G/1 QUEUE

In this and the following two sections we study systems that fall in the domain of intermediate queueing theory. This classification refers to those systems in which we permit either (but not both) the interarrival time or the service time to be nonexponentially distributed; the case when both these random variables are nonexponential forms part of advanced queueing theory which we discuss in Section 1.10. For the M/G/1 system we cannot give explicit distributions for the number in system or for the time in system as we did for the M/M/1 system [specifically, see Eqs. (1.56) and (1.64) above]. Rather, we find expressions for the transforms of these distributions.

The M/G/1 system is characterized by a Poisson arrival process at a mean rate of λ arrivals per second and with an arbitrary or general service time distribution of form $B(x)$ with a mean service time of $\bar{x}$ sec and with kth moment equal to $\overline{x^k}$. Due to the Poisson arrival process and due to the fact that the number in the system changes by at most one, we again have $p_k = r_k = d_k$.

The basic (difference) equation describing the relationship among random variables for this first-come–first-serve M/G/1 system is

$$q_{n+1} = \begin{cases} q_n - 1 + v_{n+1} & \text{if } q_n > 0 \\ v_{n+1} & \text{if } q_n = 0 \end{cases} \tag{1.81}$$

where q_n is the number of customers left behind by the departure of customer C_n, and v_n is the number of customers who enter during his

service time (x_n). The sequence $\{q_n\}$ forms a (discrete-state continuous-time) Markov chain. The entire transient and equilibrium behavior for the system is contained in this equation, and from it we may derive most of our results for M/G/1.

By far the most well-known result for the M/G/1 system is the Pollaczek–Khinchin (P-K) mean value formula, which gives the following compact expression for the (equilibrium) average waiting time in the queue:

$$W = \frac{\lambda \overline{x^2}/2}{(1-\rho)} \tag{1.82}$$

The numerator term, denoted by $W_0 = \lambda \overline{x^2}/2$, is, in fact, equal to the expected time that a newly arriving customer must spend in the queue while that customer (if any) which he finds in service completes his remaining required service time.* From this formula one may easily calculate T using Eq. (1.29); combining that result with the results quoted in Eqs. (1.31) and (1.32) we easily come up with the P-K mean-value formula for number in system as

$$\bar{N} = \rho + \frac{\lambda^2 \overline{x^2}/2}{1-\rho} \tag{1.83}$$

* This quantity is related to the concept of *residual life*, which we will use in this book. To elaborate, let us consider the sequence of instants located on the real-time axis such that the set of distances between adjacent points is a set of independent, identically distributed random variables whose density we shall denote by $f(x)$ (that is, we are dealing with a renewal process). Let m_n denote the nth moment of these interval lengths. Let us now select a point along the time axis at random; the interval in which this point falls will be referred to as the "sampled" interval. The length of the sampled interval is known as the *lifetime* of the interval, the time from the start of the sampled interval to this point is known as the *age* of the interval, and the distance from this selected point until the end of the sampled interval is known as the *residual life* of the interval. We are concerned with the statistics of the residual life. The pdf for residual life is given by $\hat{f}(x) = [1-F(x)]/(m_1)$ and the Laplace transform of this density is given by $\hat{F}^*(s) = [1-F^*(s)]/(sm_1)$; the notation here is that $F(x) = \int_0^x f(y)\,dy$ and $F^*(s)$ is the Laplace transform associated with the pdf $f(x)$. Perhaps the most significant statistic is the *mean residual life*, given by $m_2/2m_1$; that is, the expected value of the remaining length of the interval is merely the second moment over twice the first moment of the interval lengths themselves. Also, the pdf for the lifetime of the sampled interval is $xf(x)/m_1$.

The last quantity we wish to describe is the probability that the length of an interval (or that the value of any random variable) lies between x and $x+dx$ given that it exceeds x; dividing this probability by dx, we have a quantity referred to as the *failure rate* of the random variable, given by $f(x)/[1-F(x)]$, where f and F refer to the pdf and the PDF of the random variable itself.

One sees that W_0 is merely the mean residual life of a service time (i.e., the average remaining service time) ($\overline{x^2}/2\bar{x}$) times the probability ($\rho = \lambda\bar{x}$) that, in fact, someone is occupying the service facility.

As mentioned above, the best we can do regarding the distributions of the various performance measures is to give the transforms associated with these random variables. Specifically, then, we recall the definition of the z-transform for the distribution p_k to be $Q(z) = \sum_{k=0}^{\infty} p_k z^k$ and find that it is given through

$$Q(z) = B^*(\lambda - \lambda z)\frac{(1-\rho)(1-z)}{B^*(\lambda-\lambda z)-z} \tag{1.84}$$

where $B^*(\lambda - \lambda z)$ is the Laplace transform of the service time density $b(x)$ evaluated at the point $s = \lambda - \lambda z$. This last is referred to as the P-K transform equation for the number in system, and from it we easily derive Eq. (1.83).† The Laplace transform of the waiting time pdf is merely

$$W^*(s) = \frac{s(1-\rho)}{s - \lambda + \lambda B^*(s)} \tag{1.85}$$

and for the time in system we have

$$S^*(s) = B^*(s)\frac{s(1-\rho)}{s-\lambda+\lambda B^*(s)} \tag{1.86}$$

These last two equations are also referred to as P-K transform equations. Due to the independence of service times, we see that Eq. (1.86) is related to Eq. (1.85) through the obvious relationship $S^*(s) = B^*(s)W^*(s)$, that is, the transform for the pdf of the sum of two independent random variables is equal to the product of the transforms of the pdf of each separately. From Eq. (1.85) we easily obtain W in Eq. (1.82) by differentiation as usual; similarly, the second moment (and therefore, the variance of the waiting time, denoted by σ_w^2) may be obtained to give

$$\sigma_w^2 = W^2 + \frac{\lambda \overline{x^3}}{3(1-\rho)} \tag{1.87}$$

Because of the Poisson arrival process, one immediately finds that the idle time I is distributed exponentially, that is,

$$P[I \le y] = 1 - e^{-\lambda y} \tag{1.88}$$

The busy-period duration has a pdf whose transform $G^*(s)$ is given through the functional equation

$$G^*(s) = B^*(s + \lambda - \lambda G^*(s)) \tag{1.89}$$

† For the case of bulk arrivals as discussed in introducing Eq. (1.68) above, the M/G/1 system gives an expression for $Q(z)$ identical to that in Eq. (1.84), except that $B^*(s)$ is evaluated at the point $s = \lambda - \lambda G(z)$ rather than as above; $G(z)$ is as given for Eq. (1.68).

which, in general, cannot be solved. However, we may determine various moments of the busy period through the moment-generating properties of this transform, and so, for example, g_1 (the mean duration of the busy period) and σ_g^2 (the variance of this duration) are given by

$$g_1 = \frac{\bar{x}}{1-\rho} \tag{1.90}$$

$$\sigma_g^2 = \frac{\sigma_b^2 + \rho(\bar{x})^2}{(1-\rho)^3} \tag{1.91}$$

where σ_b^2 is the variance of the service time. Similarly, the z-transform for the number served during the busy period, which we denote by $F(z)$, is given functionally by

$$F(z) = zB^*[\lambda - \lambda F(z)] \tag{1.92}$$

with mean and variance for this number given respectively by

$$h_1 = \frac{1}{1-\rho} \tag{1.93}$$

$$\sigma_h^2 = \frac{\rho(1-\rho) + \lambda^2\overline{x^2}}{(1-\rho)^3} \tag{1.94}$$

An important stochastic process, which we have so far neglected, is the unfinished work $U(t)$ in the system at time t. This is a Markov process whose value represents the time required to empty the system of all customers present at time t, assuming that no new customers enter the system after time t; that is, $U(t)$ is the system backlog expressed in time units.

For a first-come–first-serve system, the unfinished work also represents the waiting time of an arrival *if* it were to enter at time t, and so $U(t)$ is sometimes referred to as the "virtual" waiting time; in the case of a first-come–first-serve system with Poisson arrivals (M/G/1), the unfinished work has the same statistics as the true waiting time for arrivals. We shall deal with this function in numerous places throughout the balance of this book. For the moment we wish to quote two important results regarding its distribution. For this purpose we define

$$F(w, t) = P[U(t) \le w] \tag{1.95}$$

and we may then cite the well-known Takács integrodifferential equation, namely,

$$\frac{\partial F(w, t)}{\partial t} = \frac{\partial F(w, t)}{\partial w} - \lambda F(w, t) + \lambda \int_{x=0}^{w} B(w-x)\, d_x F(x, t) \tag{1.96}$$

which defines the transient behavior of the unfinished work distribution. Defining the double Laplace transform $F^{**}(r, s)$ for $F(w, t)$, where r carries out the transform in the w-domain and s in the t-domain, we have the following transform equation for this time-dependent behavior:

$$F^{**}(r, s) = \frac{(r/\eta)e^{-\eta w_0} - e^{-r w_0}}{\lambda B^*(r) - \lambda + r - s} \tag{1.97}$$

Here η is the unique root (for r) of the equation $s - r + \lambda - \lambda B^*(r) = 0$ in the region $\mathrm{Re}\,(s) > 0$, $\mathrm{Re}\,(r) > 0$, and w_0 is the initial value of the unfinished work at time 0, that is, $U(0) = w_0$. We make use of these transient results in Chapter 2.

Much more can be said about the M/G/1 system, but for purposes of this primer we have said enough. In the natural order of things we should next consider the system M/G/m, but unfortunately there are very few substantive results that can be given for this system. On the other hand, the limiting case for the M/G/∞ system is itself in some ways a trivial system since no queueing ever takes place; indeed, a very lovely result for the number of busy servers (that is the number of customers in the system) is given simply by

$$p_k = \frac{\rho^k}{k!} e^{-\rho} \tag{1.98}$$

We note that this result is independent of the form for $B(x)$, depending only upon its first moment. Similarly we can immediately write down that $T = \bar{x}$ and $s(y) = b(y)$.

It is possible to interpret some of the above transforms as probabilities using the *method of collective marks*. The concept is to assume that each entering customer is "marked" independently with probability $(1-z)$. Then we may interpret the generating function $P(z, t) = E[z^{N(t)}]$ for an arrival process [e.g., for Poisson arrivals, $P(z, t) = e^{\lambda t(z-1)}$] as being equal to the probability that no customers arriving in $(0, t)$ are marked. Similarly, consider any interval whose duration is given by a random variable X whose pdf has a Laplace transform, say, $X^*(s)$; if we further consider an independent Poisson arrival process (at mean rate λ) and ask for the probability P that no arrivals are marked that enter during the interval X, then $P = X^*(\lambda - \lambda z)$. Again consider an interval and an independent Poisson process as above; let us think of the epochs generated by the Poisson process as "catastrophes." If we ask for the probability Q that no catastrophes occur in the random interval, then $Q = X^*(\lambda)$. Thus we are able to give interesting probabilistic interpretations for many of the basic transform expressions that we encounter in queueing theory.

1.8. THE G/M/1 QUEUE

The G/M/1 system is in fact the "dual" of the M/G/1 system. Surprisingly, G/M/1 yields to analysis more easily than M/G/1 and so we can quote distributions directly. The system, of course, corresponds to the case of an arbitrary interarrival time whose PDF is given by $A(t)$ and with pdf $a(t)$ the transform of which is denoted by $A^*(s)$; service times are distributed exponentially with mean $1/\mu$.

The basic recurrence relation that governs the behavior of G/M/1 (and also G/M/m), similar to that for M/G/1 given in Eq. (1.81), is

$$q'_{n+1} = q'_n + 1 - v'_{n+1} \tag{1.99}$$

where q'_n is the number of customers found in the system by C_n and v'_{n+1} is the number of customers served between the arrival of C_n and C_{n+1}. The sequence $\{q'_n\}$ forms a Markov chain. Many of the G/M/m results follow from this equation.

All our results are expressed in terms of a root σ that is the unique root in the range $0 \le \sigma < 1$ of the functional equation

$$\sigma = A^*(\mu - \mu\sigma) \tag{1.100}$$

Once σ is evaluated, the following results are immediately available. The distribution for the number of customers found in the system by a new arrival is given by

$$r_k = (1-\sigma)\sigma^k \qquad k = 0, 1, 2, \ldots \tag{1.101}$$

The PDF for waiting time is given by

$$W(y) = 1 - \sigma e^{-\mu(1-\sigma)y} \qquad y \ge 0 \tag{1.102}$$

and the mean waiting time is

$$W = \frac{\sigma}{\mu(1-\sigma)} \tag{1.103}$$

It is remarkable that the waiting times are exponentially distributed, independent of the form of the interarrival time distribution (except insofar as it affects the value for σ).

1.9. THE G/M/m QUEUE

In contrast to the M/G/m system, we find that the G/M/m system does in fact yield to analysis, the results for which we quote in this section. The G/M/m system, of course, has arbitrarily distributed interarrival times and a single queue served first-come–first-serve by m servers, each of which

has an exponentially distributed service time of mean $1/\mu$. As with the system G/M/1, σ is a key parameter and in this case it is found as the unique solution in the range $0 \le \sigma < 1$ for the equation

$$\sigma = A^*(m\mu - m\mu\sigma) \tag{1.104}$$

We have that the distribution of queue size found by a new arrival, conditioned by the fact that this arrival must queue, is given by

$$P[\text{queue size} = n \mid \text{arrival queues}] = (1-\sigma)\sigma^n \qquad n \ge 0 \tag{1.105}$$

We note here as with the G/M/1 system that the queue size is geometrically distributed. As earlier, we define r_k as the probability that a newly arriving customer finds k in the system ahead of him; in terms of these probabilities we define

$$R_k = \begin{cases} r_k/J & 0 \le k \le m-2 \\ \sigma^{k-m+1} & m-2 < k \end{cases} \tag{1.106}$$

We must evaluate J and the $m-1$ terms R_k for $0 \le k \le m-2$. The equation for J is given by

$$J = \frac{1}{[1/(1-\sigma)] + \sum_{k=0}^{m-2} R_k} \tag{1.107}$$

and the values for the terms R_k are given through the set of equations

$$R_{k-1} = \frac{R_k - \sum_{i=k}^{m-2} R_i p_{ik} - \sum_{i=m-1}^{\infty} \sigma^{i+1-m} p_{ik}}{p_{k-1,k}} \tag{1.108}$$

where the transition probabilities p_{ij} are nontrivial and are calculated through the following four equations, depending on the range of the subscripts i and j:

$$p_{ij} = 0 \qquad j > i+1 \tag{1.109}$$

$$p_{ij} = \int_0^\infty \binom{i+1}{j} [1 - e^{-\mu t}]^{i+1-j} e^{-\mu t j}\, dA(t) \qquad j \le i+1 \le m \tag{1.110}$$

$$\beta_n = p_{i,i+1-n} = \int_{t=0}^\infty \frac{(m\mu t)^n}{n!} e^{-m\mu t}\, dA(t) \qquad 0 \le n \le i+1-m,\ m \le i \tag{1.111}$$

$$p_{ij} = \int_0^\infty \binom{m}{j} e^{-j\mu t} \left[\int_0^t \frac{(m\mu y)^{i-m}}{(i-m)!} (e^{-\mu y} - e^{-\mu t})^{m-j} m\mu\, dy \right] dA(t) \qquad j < m < i+1 \tag{1.112}$$

(Who said it would be easy!) Once these constants are evaluated we may then calculate the average waiting time as

$$W = \frac{J\sigma}{m\mu(1-\sigma)^2} \tag{1.113}$$

The PDF of the waiting time is given through

$$W(y) = 1 - \frac{\sigma e^{-m\mu(1-\sigma)y}}{1+(1-\sigma)\sum_{k=0}^{m-2} R_k} \qquad y \geq 0 \tag{1.114}$$

Whereas these last two equations require the calculation of difficult constants, the waiting time pdf conditioned on the fact that the customer must queue is simply given by

$$w(y \mid \text{arrival queues}) = (1-\sigma)m\mu e^{-m\mu(1-\sigma)y} \qquad y \geq 0 \tag{1.115}$$

This only requires the calculation of σ. Note that even for the G/M/m system we have an exponentially distributed conditional waiting time.

1.10. THE G/G/1 QUEUE

Advanced queueing theory deals with the system G/G/1 and things beyond (for example, G/G/m, about which we can say so very little—recall that even the system M/G/m confounded us). In this section we give some of the principal well-known results for G/G/1 and describe a method of attack that sometimes yields the required solution or at least some simplified measures of performance. In addition we present a point of view that describes the underlying operations involved in solving the G/G/1 system.

As mentioned in the first section of this chapter, the random variables that drive any queueing system are the interarrival times t_n and the service times x_n. In the general formulation of the G/G/1 system, we find that these random variables do not appear separately in the solution but in fact always appear as a difference; thus we are led to consider a new random variable associated with the nth customer C_n, namely,

$$u_n = x_n - t_{n+1} \tag{1.116}$$

This random variable represents the difference between the amount of work (x_n) that C_n demands of the system and the "breathing space" (t_{n+1}), or time, between the arrival of this demand and the arrival of the next demand by C_{n+1}; hopefully this difference will be negative on the average so that there will be more breathing space than load on the system. In fact

if we take the average of Eq. (1.116) we find

$$E[u_n] = \bar{t}(\rho - 1) \tag{1.117}$$

which, first of all, is independent of n (as we expected) and, second, will have a negative mean value so long as $\rho < 1$; this is no different than requiring that $R < C$ if our system is to be stable. Associated with the random variable u_n, whose generic form we now write as $\tilde{u}$, we have its PDF $C(u)$, its pdf $c(u)$ and the Laplace transform of this pdf, which we denote by $C^*(s)$. Expressing these last two in terms of the pdf's and Laplace transforms thereof for the interarrival time and service times we have

$$c(u) = \int_{-\infty}^{\infty} b(u+t)a(t)\,dt \tag{1.118}$$

and

$$C^*(s) = A^*(-s)B^*(s) \tag{1.119}$$

The integral in Eq. (1.118) is, of course, the convolution integral between $a(-u)$ and $b(u)$, which we henceforth denote by $c(u) = a(-u) \circledast b(u)$. Thus once we know the interarrival time and service time pdf we also have the pdf for our random variable $\tilde{u}$.

Of basic interest to the G/G/1 system is the behavior of the waiting time w_n for customer C_n. This random variable is related to others in the sequence through the following difference equation, in which we see the basic role played by the random variable u_n:

$$w_{n+1} = \max[0, w_n + u_n] \tag{1.120}$$

This is the key defining equation for G/G/1 [as was Eq. (1.81) for M/G/1 and Eq. (1.99) for G/M/m]. The sequence $\{w_n\}$ forms a (continuous-time continuous-state) Markov process (in fact, it is an imbedded Markov process). The maximum operator shown above is often rewritten in the following fashion: $(x)^+ = \max(0, x)$. In the case of a stable system ($\rho < 1$) there will exist a limiting random variable representing the equilibrium waiting time, which we denote by $\tilde{w}$. It can be seen from Eq. (1.120) that $\tilde{w}$ must have the same distribution as $(\tilde{w} + \tilde{u})^+$; the pdf that satisfies this condition will be the unique solution for the waiting time pdf. Let us denote the pdf for w_n by $w_n(y)$. The (nonlinear) functional equation that defines this pdf is given through

$$w_{n+1}(y) = \pi(w_n(y) \circledast c(y)) \tag{1.121}$$

where $\circledast$ is the convolution operator and π is a special operator that modifies the pdf of its argument by replacing all of the probability associated with negative values of y (the argument of the pdf) with an impulse at $y = 0$

whose area equals this probability. The pdf $w(y)$ for our limiting random variable $\tilde{w}$ must, from Eq. (1.121), satisfy the following basic equation:

$$w(y) = \pi(w(y) \circledast c(y)) \tag{1.122}$$

whose solution will be the equilibrium density for the waiting time in G/G/1. Equation (1.122) states that this equilibrium pdf must be such that when it is convolved with $c(y)$ and when the resulting density has all of its probability on the negative half-line moved to an impulse at the origin, then we must have a resulting pdf that is the same as the $w(y)$ with which we began.

Another way to describe the random variable $\tilde{w}$ is through the equation

$$\tilde{w} = \sup_{n \geq 0} U_n \tag{1.123}$$

where $U_n = u_0 + u_1 + \cdots + u_{n-1}$ $(n \geq 1)$ and $U_0 = 0$.

A random variable related to w_n that forms the "other half" for w_n is

$$y_n = -\min[0, w_n + u_n] \tag{1.124}$$

Thus we see that

$$w_{n+1} - y_n = w_n + u_n \tag{1.125}$$

Taking expectations of this equation in the limit as $n \to \infty$, we obtain

$$\bar{y} = -\bar{u} \tag{1.126}$$

Another defining relationship for the waiting time PDF is given by the well-known Lindley's integral equation:

$$W(y) = \begin{cases} \int_{-\infty}^{y} W(y-u)\, dC(u) & y \geq 0 \\ 0 & y < 0 \end{cases} \tag{1.127}$$

This equation is of the Wiener–Hopf type. We now let $\Phi_+(s)$ denote the Laplace transform for the waiting time PDF $W(y)$; note that this is the transform for the PDF and not for the pdf $w(y)$, whose transform we had previously denoted by $W^*(s)$ and which is related to this new transform through the equation $W^*(s) = s\Phi_+(s)$. We wish to solve for $\Phi_+(s)$. The procedure we are about to describe is formally correct for those G/G/1 systems for which $A^*(s)$ and $B^*(s)$ may be written as rational functions of s. In this case our task is to find a suitable representation of the following form:

$$A^*(-s)B^*(s) - 1 = \frac{\Psi_+(s)}{\Psi_-(s)} \tag{1.128}$$

where for $\mathrm{Re}(s) > 0$, $\Psi_+(s)$ must be an analytic function of s that contains no zeros in this half-plane; similarly, for $\mathrm{Re}(s) < D$, $\Psi_-(s)$ must be an analytic function of s and be zero-free (where $D > 0$). In addition, we require for $|s|$ approaching infinity that the behavior of $\Psi_+(s)$ should be $\Psi_+(s) \cong s$ for $\mathrm{Re}(s) > 0$ and that the behavior of $\Psi_-(s)$ should be $\Psi_-(s) \cong -s$ for $\mathrm{Re}(s) < D$. Having accomplished this "spectrum factorization" we may write our solution for $\Phi_+(s)$ as

$$\Phi_+(s) = \frac{K}{\Psi_+(s)} \tag{1.129}$$

where the constant K may be evaluated through

$$K = \lim_{s \to 0} \frac{\Psi_+(s)}{s} \tag{1.130}$$

This constant represents the probability that an arriving customer need not queue. We note that once we have found $\Phi_+(s)$ then we have found the transform for the waiting time PDF, which is what we were seeking.

Although we have described a procedure above for calculating the waiting time pdf, we have not been able to extract the properties of this solution and in fact we have not even given an expression for the average waiting time W in the G/G/1 system. Sad to say, this quantity is, in general, unknown! Its value can be expressed, however, in terms of other system variables as follows. For example, the average waiting time is simply the negative sum of the mean residual life of the random variable $\tilde{u}$ and of $\tilde{y}$ (which is the limiting random variable for the sequence y_n); that is,

$$W = -\frac{\overline{u^2}}{2\bar{u}} - \frac{\overline{y^2}}{2\bar{y}} \tag{1.131}$$

It can be shown that the mean residual life for $\tilde{y}$ is exactly equal to the mean residual life for the random variable I, which denotes the length of an idle period in G/G/1; this last observation coupled with the easy evaluation of the first two moments of the random variable $\tilde{u}$ yields the following expression for the mean wait in G/G/1:

$$W = \frac{\sigma_a^2 + \sigma_b^2 + (\bar{t})^2(1-\rho)^2}{2\bar{t}(1-\rho)} - \frac{\overline{I^2}}{2\bar{I}} \tag{1.132}$$

where σ_a^2 and σ_b^2 are, respectively, the variance of the interarrival time and service time. We shall make use of this last formula in evaluating bounds on the mean waiting time in Chapter 2.

We include no exact results for the G/G/m queue, but refer the reader to the approximations and bounds in Chapter 2. An elegant approach to

the exact analysis of G/G/m has been given by Kiefer and Wolfowitz [KIEF 55] involving the (usually impossible) task of solving an integral equation (which reduces to Lindley's Integral Equation for G/G/1). More recently, de Smit [DESM 73] has extended the theory due to Pollaczek [POLL 61] for G/G/m and has elaborated upon the G/M/m and G/H_R/m queues.

This completes our very rapid summary of the elements of queueing theory. We will need much of this material in the following chapters. It should be clear that a number of important behavioral properties for these queueing systems remain as yet unsolved. Nevertheless we are faced in the real world with applying the tools from queueing theory to solve immediate problems. The balance of this textbook discusses such problems and methods for applying the theory developed. Consequently, we begin with a rather advanced chapter in queueing systems where the goal is *not* to extend the rigorous theory as summarized here but rather to find *effective approximation methods* that permit one to use the theory in a true engineering sense.

REFERENCES

BROC 48 Brockmeyer, E., H. L. Halstrøm, and A. Jensen, "The Life and Works of A. K. Erlang," *Transactions of the Danish Academy of Technology and Science*, **2,** (1948).

DESM 73 de Smit, J. H. A., "Some General Results for Many Server Queues," pp. 153–169 and, "On the Many Server Queue with Exponential Service Times," *Advances in Applied Probability*, **5,** No. 1 (April 1973).

KIEF 55 Kiefer, J., and J. Wolfowitz, "On the Theory of Queues with Many Servers," *Transactions of the American Mathematics Society*, **78,** 1–18 (1955).

KLEI 75 Kleinrock, L., *Queueing Systems, Volume I: Theory*, Wiley-Interscience (New York), 1975.

POLL 61 Pollaczek, F., *Théorie Analytique de Problèms Stochastiques Relatifs à un Groupe de Lignes Teléphonique avec Dispositif d'attente*, Gauthiers-Villars (Paris), 1961.

TELE 70 *Telephone Traffic Theory Tables and Charts, Part 1*, Siemens Altiengesellschaft, Telephone and Switching Division (Munich), 1970.

2

Bounds, Inequalities, and Approximations

An exciting "new" branch of queueing theory is emerging that deals with methods for finding approximate or bounding behaviour for queues.* It is not hard to convince oneself that queueing theory is rather difficult and that exact results are hard to obtain; in fact, *many* of the interesting queueing phenomena have not as yet yielded to exact analysis (and perhaps never will!). Moreover, in those simpler systems where exact results can be obtained, their form is sometimes so complex as to render them ineffectual for practical applications.

If one examines why we study queueing theory in the first place, one readily admits that it is to answer questions regarding real queues in the real world. The mathematical structures we have created in attempting to describe these real situations are merely idealized fictions, and one must not become enamoured with them for their own sake if one is really interested in practical answers. We must face the fact that authentic queueing problems seldom satisfy the assumptions made throughout most of the literature available on queueing theory: stationarity is rare, independence occurs only occasionally, and ergodicity is not only unlikely but is also impossible to establish with measurements over a finite time! Therefore if our mathematical models are so crude, we should be willing to accept much less than an exact solution to the systems of equations they give rise to; rather, we should be happy to accept approximate solutions to these "approximate" mathematical models and hope that such solutions provide information about the behavior of real-world queues. Even more important is the search for "robust" qualitative behavior of queues which provides "rules of thumb" for estimating the

* Perhaps the first approximations used in queueing theory date back to Erlang himself through his introduction of the method of stages (see [KLEI 75], Section 4.2); he tried to approximate the underlying distributions of a queueing system with tractable analytic functions. The reader is referred to [SYSK 62] and to the elucidation of Erlang's work and era in [BROC 48] for some of the historical flavor.