CHAPTER 3

Traffic Models

As verified by numerous measurements, the holding times of ordinary telephone calls are exponentially distributed random variables, with a mean of a few minutes. Service requests from outside the network also have exponentially distributed interarrival times and are represented by Poisson processes. From a theoretical point of view, these features simplify network analysis, which is reduced to a straightforward problem in the theory of Markov processes.

Unfortunately, as is clear from the discussion of Chapter 2, the standard techniques of Markov processes are totally inadequate for the practical analysis of any network of reasonable size operating under a usual routing technique; solutions to the analysis problem can only be heuristic. Nevertheless, the use of probabilistic methods is unavoidable because of the stochastic character of the demand for telecommunication services. In this chapter, we study some simple systems that are amenable to the methods of queuing theory and are of interest in their own right. More importantly for our purpose, these systems form the elements on which all methods of network analysis are based.

There is extensive literature on the subject of traffic models for circuit-switching networks, most contained in the proceedings of the International Teletraffic Congress (ITC). Although the subject of this book is networks, some knowledge of traffic is required, if only because of the influence of traffic on the methods used to analyze and synthesize networks. Thus a short overview of the more widely used traffic models is in order, sufficient to give enough background for the rest of the material. (Excellent sources for the mathematical foundation of teletraffic are [1,2].) A knowledge of elementary stochastic processes and elementary queuing theory is assumed; a standard textbook on the subject, such as [3], should be consulted if this is not the case.

Teletraffic models are queues of the general type $G/M/N/\infty/N$, that is, with a general arrival process, exponential service times, N servers, an infinite source population and no waiting room. The cases of interest are thus determined by the form of G, the arrival process. Because the theory for general arrival is too complex to use in networks, the class of G is limited to special cases. In the first such case, the queue is analyzed as a birth-and-death (BD) process with a given state-dependent arrival rate and exponential service. Another, more recent approach studies arrival processes of the renewal type; this

approach is characterized by the distribution function of the interarrival times, which is sufficient to define the process.

Standard methods of queuing theory can then be used to compute the distribution of busy servers in the group — in particular, the blocking probability. Because of the presence of alternate routing, this analysis is generally not sufficient, and other quantities are needed to analyze circuit-switched networks.

In the simplest case, the so-called *responsive system* case, the number of servers is infinite. Even if the analysis for finite N turns out to be too difficult or cumbersome to use, the results available for the infinite case often can be used instead. This is related to the notion of *offered traffic*, another way to characterize the arrival process that is often used in studying of circuit switching. For a $G/M/N/\infty/N$ queue, the offered traffic is defined as follows. First, imagine a parallel system of infinite capacity associated with the real system of size N. Whenever a call arrives at the real system, another call is set up in the parallel system. The actual holding time for the parallel call is drawn from an exponential distribution with the same mean as for the real calls, but independently of the actual holding time of the real calls. The offered-traffic process is the process whose state variable is the number of calls in progress on the parallel group at time t; the offered traffic is defined as the mean of this process. The calculation of the moments of the offered-traffic distribution plays an important role in network analysis, the reason why so much attention is given to the responsive system in forthcoming sections.

Because of the routing algorithms generally used in circuit-switched networks, customers blocked on one system are frequently routed to another system, where they attempt to be served. This process, called *overflow*, is a feature of virtually all circuit-switched networks. Because these blocked calls are being sent to some other part of the network, they also contribute to the arrival process of some other groups. Thus, in addition to the analysis of the $G/M/N/\infty/N$ queue, a great deal of attention must be paid to the *overflow* process, that is, the process generated by calls that cannot be accepted on the system.

Similarly, customers that can be served on a group often must attempt to obtain service in additional groups before their service period begins — for instance, whenever a call must traverse two or more links in tandem in order to reach its destination. This situation also contributes to the arrival process of other links, posing some additional difficulties that are discussed shortly. Unfortunately, the theory for carried calls in tandem is not nearly as developed as for the corresponding case of overflow.

The chapter is organized as follows. First, we examine the simplest case for $G = M$, that is, the Poisson arrival. The behavior of this system is well known and can be derived from elementary queuing theory. A detailed discussion of the overflow process is also presented, since overflow is better known and is the basis for many heuristics.

Then, because of the alternate routing of blocked calls and the multilink routing of accepted calls, we consider the case of non-Poisson arrivals. Needless to say, this theory is much more difficult, and the analytic results much more complex. Two classes of queues are considered, depending on the nature of their arrival processes. For state-dependent arrival rates, we use the methods of birth-and-death processes, while, for an arrival process of the general renewal type, we can rely on the known results of the $GI/M/N/\infty/N$ queue. In some cases, these processes can be characterized exactly in terms of the moments of the offered traffic. We compute the congestion probability, and in some cases the parameters of the overflow and carried processes. Finally, we introduce the notion of conserved flows, discussing in depth their relation to the queuing models presented previously and their possible use in network analysis.

This leads to the second part of the chapter, where we explain how the models based on queuing theory are used in actual problems for non-Poisson arrivals. Particular emphasis is placed on the characterization of processes by the offered traffic and on moment-matching techniques, which are the most frequently used network analysis methods. We see under what conditions the queuing models can be used to represent a traffic characterized by a small number of moments. We then consider the possibility of inverting nonlinear systems and the computational efficiency of these techniques.

Finally, we look at the traffic problems arising from the fact that the traffic offered to a link is a superposition of various streams, each having its own history. Models for the calculation of individual stream blocking are reviewed, with an emphasis on the simplifying assumptions required to make these models work. In addition, practical methods usable in large networks are described.

The following symbols are introduced in this chapter:

N = The number of servers in the group.

p_n = The stationary probability of being in state n.

$\overline{n}$ = The average value of n. When the group is finite, this is called the *carried* traffic and is denoted by $\overline{M}$. When the calls are offered to an infinite group, this is called the *offered traffic* and is denoted by M. M can also be defined as the expected number of arrivals during an average holding time.

E = The time congestion, simply $P(N)$. This is the fraction of the time when all the servers are busy. We say that the system is in the *blocked* state, and calls finding the system in that state are said to be blocked. From the point of view of traffic theory, calls that are blocked are lost and do not return. In practice, in a network, blocked calls may be sent to other groups, where they appear as requests for service.

B = The call congestion. This is the probability that an arriving call will find the system in state N. Note that, for a general arrival process, $E \neq B$.

M, V, Z = Mean, variance, and peakedness of distributions, respectively. The peakedness is defined as $Z = V/M$. Processes are said to be *peaked* or *smooth* according to whether $Z > 1$ or $Z < 1$, respectively. The M, V notation is used to emphasize the non-Poisson character of the traffic. Unless otherwise noted, a Poisson traffic is denoted by the symbol A.

α_q, μ_q, M_q, β_q = The q^{th} moment, central moment, factorial moment, and binomial moment of a distribution p, respectively.

Several other conventions are also used when representing traffic. Quantities dealing with overflow are of the form $\hat{x}$, while the corresponding quantities for carried traffic are of the form $\overline{x}$. It is sometimes necessary to distinguish a distribution at an arbitrary time from the corresponding distribution seen by a call arriving in the system. Variables have an asterisk in the first case, and none in the second. For instance, β_q^* denotes the q^{th} binomial moment of the busy-circuit distribution at an arbitrary instant of time, and $\overline{M}_1$ denotes the first factorial moment of the distribution of busy circuits at the time of a call arrival.

3.1 Poisson Arrivals

The first system we analyze consists of a group of N servers offered calls according to a Poisson process. Many of the results discussed here are well known, although some are limited to the area of teletraffic. Our main purpose is to illustrate how the standard questions concerning traffic models have been answered in the simplest case, also showing the limitation of these techniques. First, we characterize the process, in terms of both its arrival rate and the interarrival time distribution, and introduce the notion of Poisson traffic. Next, we compute the congestion functions, characterizing the overflow and carried processes in terms of the distribution of busy circuits in a group of infinite size.

Characterization of the Poisson System

The Poisson system is a system where service requests arrive according to a Poisson process and where the service times have a negative exponential distribution. The Poisson process has some interesting properties, some of which we now recall.

This process is a pure birth process with an arrival rate λ independent of the system state, an initial state labeled 0 and a final, absorbing state labeled

N which may be at infinity. We compute the probability of having k customers in the system at time t when there is no finite absorbing state by the standard equation (A.14); we get

$$\frac{dp_k(t)}{dt} = -\lambda p_k(t) + \lambda p_{k-1}(t), \quad k \geq 1$$
$$\frac{dp_0(t)}{dt} = \lambda p_0(t)$$

These equations can be solved recursively, yielding

$$p_j(t) = e^{-\lambda t}\frac{(\lambda t)^j}{j!}. \tag{3.1}$$

We can readily compute the generating function

$$G(z) = \sum_{k=0}^{\infty} p_k(t) z^k$$
$$= e^{\lambda t(z-1)}.$$

If we let $N(t)$ represent the number of customers in the system at time t, we can compute the mean $\overline{N}(t)$ and the variance $\sigma^2(t)$ of $N(t)$ from $G(z)$ as

$$\overline{N}(t) = \lambda t$$
$$\sigma^2(t) = \lambda t$$

which is an important property of Poisson processes.

The Poisson process has other interesting properties; one is the fact that the interval between arrivals has a negative exponential distribution. Let τ be the random variable that represents the time between arrivals, and $A(t)$ and $a(t)$ its distribution and density functions. Since $A(t)$ is the distribution of interarrival times, we have

$$A(t) = 1 - Pr(\tau > t).$$

$Pr(\tau > t)$ is the probability that there is no arrival before t, and is given by

$$Pr(\tau > t) = 1 - p_0(t).$$

Replacing in Eq. (3.1) for $j = 0$, we have

$$A(t) = 1 - e^{-\lambda t}, \quad t \geq 0$$

and the density

$$a(t) = \lambda e^{-\lambda t}.$$

Thus we see that the specification of an arrival rate fixes the interarrival time distribution of the Poisson process — in the present case, the negative exponential.

Let us now return to the Poisson system, which we have defined as a birth-and-death process with Poisson arrivals and independent negative exponential holding times. It is defined mathematically by taking $\lambda_k = \lambda$ — that is, as a state-independent arrival rate — and a linear departure process with $\mu_k = k\mu$. In terms of queuing theory, the system is described by a $M/M/\infty/\infty$ queue. We can solve Eq. (A.14) for this case and obtain the stationary probability of having j customers in the system:

$$p_j = e^{-A}\frac{A^j}{j!}$$

$$A \triangleq \frac{\lambda}{\mu}.$$

From this, we get the first two moments of the busy-circuit distribution:

$$M = V = A$$

Recall that the offered traffic corresponding to an arrival process is characterized by the moments of the busy-server distribution in an infinite system. We say that an offered traffic is a *Poisson traffic* whenever its mean is equal to its variance. We see from the preceding results that a Poisson traffic is produced when the arrival process is a Poisson process (and of course when the holding times are exponentially distributed). The peakedness $Z = 1$ is an important feature of the Poisson traffic that severely limits the use of this traffic as a model for call arrivals, since this feature applies only to an arrival process whose mean is equal to its variance.

Congestion Functions

A group of size $N < \infty$ is modeled as an $M/M/N/N$ queue. The distribution of busy servers can be found by noting that the BD equation has the same form as in the infinite case, but with the added condition that $p_k = 0$ whenever $k > N$. The state probabilities still have the form $p_k = p_0 A^k/k!$, but the normalization condition is different since it is only a partial sum from 0 to N. Thus the state probability distribution is given by the truncated Poisson distribution

$$p_k = \frac{A^k/k!}{\sum_{i=0}^{N} A^i/i!} = \frac{A^k/k!}{e_N(A)}, \tag{3.2}$$

where the function $e_N(A) = \sum_{j=0}^{N} A^j/j!$ is called the *incomplete exponential function*. The probability that all the servers are busy, given by the well-known Erlang B function,

$$E(A,N) = \frac{A^N/N!}{e_N(A)}, \tag{3.3}$$

is the time congestion of the system. In the present case, this probability is also the call congestion because of the exponential distribution of arrivals, which is memoryless. A useful relation is

$$\frac{1}{E(A,N)} = 1 + \left(\frac{N}{A}\right)\frac{1}{E(A,N-1)}, \tag{3.4}$$

which can be used for the evaluation of $E(A,N)$ over integers, and which is numerically stable. Although the recursive formula (3.4) is quite sufficient for small integer values, it becomes too time consuming when a large group must be evaluated; it probably should not be used as is within network-design algorithms if numerical efficiency is important. The question of efficient algorithms for the numerical evaluation of the Erlang B function is somewhat complex, and considering it would lead us far outside the scope of this discussion. We therefore refer the interested reader to the methods described in [4].

As we see later, it is often necessary to generalize the Erlang B function to fractional values, especially when dimensioning networks. The function can be extended to the complex half-plane by

$$E(A,N)^{-1} = A\int_0^\infty e^{-Ay}(1+y)^N dy, \quad \text{Re } A > 0 \tag{3.5}$$

The mathematical properties of this function and its derivatives have been studied extensively by Jagerman [5].

Overflow Process

We have characterized the Poisson process and computed its congestion functions. Remember, however, that blocked calls generally appear as requests for service somewhere else in the network. For this reason, we must characterize this overflow process. We do this via the distribution of busy circuits in the infinite group, which corresponds to the definition of offered traffic.

Following Kosten [6], we use the BD approach. The calls that are blocked in the primary group are offered to a secondary group of infinite size; we are interested in the nature of this arrival process. Although we wish to consider the distribution in the secondary group, the number of customers in this system alone is not sufficient to define a state of the Markov chain. Such a state can be defined only if we consider *both* the primary and secondary groups at the same time, which leads us to the study of a two-dimensional process — a fundamental limitation of the BD approach in the study of overflow. Nevertheless, because of the relative simplicity of the system, it is possible to solve the Kolmogorov equations exactly for the combined primary and secondary group, from which various quantities of interest can be derived. Because of their fundamental importance in practical situations, these equations are derived following a method proposed by Riordan as described in [6].

The joint probability distribution $p_{n,m}$ of having n busy servers in a primary group of size N, and m in a secondary group of infinite size is found by writing the equilibrium equations with $A = \lambda/\mu$:

$$(A + n + m)p_{n,m} - (n + 1)p_{n+1,m} - (m + 1)p_{n,m+1} - Ap_{n-1,m} = 0 \tag{3.6}$$

$$(A + N + m)p_{N,m} - Ap_{N,m-1} - (m + 1)p_{N,m+1} - Ap_{N-1,m} = 0 \tag{3.7}$$

Once again, these equations can be written by inspecting the state-transition diagram to and from state (n, m). Because we are interested in the distribution of customers in the secondary system, we concentrate on the calculation of the factorial moments of this distribution, given that there are n customers in the primary group. For these moments, and for fixed n, we define the generating function

$$M(n, z) = \sum_{k=0}^{\infty} M_k(n)\frac{z^k}{k!}$$

$$= \sum_{k=0}^{\infty}(1 + z)^k p_{n,k}$$

This last relation is replaced in Eqs. (3.6) and (3.7), which then become differential equations for $M(n, z)$:

$$(A + n + \frac{d}{dz})M(n, z) - (n + 1)M(n + 1, z) - AM(n - 1, z) = 0$$

$$(N - Az + z\frac{d}{dz})M(N, z) - AM(N - 1, z) = 0$$

Replacing with the definition of the generating function, and equating equal powers of z, we obtain a recurrence relation for the factorial moments:

$$(A + n + k)M_k(n) - (n + 1)M_k(n + 1) - AM_k(n - 1) = 0 \tag{3.8}$$

$$(N + k)M_k(N) - AkM_{k-1}(N) - AM_k(N - 1) = 0 \tag{3.9}$$

Equations (3.8) and (3.9) form a recurrence on n and can be solved by introducing another generating function:

$$F_j(y) \triangleq \sum_{n=0}^{\infty} M_j(n)y^n.$$

Using Eq. (3.8), we once again obtain a differential equation for F,

$$\left[A + k - Ay + (y - 1)\frac{d}{dy}\right] F_k(y) = 0,$$

which can be integrated to yield

$$F_k(y) = Ce^{Ay}(1-y)^{-k}, \tag{3.10}$$

where C is an integration constant given by the condition

$$C = M_k(0).$$

Having the expression for $F_k(y)$, we can expand the right-hand side of Eq. (3.10) to get the moments directly:

$$M_k(n) = M_k(0)S(n,k), \tag{3.11}$$

where $S(n,k)$ is the Brockmeyer polynomial (see [7]). The moments are uniquely determined from Eq. (3.11) up to the values $M_k(0)$; they can be calculated from Eq. (3.9) and the normalizing condition $\sum_{k=0}^{N} M_k(0) = 1$. We have

$$[(N+k)S(N,k) - AS(N-1,k)]\, M_k(0) = AkS(N,k-1)M_{k-1}(0),$$

and using the recurrence relations

$$\begin{aligned}(N+k)S(N,k) - AS(N-1,k) &= (N+k-A)S(N,k) + A\,[S(N,k) - S(N-1,k)] \\ &= (N+k-A)S(N,k) + AS(N,k-1) \\ &= kS(N,k+1)\end{aligned}$$

we obtain finally

$$\begin{aligned}M_k(0) &= A\frac{S(N,k-1)}{S(N,k+1)}M_{k-1}(0) \\ &= A^k\frac{S(N,1)S(N,0)}{S(N,k+1)S(N,k)}M_0(0)\end{aligned}$$

After using the normalization condition, we obtain

$$M_k(n) = A^k\frac{S(N,0)S(n,k)}{S(N,k+1)S(N,k)}$$

and the moments of the distribution in the secondary group:

$$M_k = A^k\frac{S(N,0)}{S(N,k)}.$$

From the moments, we immediately obtain the joint state probability

$$p_{n,m} = \sum_{j=m}^{\infty}(-1)^{j-m}\binom{j}{m}\frac{A^j}{j!}\frac{R(N,0)R(n,j)}{R(N,j+1)R(N,j)}.$$

The marginal distribution in the primary group is the Erlang B distribution, while the marginal distribution in the secondary group is given by

$$p_m = \sum_{j=m}^{\infty}(-1)^{j-m}\binom{j}{m}\frac{A^j}{j!}\frac{R(N,0)}{R(N,j)}. \tag{3.12}$$

From this distribution, all the moments of the traffic can be computed. Of particular interest are the mean M and the variance V of the traffic carried in the secondary group, which are the required characterization of the overflow process and are of fundamental importance in many applications. They are given by

$$M = AE(A, N) \tag{3.13}$$

$$V = M\left(1 - M + \frac{A}{(N + 1 - A + M)}\right) \tag{3.14}$$

Although not obvious from Eq. (3.14), the peakedness of the overflow traffic is always greater than one — an important point for the study of overflow systems with peaked offered traffic. From this follows the conclusion that the overflow process is not Poisson.

Carried Processes for Poisson Arrivals

A complete study of Poisson arrivals requires knowledge of the carried process, of which we have a complete description by the probability distribution (3.2). From this, we obtain the generating function and the first two moments of the carried traffic:

$$G(z) = \frac{e_N(Az)}{e_N(A)}$$

$$\overline{M} = A(1 - E(A, N)) \tag{3.15}$$

$$\overline{V} = \overline{M} - A(N - \overline{M})E_N \tag{3.16}$$

The derivation is left as an exercise (Problem 3.1). Note that, for the carried traffic, we always have $\overline{Z} < 1$, another indication that this process is not Poisson either.

3.2 Other State-Dependent Arrival Processes

Since traffic flows inside a network are not Poisson in general, let us now consider other arrival processes. First, we review some frequently used models that can be analyzed with the two-dimensional BD methods that have state-dependent arrival rates. We characterize the arrival process and congestion functions, but in most cases do not characterize the overflow and the carried process explicitly.

Overflow Arrivals

As we have emphasized, traffic that overflows from a link is not lost, but is often offered to some other link in a network. It seems reasonable, then, to examine

in the first place a link where the arrival process is precisely the overflow from a group of size N being offered Poisson traffic. As we already saw, analysis of the Kosten system gives the moments of the distribution of blocked calls in a group of infinite size, that is, characterizes the offered traffic when it is produced by overflow. We wish to analyze *this* traffic when it is offered to a group of size L, that is, its congestion function and the characterization of *its* overflow.

This amounts to a Kosten-like system where the secondary group is now of finite size L. In this case, we are interested in the busy-circuit distribution in the secondary group, as well as the distribution of the traffic that overflows from it onto a group of infinite size.

This system was analyzed by Brockmeyer [6,7], who gives a complete solution for the joint state probability in the primary and secondary groups. The solution technique is virtually identical to the one for the Kosten system, with some minor modifications of the state equation to take into account the finite size of the secondary group. For this reason, we do not go into a detailed derivation here, only quoting the main results. The actual derivation is left as an exercise (see Problem 3.5). The joint probability of having n busy servers in the primary and l in the secondary group is given by

$$p_{n,l} = \sum_{j=0}^{L-l} (-1)^j \binom{l+j}{l} S(n-j, l+j) p_{0,l+j} \tag{3.17}$$

$$p_{0,l} = \frac{1}{S(N+L,1)} \sum_{k=l}^{L} (-1)^{k-l} \binom{k-1}{l-1} \frac{1}{S(N,k)} \sum_{l=k}^{L} \binom{l-1}{k-1} S(N+l,0)$$

$$p_{0,0} = \frac{1}{S(N+L,1)}$$

The marginal state distribution for the secondary system is given by

$$p_l = \sum_{j=0}^{L-l} (-1)^j \binom{l+j}{l} S(N-j, l+1+j) p_{0,l+j}, \tag{3.18}$$

from which the mean and variance of the traffic carried in a finite group offered calls produced by an overflow source are given by

$$\overline{M} = A(E_N - E_{N+L}) \tag{3.19}$$

$$\overline{V} = A\left[\frac{A(E_N - E_{N+L})}{N+1-A(1-E_N)} - LE_{N+L}\right] + \overline{M}(\overline{M}-1) \tag{3.20}$$

Note that these formulas reduce to Eqs. (3.13) and (3.14) of the Kosten system whenever $L \to \infty$. In this case, it is not obvious whether the traffic carried in the secondary group is smooth or peaked. The time and call congestions for

the secondary group are given by

$$E = E_{N+L}\frac{S(N, L+1)}{S(N, L)} \tag{3.21}$$

$$B = \frac{E_{N+L}}{E_N} \tag{3.22}$$

As for the characterization of the traffic blocked on the secondary system, note that the distribution of the overflow is nothing but the distribution of overflow from a Kosten system of size $N + L$, and as such poses no difficulty. One gets

$$\hat{M} = AE(A, N+L) \tag{3.23}$$

$$\hat{V} = \hat{M}\left(1 - \hat{M} + \frac{A}{N + L + 1 - A(1 - E(A, N+L))}\right) \tag{3.24}$$

As expected, the second overflow from the primary group is also peaked, and cannot be represented by a Poisson process. This is about as far as we can go in the sequence of overflows. Calculating the overflow of this traffic would require a three-dimensional BD process, which would probably be of little practical use because of its complexity. Instead, let us now analyze other arrival processes, the usefulness of which will become apparent when we consider approximate methods in Section 3.4.

Linear Arrival Rates

Consider an arrival process defined by a linear arrival rate:

$$\lambda_k = [\alpha + (k - \alpha)\beta]$$

$$\mu_k = k\mu$$

The usefulness of this process was pointed out first by Wilkinson [6], who noted that the distribution of busy circuits in a lightly loaded group receiving overflow traffic was closely approximated by a Pascal distribution, also known as a *negative binomial distribution*. We now show that the linear arrival rate does yield such a distribution and, as such, might be used to model the arrival process corresponding to overflow traffic. Replacing in Eq. (A.16), we get, after some simple algebraic manipulations,

$$p_k = p_0\left(\frac{\beta}{\mu}\right)^k \frac{1}{k!}a(a+1)\ldots(a+k-1), \tag{3.25}$$

where

$$a \triangleq \frac{\alpha}{\beta}(1-\beta).$$

The value of the normalization constant is easily obtained by the use of Eq. (A.3), yielding

$$p_k = \frac{1}{\left(1 - \frac{\beta}{\mu}\right)^{-a}} \left(\frac{\beta}{\mu}\right)^k \frac{a(a+1)\dots(a+k-1)}{k!}. \tag{3.26}$$

This distribution can be identified as the Pascal distribution by noting that

$$\frac{x(x+1)\dots(x+k-1)}{k!} = \binom{x+k-1}{k}.$$

The Pascal distribution can be obtained for integer x as the probability distribution for the waiting time in a sequence of Bernouilli trials with success probability p. It is written as

$$p_k = \binom{n+k-1}{k} p^n q^k, \quad p+q=1, \quad 0<p<1$$

and has the first two moments

$$E(x) = \frac{nq}{p}, \quad var(x) = \frac{nq}{p^2}$$

If we identify $q = \beta/\mu$, $a = n$, and $p = 1 - \beta/\mu$, and set $\mu = 1$, the first two moments are expressed simply as

$$M = \alpha \tag{3.27}$$

$$V = \frac{\alpha}{1-\beta} \tag{3.28}$$

As seen from these expressions, we must have $\alpha \geq 0$ and $0 < \beta < 1$ when $\mu = 1$ to ensure that the parameters do not become negative. We also see that the Pascal distribution can be used to represent peaked traffic only, but that the value of peakedness can be arbitrary as long as it is larger than one. A final advantage of the Pascal distribution is that it is very simple to compute the distribution parameters from the moments, that is, to invert Eqs. (3.27) and (3.28). This fact is of considerable importance for the use of the Pascal model in approximate methods.

Given that the negative binomial distribution can model an arbitrary peaked process (at least up to the first two moments), we might ask whether a similar process could be used to represent a smooth traffic, that is, one having $z \leq 1$. Since the negative binomial distribution is related to peaked arrivals, we might suspect that the ordinary binomial distribution is what we need for smooth processes. The answer to this question is left as an exercise (Problem 3.4), where it is shown that the standard binomial distribution can indeed be used for smooth traffic, but with some restrictions. Within these restrictions, the binomial distributions, both negative and standard, provide a

unified framework with which to model non-Poisson traffic, the main reason these distributions have been so popular in network analysis.

In principle, the overflow process could be calculated in the same way as for the Kosten system. The only difference would be to replace the original Poisson arrival rate by the linear rate and to solve the Kolmogorov equations. This seems not to have been done, however; the Pascal distribution has been used mostly in the area of approximations.

Balking Systems

An implicit assumption in the BD analysis of queues is that a customer arriving in the queue and finding a free server will immediately enter service. These systems are called *full availability* for obvious reasons. In some cases, however, this assumption is not really justified: A customer may find a free server but not enter service, and is lost. In this case, we say that the system has *limited availability*, and that there is a nonzero probability of *balking*. Such systems occur quite frequently in networks; the simplest example is the case of two links in series where a call must find a free server on both links in order to begin service. If there is a free server on the first link and none on the second, an analysis of the first link taken in isolation would describe this link as a limited-availability system in order to take into account the presence of the second link.

In general, the balking probability is state dependent, and analyzing it is quite complex since the probability depends on the particular form of this dependence. One simple case, however, offers some insight into the operation of limited availability systems: where the balking probability β is independent of the state of the system. Assuming a Poisson arrival of rate λ, the BD model can be used with an arrival rate

$$\lambda_k = \begin{cases} \lambda(1-\beta) & \text{if } k \leq N \\ 0 & \text{otherwise} \end{cases}$$

$$\mu_k = \begin{cases} k\mu & \text{if } k \leq N \\ 0 & \text{otherwise} \end{cases}$$

The analysis proceeds exactly as for the Poisson case. The state probability is

$$p_k = p_0 \frac{[A(1-\beta)]^k}{k!},$$

where as usual $A = \lambda/\mu$. In other words, we find the truncated Poisson distribution, but with a reduced arrival rate $A(1-\beta)$, that is, with an offered traffic that is reduced by the balking probability. Returning to the simple case of two links in series, we can assume that $\beta = B_2$, where B_2 is the blocking probability on the second link. This is only an approximation, of course, since the balking probability is state dependent because of the correlation induced between

the links by the common traffic. Nevertheless, if we make the independence assumption, then B_1, the blocking probability of the first link, becomes

$$B_1 = E[A(1 - B_2), N_1],$$

which is the apparently paradoxical result that the blocking probability and the traffic offered to the first link depend on the blocking probability on the second link. If we also note that the traffic offered to the second link depends on the blocking on the first link, we are faced with our first encounter with a recurrent problem in the analysis of networks. The two-link system, even with the independence assumption, is described in terms of a set of two nonlinear equations, and knowledge of the blocking probability requires this system to be solved. This topic is discussed in depth in Chapter 4, where numerous methods for the analysis of complex systems are proposed.

The apparent paradox resides in the picture of the offered traffic as a flow that is thinned sequentially as it passes from one link to the next. This situation is only a result of the decomposition of the path into two presumably independent components and of the attempt to take into account the presence of the second component when analyzing the first one. The real analysis should be done on the complete path using a three-dimensional birth-and-death process, which leads us outside the scope of this chapter. We consider the question in Chapter 4, where we look at decomposition methods for network analysis.

Throttled Arrivals

We now present *throttled arrivals*, another important BD process that is being increasingly used in networks [8]. In this system, two independent streams of Poisson traffic, say type 1 and type 2, are offered to a single group of size N. The first stream, say stream 1, can use any free server, while stream 2 can use a server only when the total number of busy servers is less than some fixed number $m \leq N$. The threshold m is called the *protection level*. The arrival rate is given by

$$\lambda_k = \begin{cases} \lambda_1 + \lambda_2 & \text{if } k < m \\ \lambda_1 & \text{otherwise} \end{cases}$$

$$\mu_k = k\mu.$$

The state probabilities, found by Eq. (A.15), are given by

$$p_k = \begin{cases} p_0 \left(\dfrac{\lambda_1 + \lambda_2}{\mu}\right)^k \dfrac{1}{k!} & \text{if } k < m \\ p_0 \dfrac{1}{n!} \dfrac{1}{(n+1)(n+2)\ldots(k)} \left(\dfrac{\lambda_1 + \lambda_2}{\mu}\right)^n \left(\dfrac{\lambda_1}{\mu}\right)^{k-n} & \text{otherwise} \end{cases}$$

In this system, p_0 is given by the ordinary normalization condition. Given the state probabilities, it is possible to compute all the moments of the traffic, in

particular of the variance. Because the end result is quite complex, we do not pursue the computation of moments.

Note also that the throttled-arrivals model has two blocking probabilities of interest:

B = The probability that all of the n trunks are occupied.

B' = The probability that more than $m-1$ trunks are busy.

These probabilities, which determine how much of the two streams will be blocked, are used in many network analysis algorithms. We have, for integer m and N [8],

$$B = p_0 \frac{a^N}{N!}(1-r)^{N-m} \tag{3.29}$$

$$1 - B' = p_0 \sum_{j=0}^{m-1} \frac{a^j}{j!} \tag{3.30}$$

$$p_0^{-1} = \sum_{k=0}^{m} \frac{a^k}{k!} + \sum_{k=m+1}^{N} \frac{a^k}{k!}(1-r)^{k-m} \tag{3.31}$$

where

a = Mean of total traffic offered to the group, $a = (\lambda_1 + \lambda_2)/\mu$.

r = Ratio of type two to total traffic, $r = \lambda_2/a\mu$.

m = Protection level.

This model can be used whenever we want to offer better service to one type of call, here type 1, by restricting the availability of servers to the other type of calls. This is particularly important as a way to stabilize networks operating with nonhierarchical routing.

The Interrupted Poisson Process

The interrupted Poisson process (IPP) has been used to generate non-Poisson traffic in simulations and also to evaluate network models. The calls are produced by a Poisson source of intensity λ independent of the system state. This source, however, is randomly interrupted for an interval with negative exponential distribution, and then turned on for an interval also exponentially distributed. Let

λ = The intensity of the Poisson source.

$1/\mu$ = The mean service time of the customers.

$1/\gamma$ = The mean on-time of the source.

$1/\omega$ = The mean off-time of the source.

k = The number of customers in the system.

m = The state of the switch; $m = 1$ indicates that the switch is on, while $m = 0$ indicates that it is off.

$p_{k,m}$ = The stationary probability of being in state (k, m).

We are interested in the distribution of busy servers in systems of infinite capacity, that is, in characterizing the offered traffic that can be produced by such a system.

The generating function of the occupancy distribution has been computed in [9] based on a two-dimensional Markov chain. The equilibrium equations can be written by examining the transition rates from and to state (k, m). They are written separately for the two possible values of m:

$$(k\mu + \omega)p_{k,0} = \gamma p_{k,1} + (k+1)\mu p_{k+1,0}, \quad k \geq 0 \tag{3.32}$$

$$(k\mu + \gamma + \lambda)p_{k,1} = \omega p_{k,0} + (k+1)\mu p_{k+1,1} + \lambda p_{k-1,1}, \quad k \geq 1 \tag{3.33}$$

$$(\gamma + \lambda)p_{0,1} = \omega p_{0,0} + \mu p_{1,1} \tag{3.34}$$

Equation (3.32) is obtained by noting that, if the switch is off ($m = 0$), the state $(k, 0)$ can be entered in only one of two ways: Either a call terminates from state $(k+1, 0)$ (term $(k+1)\mu p_{k+1,0}$) or the switch goes from the on to the off state with no change in the number of customers (term $\gamma p_{k,1}$). Similarly, from the state $(k, 0)$, only two other states can be entered: A call terminates from state $(k, 0)$ (term $k\mu p_{k,0}$) or the switch goes from the off to the on state with no change in the number of customer (term $\omega p_{k,0}$). The equilibrium between these flow rates out of and into state $(k, 0)$ yields Eq. (3.32). A similar argument permits the other equations to be constructed straightforwardly. The precise construction of the transition diagrams is left as an exercise (Problem 3.6).

The distribution of busy circuits is calculated by the generating function method. Let

$$G(z) = \sum_{k=0}^{\infty} p_k z^k$$

where

$$= G_1(z) + G_2(z),$$

and

$$G_i(z) \triangleq \sum_k p_{k,i} z^k, \quad i = 0, 1$$

$$p_k \triangleq p_{k,0} + p_{k,1}.$$

From this, it is obvious that $G_1(1) = \sum_k p_{k,1}$ is the probability that the switch is on, and is simply $\omega/(\gamma + \omega)$. Similarly, $G_0(1)$ is the probability that the switch is off, and is given by $\gamma/(\gamma + \omega)$. The solution technique consists of writing a differential equation for the generating function, from which one can obtain the factorial moments of the steady state distribution, as well as the probabilities themselves. Using the well-known properties of generating functions, $kp_k \longleftrightarrow zG'(z)$ and $p_{k-j} \longleftrightarrow z^j G(z)$, where the symbol $\longleftrightarrow$ means "is the transform of," we get from Eqs. (3.32) and (3.33) the coupled differential equations

$$\mu(z-1)G_0'(z) + \omega G_0(z) - \gamma G_1(z) = 0 \tag{3.35}$$

$$\mu(z-1)G_1'(z) + (\gamma + \lambda - \lambda z)G_1(z) - \omega G_0(z) = 0 \tag{3.36}$$

These equations can be uncoupled by a second derivation and substitution, yielding

$$\mu(z-1)G_0''(z) + [\mu + \gamma + \omega - \lambda(z-1)]\,G_0'(z) - \frac{\lambda}{\mu}\omega G_0(z) = 0 \tag{3.37}$$

$$\mu(z-1)G_1''(z) + [\mu + \gamma + \omega - \lambda(z-1)]\,G_1'(z)\frac{\lambda}{\mu}(\gamma + \omega)G_1(z) = 0 \tag{3.38}$$

Changing variables $x = (z-1)(\lambda/\mu)$, we get

$$xF_0''(x) + (a-x)F_0'(x) - bF_0(x) = 0$$

$$xF_1''(x) + (a-x)F_1'(x) - (1-b)F_1(x) = 0$$

where

$$a = 1 + \frac{\gamma + \omega}{\mu}, \qquad b = \frac{\omega}{\mu}.$$

The solution to these equations is given in terms of the confluent hypergeometric function. We get

$$G(z) = \frac{\gamma}{\gamma + \omega}\,{}_1F_1\left[b; a; \frac{\lambda}{\mu}(z-1)\right] + \frac{\omega}{\gamma + \omega}\,{}_1F_1\left[1 + b; a; \frac{\lambda}{\mu}(z-1)\right],$$

where

$${}_1F_1(x; y; z) \triangleq \sum_{k=0}^{\infty} \frac{x(x+1)\ldots(x+k-1)z^k}{y(y+1)\ldots(y+k-1)k!}, \tag{3.39}$$

where the integration constants are found from the two conditions $G_1(1) = \omega/(\omega + \gamma)$ and $G_0(1) = \gamma/(\omega + \gamma)$. Once we have the generating function, we can derive all quantities of interest from it. In particular, the factorial moments and the probabilities can be read directly from the definition of the hypergeometric functions (3.39); after setting $\mu = 1$, we get

$$M_q = \lambda^q \frac{\omega(\omega-1)\ldots(\omega+q-1)}{(\gamma+\omega)(\gamma+\omega+1)\ldots(\gamma+\omega+q-1)},$$

which satisfies the useful recurrence relation

$$M_{q+1} = \lambda \frac{\omega + q}{(\gamma + \omega + q)} M_q,$$

from which we can write the distribution

$$p_m = \frac{\lambda^m}{m!} \sum_{j=m}^{\infty} \frac{(-\lambda)^{j-m}}{(j-m)!} \frac{\omega(\omega+1)\dots(\omega+j-1)}{\gamma + \omega(\gamma+\omega+1)\dots(\gamma+\omega+j-1)}.$$

The mean and the variance of this distribution are given by

$$M = A \frac{\omega}{\omega + \gamma} \tag{3.40}$$

$$V = M \left[\frac{A(\omega+1)}{(\omega+\gamma+1)} - M + 1 \right] \tag{3.41}$$

which characterize the traffic produced by an IPP generator. It is not hard to show that this traffic is always peaked if $\gamma > 0$.

Following the usual approach, we would investigate the character of the overflow and the carried traffic in a system of finite size. No formulas specifically designed for the IPP case, however, seem to exist. Instead, the parameters of such traffic are computed using the general $GI/M/N$ theory (see Section 3.3).

Most of the results obtained so far utilize techniques based on BD processes and state-dependent arrival rates. Before ending this discussion, we should mention some remarkable results due to Wallström [2], who gives a complete characterization of the offered traffic for an arrival process with an *arbitrary* state-dependent coefficient in terms of all the binomial moments of the distribution. Two classes of traffic have been examined: those where the dependence is on the total number of calls in the system, and those where the dependence is on the number of calls in the primary group only. Although these results constitute a *tour de force*, their application to network problems is somewhat limited because of the very heavy computational requirements implied by the technique. Also, the results cannot be used recursively to compute the overflow cascade since the dimension of the BD process increases with the number of overflow stages. For this reason, we believe that a full demonstration of these results, which is quite involved, would be of limited use in the present context; we refer the interested reader to reference [2].

3.3 Renewal Arrival Process

The solution of the single-overflow system with Poisson offered traffic by the BD methods, given by Kosten [10] and Brockmeyer [7], requires the use of a two-dimensional BD process. This is because the secondary group cannot be characterized only by the coefficients of the arrival process, but also depend on the state of the primary group. This fundamental factor limits the use of the BD approach in traffic models in the presence of overflow. Recall that

many routing techniques allow *many* overflows for a blocked call. The exact analysis of a system of k groups operating in such a cascade requires a k-dimensional BD process, which is unmanageable for values as small as $k = 3$. The solution to this problem of cascade overflow, which has received a great deal of attention over the years, requires a new approach to teletraffic, one based on the theory of renewal stochastic processes. The results obtained this way are theoretically important because they give a complete solution to an arbitrary cascade of overflows, provided the first arrival process is a renewal process. These results, although not easily used in network calculations as such, provide a new collection of traffic models that are used in approximate methods such as those discussed in Section 3.4.

For these reasons, we now present results for the case when the traffic offered to a group of N servers is described by an arbitrary renewal process, with an exponential service time of mean $1/\mu$. A short derivation of the basic results from renewal theory can be found in Section A.3.

Congestion Functions and Carried Traffic

First, let us consider the distribution of busy servers in the group. Because of the independence of interarrival times, some time instants play an important role in the analysis of the $GI/M/N/N$ queue. These time instants occur when a new call arrives, since the evolution of the system after that time is independent of the previous history of the process. These points, called *regeneration points*, are labeled t_r.

The process of interest, denoted N_r, is the number of busy circuits when the r^{th} call arrives. If t_0 is a regeneration point, the renewal property means that $Pr\{N(t_r) \mid N(t_0)\} = Pr\{N(t) \mid N(\tau) \quad \forall\ \tau \leq t_0\}$. That is, there is an embedded Markov chain defined by the process N_r corresponding to the arrivals. A great deal of information can be learned from this Markov chain, in particular, the value of p_j, the stationary probability of being in state j at the time a call arrives in the system.

The calculation of the probabilities is straightforward if we have the transition matrix $Q_{i,j}$, the transition probability from i to j. We have

$$Q_{i,j} = Pr\{N_{r+1} = j \mid N_r = i\}$$

$Q_{i,j}$ is the conditional probability that an arriving call sees j busy circuits, given that the preceding call saw i busy circuits when it arrived. This expression can be evaluated using standard results for the age of an interval under renewal input. Immediately after the r^{th} arrival, the state was $i + 1$. Just before the $r + 1^{\text{th}}$ arrival, the state was j. There must have been $i + 1 - j$ terminations during this interval. The conditional distribution of states during this interval

is given by

$$Q_{i,j} = \binom{i+1}{j} \int_0^\infty e^{-\mu t j}(1-e^{-\mu t})^{i+1-j} dF(t), \quad j-1 \le i \le N-1 \tag{3.42}$$

and

$$Q_{i,j} = 0, \quad i < j-1$$
$$Q_{N,j} = Q_{N-1,j}.$$

The derivation of this equation is given in Section A.3.

We are now interested in the stationary probability of being in state j when a call arrives, which we denote p_j. This can be computed by the standard methods of Markov chains by solving the equation $\sum_{i=0}^{N} p_i Q_{i,j} = p_j$, with the normalization $\sum_{j=0}^{N} p_j = 1$ [1]. If we call $\psi(z) = \sum_{j=0}^{\infty} p_j z^j$ the z transform of p_j, and $\Phi(s)$ the Laplace-Stieltjes transform of $F(t)$, we have

$$\begin{aligned}
\psi(z) &= \sum_{j=0}^{N} \sum_{i=j-1}^{N-1} p_i z^j Q_{i,j} + p_N \sum_{j=0}^{N} z^j Q_{N,j} \\
&= \sum_{i=0}^{N-1} p_i \sum_{j=0}^{i+1} z^j Q_{i,j} + p_N \sum_{j=0}^{N} z^j Q_{N,j} \\
&= \sum_{i=0}^{N-1} p_i \int_0^\infty (1-a+az)^{i+1} dF(t) \\
&\quad + p_N \int_0^\infty (1-a+az)^N dF(t),
\end{aligned} \tag{3.43}$$

where

$$a = e^{-\mu t}$$

We now compute the binomial moments of the distribution. We know from the properties of z transforms that

$$\begin{aligned}
\beta_q &\triangleq \sum_{j=q}^{N} \binom{j}{q} p_j \\
&= \frac{1}{n!} \left(\frac{d^q \psi(z)}{dz^q} \right)_{z=0}
\end{aligned}$$

We differentiate Eq. (3.43) n times. For the binomial moments, we get

$$\beta_q = \Phi(q\mu) \left[\sum_{i=0}^{N-1} p_i \binom{i+1}{q} + p_N \binom{N}{q} \right]. \tag{3.44}$$

The recurrence for the coefficients is

$$[1 - \Phi(q\mu)]\beta_q = \Phi(q\mu)\beta_{q-1} - p_N \binom{N}{q-1} \Phi(q\mu), \tag{3.45}$$

which is left as an exercise (Problem 3.11). The solution for the moments can be obtained in closed form starting the recurrence at 0. We obtain the binomial moments of the busy-circuit distribution at the time of a call arrival:

$$\overline{\beta}_i = h_i \frac{\sum_{j=i}^{N} \binom{N}{j} \frac{1}{h_j}}{\sum_{j=0}^{N} \binom{N}{j} \frac{1}{h_j}},$$

where

$$h_n = \prod_{j=1}^{n} \frac{\Phi(j\mu)}{1 - \Phi(j\mu)} \quad n \geq 1, \quad h_0 = 1$$

Given these expressions for the moments, we can get the state probabilities from Eq. (A.13), and in particular the call congestion, which is given by

$$B = p_N = \frac{1}{\sum_{j=0}^{N} \binom{N}{j} \frac{1}{h_j}}. \tag{3.46}$$

The first two moments of this distribution, which is not to be confused with the carried traffic, are given by

$$\overline{\alpha}_1 = (1 - B)\frac{\Phi(\mu)}{1 - \Phi(\mu)}$$

$$\overline{\alpha}_2 = \frac{(\overline{\alpha}_1 + 1)\Phi(\mu) + \overline{\alpha}_1\Phi(2\mu) - B\left[\Phi(\mu) + 2N\Phi(2\mu)\right]}{1 - \Phi(2\mu)}.$$

Having solved the distribution of the embedded Markov chain, we can now obtain the distribution at an arbitrary instant in time and, from this, the time congestion and the distribution of carried traffic. Here, we use the asterisk superscript to emphasize that the quantities of interest refer to the distribution at an arbitrary instant in time. Let q_j be the stationary distribution of busy circuits at an arbitrary time t, Let $Q^*_{i,j}$ be the conditional probability that there are j busy circuits at time t, given that there were i busy circuits when the interval between these events began. This is the transition matrix, which can be expressed in terms of the transition matrix of the embedded Markov chain. By a similar argument, we have

$$Q^*_{i,j} = \binom{i+1}{j} \int_0^\infty (1-a)^{i+1-j} a^j dF^*(t), \quad j - 1 \leq i \leq N - 1$$

$$Q^*_{N,j} = Q^*_{N,j}$$

$$Q^*_{i,j} = 0, \quad i < j - 1$$

The only difference from the Markov chain is in the expression of $F^*(t)$, the age of the interval since the last arrival. Because we are now looking at the interval at an arbitrary point in time, the value of this distribution is given by Eq. (A.24), as discussed in Section A.3.

$$F^*(t) = \lambda \int_0^t [1 - F(\tau)] d\tau$$

From this, we get

$$q_j = \sum_{i=j-1}^{N} p_i Q^*_{i,j}.$$

This is the required distribution, but it can be put in a more convenient form through the binomial moments. The calculation is the same as for the embedded Markov chain; we get

$$\overline{\beta}^*_i = \frac{A}{i} \left[\overline{\beta}_{i-1} - \binom{N}{i-1} B \right]. \tag{3.47}$$

From this, we immediately get the time congestion $E = q_N$ and the first two parameters of the carried traffic:

$$E = B \left(\frac{A}{N} \right) \frac{1 - \Phi(N\mu)}{\Phi(N\mu)}. \tag{3.48}$$

This equation establishes a general relation between the call and time congestion for arbitrary renewal input. It is not hard to show that, for Poisson arrivals, these are equal. The converse is somewhat more difficult, and is left as an exercise (Problem 3.12). The first two moments of the carried traffic are given by

$$\overline{\alpha}^*_1 = \overline{M} = A(1 - B) \tag{3.49}$$

as one would expect, and

$$\overline{\alpha}^*_2 = \overline{M} \left[1 - \overline{M} + h_1 - N \frac{B}{1 - B} \right]. \tag{3.50}$$

These results can be used to derive the characterization of a renewal process in terms of the first moment of the traffic carried in a group of infinite size, in other words, the parameters of the offered traffic generated by the renewal process. This is done by taking $N \to \infty$ in Eqs. (3.45) and (3.47); we get, respectively,

$$\beta_n = h_n \tag{3.51}$$

$$\beta^*_n = A \frac{h_{n-1}}{n} \tag{3.52}$$

From these values, we can compute the mean and variance of the offered traffic corresponding to F (i.e., the parameters of the traffic carried in the infinite group):

$$M = M_1^* = -\frac{1}{\mu\Phi'(0)} = A \tag{3.53}$$

$$V = M_2^* - (M_1^*)^2 = M\left(\frac{1}{1-\Phi(\mu)} - M\right) \tag{3.54}$$

Note that Eq. (3.53) is in fact a theorem. It states that the expected number of busy circuits in the infinite group is precisely the intensity of the input stream. This point is generally taken for granted, since circuit occupancy is commonly used to measure input traffic, but nevertheless can be demonstrated mathematically.

As we said before, some calls accepted for service on the link may require service from other links in order to establish a connection to their destination. In this sense, the carried process can be viewed as the arrival process to some other link, just as was the case for overflow. Let us therefore end this discussion of the carried traffic by computing the interarrival time distribution of the carried calls. We follow the method described in [11]. As usual, let $F(t)$ be the distribution of interarrival times and $F_c(t)$ the distribution of the interval τ_n between a call accepted at t_n and the next accepted call. Let $N(t)$ be the number of calls in progress at time t. Then,

$$F_c(t) = Pr\left\{\tau_n \leq t \mid N(t_n^-) < N\right\}.$$

We must consider separately two cases immediately after t_n. After the call is accepted, the group may be full or some free circuit may still be available. Corresponding to these two cases, we define

$$F_c[t \mid N(t_n^+) < N] = Pr\left\{\tau_n \leq t \mid N(t_n^+) < N\right\}$$
$$\overline{F}_c(t) = Pr\left\{\tau_n \leq t \mid N(t_n^+) = N\right\}$$

Now, if the group is not full after the call is accepted, the interarrival distribution is just given by

$$F_c[t \mid N(t_n^+) < N] = F(t),$$

so that

$$F_c(t) = F(t)\left[1 - \frac{p_{N-1}}{1-p_N}\right] + \overline{F}_c(t)\left[\frac{p_{N-1}}{1-p_N}\right],$$

where p_j is the probability that an arriving call will see j busy circuits. In particular, p_N is the call congestion as defined in Eq. (3.46). We can now take the Laplace-Stieltjes transform of the equation, and get

$$\Phi_c(s) = \Phi(s)\left[1 - \frac{p_{N-1}}{1-p_N}\right] + \overline{\Phi}_c(t)\left[\frac{p_{N-1}}{1-p_N}\right]. \tag{3.55}$$

The only element missing from this equation is the value of $\overline{\Phi}(s)$, the transform of the interarrival distribution to the next accepted call when the currently accepted call saturates the group. The derivation of this expression can be found in [12]; it is reproduced here without derivation. We have

$$\overline{\Phi}_c(s) = [1 - \Phi(s)] \int_0^\infty e^{-st} H(t) dM(t),$$

where $H(t) = 1 - e^{-Nt}$ and $M(t)$ is the expected number of arrivals in the interval $(0, t)$. Note the dependence of the distribution on the terms p_N and p_{N-1}. This dependence means that the interarrival times depend on the state of the system, and that the process is not a renewal process. As a consequence, we cannot use this distribution to characterize completely the traffic offered to the second link in a path. We see later that this is an important difference between the carried and overflow processes.

Let us now pause to consider what has been achieved. We assumed that the arrival process to a group of size $N \leq \infty$ is an arbitrary renewal process. From this, we gave a complete description of the busy-circuit distribution in the group, in terms of both probabilities and moments. The case of the infinite group was handled as a special case; its importance will become clearer when we discuss moment-matching methods in Section 3.4.

Overflow Traffic

The next step in the study of the $GI/M/N/N$ queue is to characterize the overflow process. A very important result from the renewal approach is a complete solution to the overflow cascade. In this system, a first group of finite size is offered Poisson traffic or, more generally, an arbitrary renewal stream. Blocked calls are then offered to a second group of finite size, the overflow of which is offered to a third, and so on. While the analysis of such a system by BD methods becomes intractable very rapidly as the amount of overflow increases, the renewal method gives a recursive approach that allows a complete solution for an arbitrary number of overflows. The main advantage of the renewal method over the BD technique is that the same formulas can be used at any stage, although the actual expressions can become quite complex. Even though they generally cannot be written analytically, it is possible to have a numerical evaluation procedure that is the *same* at all stages, which is not possible with the BD approach.

To do this, we simply realize that if the input to the primary group is renewal, then the input to the secondary group is also renewal, and so on for all the groups in the cascade. This is why we want to compute the distribution function of the interarrival times of the overflow process from the knowledge of the distribution of the input interarrival times. Let us now proceed to do this,

following Pearce and Potter [13]. We also give a complete characterization of the overflow distribution.

Define $F_n(t)$, the distribution function for the time interval T separating the following two events — first, at the time when a call arrives and finds n busy circuits and, second, at the time of the next overflow. We write a set of integral equations for this distribution. Consider the system immediately after the call is accepted: There are $n+1$ calls in the system. Let Y be the interval of time starting immediately after the call is accepted during which there is no arrival. This random variable has a distribution $F(y)$, the interarrival distribution. During Y, there can be between 0 and $n+1$ call terminations. Because the calls in progress are independent, the probability of j terminations has a Bernouilli distribution with parameter $a = \exp(-\mu y)$. At the end of the interval, we are in state j, and calls start arriving. From this point on, the distribution of the interval to the next overflow is by definition $F_j(t-y)$. The event that the interval T is no greater than t is then the convolution of two events: The Y interval is no larger than y, and the state after Y is j. This is written

$$F_n(t) = \sum_{j=0}^{n+1} \int_0^t \binom{n+1}{j} (1-a)^{n+1-j} a^j F_j(t-y) dF(y), \quad 0 \le n \le N \tag{3.56}$$

which is the set of required integral equations. We are interested in the interarrival time distribution of blocked calls. The interval between two overflows has a distribution that is the distribution between two consecutive arrivals that find N busy servers. If we call this distribution $G(t)$, we have

$$G(t) = \sum_{j=0}^{N} \int_0^t \binom{N}{j} (1-a)^{N-j} a^j F_j(t-y) dF(y),$$

which requires the calculation of the $F_j(t)$s.

We must compute this distribution or, more simply, its Laplace-Stieltjes transform. Let $\Psi_n(s)$ be the Laplace-Stieltjes transform of $F_n(t)$. Using the fact that the transform of a convolution is the product of the transforms of the individual distributions, Eq. (3.56) becomes, for $0 \le n \le N$ and $\mathrm{Re}\ s \ge 0$,

$$\Psi_n(s) = \sum_{j=0}^{n+1} \Psi_j(s) \int_0^\infty e^{-st} \binom{n+1}{j} (1-a)^{n+1-j} a^j dF(t) \tag{3.57}$$

Note the condition $n \le N$ imposed on Eq. (3.57). To simplify, we remark that the solutions are the same if we lift this restriction on n but replace it with the condition that $\Psi_N(s) = 1$. Introducing the generating function,

$$\Psi(z) \triangleq \sum_{n=0}^{\infty} \frac{\Psi_n(s) z^n}{n!}, \tag{3.58}$$

we multiply Eq. (3.57) by $z^n/n!$. Summing, we obtain

$$\begin{aligned}\Psi(z) &= \sum_{j=1}^{\infty} \Psi_j(s) z^j \sum_{k=0}^{\infty} \int_0^{\infty} e^{-st} z^{k-1} \binom{k+j}{k} (1-a)^k a^j dF(t) \\ &\quad + \Psi_0(s) \int_0^{\infty} e^{-st} (1-a) e^{z(1-a)} dF(t) \\ &= \int_0^{\infty} e^{-st} \frac{d}{dz} \left\{ \Psi(az) e^{z(1-a)} \right\} dF(t) \\ &= \int_0^{\infty} e^{-st} \frac{d}{dz} \left\{ e^z \Psi(az) e^{-az} \right\} dF(t) \\ &= e^z \int_0^{\infty} e^{-st} \left[1 + \frac{d}{dz} \right] \left\{ \Psi(az) e^{-az} \right\} dF(t). \end{aligned} \tag{3.59}$$

Define

$$k(z) \triangleq \Psi(z) e^{-z} \tag{3.60}$$

Using Eq. (3.58), we get

$$k(z) = \int_0^{\infty} e^{-st} \left(1 + \frac{d}{dz} \right) k(za) dF(t),$$

which is an integral equation for $k(z)$. We can solve this by introducing yet another transform with the coefficients $k_j(s)$:

$$k(z) = \sum_{j=0}^{\infty} \frac{k_j z^n}{n!},$$

where it should be remembered that the coefficients k_j are in fact functions of s, the transform variable for t. Now replacing the value of $k(z)$ in Eq. (3.59) and setting equal the coefficients of identical powers of z, we get the recurrence relation

$$k_{n+1}(s) = k_n(s) \frac{[1 + \Phi(s + n\mu)]}{\Phi(s + (n+1)\mu)}$$

whose solution is

$$k_n(s) = k_0(s) \prod_{j=1}^{n} \frac{[1 - \Phi(s + (j-1))\mu]}{\Phi(s + j\mu)}, \tag{3.61}$$

from which we can recover the transform

$$\Psi_n(s) = \sum_{j=0}^{n} \binom{n}{j} k_j(s). \tag{3.62}$$

We know that $\Psi_n(s)$ is the Laplace-Stieltjes transform of $F_n(t)$. From the definition of the Laplace-Stieltjes transform, we see that $\Psi_n(0) = 1$; from Eq. (3.58), it follows that $\Psi(z) = e^z$ when $s = 0$. From Eq. (3.60), we can then conclude that $k_n(0) = \delta_{n,0}$. It can be verified that the solution where the normalization constant $k_0(s) = 1$ satisfies the original equation. The knowledge of Ψ_{N-1} is all that is needed for the distribution of intervals between overflows. Recall that Ψ_{N-1} is the transform of the distribution of the interval between the time a call arrives and finds $N-1$ busy servers, and the time of the next overflow. After this call is accepted, however, the interval to the next overflow is statistically the same as the inter-overflow interval, whose distribution is defined as $G(t)$. We then have

$$\hat{\Phi}(s) = \frac{\sum_{j=0}^{N-1} \binom{N-1}{j} k_j(s)}{\sum_{j=0}^{N} \binom{N}{j} k_j(s)}. \tag{3.63}$$

The knowledge of the distribution of interarrival times for the overflow process is all that is needed to characterize the overflow traffic. In theory, all the moments could be computed by replacing the $\hat{\Phi}(s)$ of Eq. (3.63) in Eqs. (3.52) and (3.51), and the blocking probability could be obtained by substitution in Eqs. (3.46–3.48). The Laplace-Stieltjes transform of the overflow of *this* traffic can again be computed by Eq. (3.63), and so on recursively any number of times.

Recursive Formulation of Overflow Moments

In practice, the expressions for the moments of the overflow traffic in terms of the arrival process can become so complex that they cannot easily be used. Consider simply the calculation of the mean of the overflow traffic. To get this quantity, we need $\hat{\Phi}'(s)$. Taking the derivative at the origin of Eq. (3.63), and using the fact that $k_r(0) = \delta_{r,0}$, we get

$$\begin{aligned}\hat{\Phi}'(0) &= \sum_{r=0}^{N-1} \binom{N-1}{r} k_r'(0) - \sum_{r=0}^{N} \binom{N}{r} k_r'(0) \\ &= \sum_{r=1}^{N} \binom{N-1}{r-1} k_r'(0)\end{aligned}$$

and, from the definition of the k_rs,

$$k_r'(0) = -\frac{\Phi'(0)}{\Phi(\mu)} \prod_{l=2}^{r} \frac{1 - \Phi[(l-1)\mu]}{\Phi(l\mu)}$$

$$= -k_{r-1}(\mu)\frac{\Phi'(0)}{\Phi(\mu)},$$

which gives

$$\hat{\Phi}'(0) = \frac{\Phi'(0)}{\Phi(\mu)} \sum_{r=0}^{N-1} \binom{N-1}{r} k_r(\mu), \tag{3.64}$$

from which we get the mean of the overflow traffic from Eq. (3.53). In the same manner, we could compute the variance of the overflow by Eq. (3.54) with Φ replaced by $\hat{\Phi}$, but there is no simple expression in terms of the transform of the arrival process.

More useful expressions are given in [14] for the moments of the overflow traffic, and in particular for the mean and variance, directly in terms of the interarrival distribution for the traffic offered to the primary group. The recursion, given in terms of the order of the factorial moments and of the size N of the primary group, can be transformed readily into a numerical procedure. We must extend the notation for the moments to $M_q(N)$, which is the q^{th} factorial moment of the overflow distribution when there are N circuits in the primary group. In the same way, transforms are indexed by N whenever necessary. First, we give some useful relations between the transforms of the arrival and overflow processes. The intensity of the renewal stream — none other than the mean offered traffic — is defined as $I = -1/\mu\Phi'(0)$; its inverse is called the *weakness* by Potter [14]. This inverse is denoted by w for an infinite group, in which case it represents the weakness of the offered traffic, or by W_N when we want to consider the overflow stream from a primary group of size N.

Because the weakness of the stream plays a central role in calculating the overflow moments, we first give a relation between weaknesses of input and overflow streams. From Eq. (3.64), we get the relation for the weakness:

$$W_N = -\mu\frac{\Phi'(0)}{\Phi(\mu)} \sum_{r=0}^{N-1} \binom{N-1}{r} k_r(\mu). \tag{3.65}$$

We now use the fact that the call congestion is the ratio of the overflow mean to the offered mean, which we write as

$$B = \frac{\Phi'(0)}{\hat{\Phi}'(0)},$$

or equivalently, from Eq. (3.65),

$$B = \frac{\Phi(\mu)}{\sum_{r=0}^{N-1} \binom{N-1}{r} k_r(\mu)}. \tag{3.66}$$

The call-congestion probability can also be written from Eq. (3.46) as

$$B = \frac{1}{\sum_{r=0}^{N} \binom{N}{r} \frac{1}{h_r(\mu)}}. \tag{3.67}$$

Combining Eqs. (3.66) and (3.67), we obtain

$$\frac{1}{\Phi(\mu)} \sum_{r=0}^{N-1} \binom{N-1}{r} k_r(\mu) = \sum_{r=0}^{N} \binom{N}{r} \frac{1}{h_r(\mu)}.$$

From this equation, the weakness of the overflow can be expressed as a function of the weakness w of the arrival process by replacing the left-hand side in Eq. (3.65), yielding

$$W_N = w \sum_{r=0}^{N} \binom{N}{r} \frac{1}{h_r(\mu)}, \tag{3.68}$$

which can be written as a recurrence for the weakness

$$\Delta^n W_N = w \sum_{r=0}^{N} \binom{N}{r} \frac{1}{h_{n+r}(\mu)}, \tag{3.69}$$

which can be proved by induction. In this equation, the operator Δ is defined as the forward difference operator. It is defined formally by the following equations:

$$\begin{aligned} \Delta^n f(n) &= (E-1)^n f(n) \\ E^k f(n) &= f(n+k) \\ (E-1)^k f(n) &= \sum_{j=0}^{k} \binom{k}{j} E^k f(n) \end{aligned}$$

Expression (3.69) provides an easy way to calculate the mean of the overflow given the transform of the arrival process by means of a recursion over the group size. In a sense, this is the generalization of the well-known recurrence of Eq. (3.4) for the Erlang B function.

We now continue the derivation of some results that will be needed in calculating all the moments of the overflow by a recurrence of the same type. From the definition of $\hat{\Phi}(s)$, and using the identity

$$\binom{n}{r-1} + \binom{n}{r} = \binom{n+1}{r},$$

we get

$$\frac{\hat{\Phi}_N(s)}{1-\hat{\Phi}_N(s)} = \frac{\sum_{r=0}^{N-1}\binom{N-1}{r}k_r(s)}{\sum_{r=0}^{N-1}\binom{N-1}{r}k_{r+1}(s)}.$$

We now use the relation

$$k_r(s+\mu) = k_{r+1}(s)\frac{\Phi(s+\mu)}{1-\Phi(s)},$$

which can be obtained directly from the definition of $k_r(s)$, to obtain

$$\hat{\Phi}_N(s+\mu) = \hat{\Phi}_N(s)\left(\frac{1-\frac{1}{\hat{\Phi}_N(s)}}{1-\frac{1}{\hat{\Phi}_{N+1}(s)}}\right). \tag{3.70}$$

Given these preliminary results, we can derive the promised recurrence equation for the moments of the overflow traffic. Setting $s=(n-1)\mu$, we have

$$\frac{1}{\hat{\Phi}_N(n\mu)} = \frac{1}{\hat{\Phi}_N((n-1)\mu)}\left[\frac{1}{\hat{\Phi}_{N+1}((n-1)\mu)}-1\right] \times \left[\frac{1}{\hat{\Phi}_N((n-1)\mu)}-1\right]^{-1}. \tag{3.71}$$

We introduce the factorial moments in this expression by replacing $\hat{\Phi}$ in terms of the moments. Using Eq. (3.52), we get

$$\frac{1}{\hat{\Phi}_N(n\mu)} = 1+n\frac{M_n(N)}{M_{n+1}(N)}.$$

Replacing in Eq. (3.71), we get

$$1+n\frac{M_n(N)}{M_{n+1}(N)} = \left[1+(n+1)\frac{M_{n+1}(N)}{M_n(N)}\right] \times \frac{M_n(N+1)}{M_{n-1}(N)}\frac{M_n(N)}{M_n(N+1)}. \tag{3.72}$$

The general theory of renewal arrivals gives a complete characterization of an arbitrary arrival stream of the renewal type. In practice, this is used for simple arrival processes, where the computation of the carried and overflow parameters is much easier in the renewal framework. We now give two examples in which the general theory is used to describe well-known processes.

The IPP Process as a Renewal Process. As we have seen above, the description of the interrupted Poisson process as a two-dimensional Markov chain is somewhat involved, and it would be difficult to describe the overflow parameters. We now show that an arbitrary IPP process is of the renewal type; we compute its interarrival time distribution. From this, using the general theory, we immediately can find the moments of the overflow and carried traffic. We follow the method outlined in [9].

First recall some definitions:

$F(t)$ = The interarrival time distribution.

W_n = The waiting time from 0 to the n^{th} arrival.

$H_n(t)$ = The distribution of W.

$N(t)$ = The number of arrivals from 0 to t.

$p_k(t) = P[N(t) = k]$.

$p_{km}(t)$ = The probability that there are k arrivals from 0 to t, given that there was an arrival at 0, when the switch is in state m. Recall that $m = 1$ (respectively 0) indicates that the switch is on (respectively off).

We have

$$H_n(t) = 1 - \sum_{k=0}^{n-1} p_k(t). \tag{3.73}$$

Taking the Laplace-Stieltjes transform of this equation, we get

$$h_n(s) = 1 - s \sum_{k=0}^{n-1} \pi_k(s) \tag{3.74}$$

$$h_n(s) \triangleq \int_0^\infty e^{-st} dH_n(t),$$

$$\pi_k(s) \triangleq \int_0^\infty e^{-st} p_k(t) dt$$

We can write a differential equation for $p_{km}(t)$ by the same techniques used to derive the stationary equations of Markov chains. We know that, in a small interval dt, the probability that the switch will go from the off state to the on state is given by ωdt. Similarly, the probability that a new customer will arrive is given by λdt. Using these arguments, we can write

$$p_{01}(t + dt) = \omega p_{00}(t) + (1 - (\lambda + \gamma))\, p_{01}(t)$$

$$\frac{p_{01}(t + dt) - p_{01}(t)}{dt} = \omega p_{00}(t) - (\lambda + \gamma) p_{01}(t).$$

Letting $dt \to 0$,

$$\frac{d}{dt}p_{01}(t) = \omega p_{00}(t) - (\lambda + \gamma)p_{01}(t).$$

By a similar argument,

$$\frac{d}{dt}p_{k1}(t) = \omega p_{k0}(t) - (\lambda + \gamma)p_{k1}(t) + \lambda p_{k-1,1}(t), \quad k = 1, \ldots$$
$$\frac{d}{dt}p_{k0}(t) = -\omega p_{k0}(t) + \gamma p_{k1}(t)$$
$$p_{01}(0) = 1$$

As usual, we take the Laplace transform of this set of equations:

$$s\pi_{01}(s) = \omega\pi_{00}(s) - (\lambda + \gamma)\pi_{01}(s) + 1 \tag{3.75}$$
$$s\pi_{k1}(s) = \omega\pi_{k0}(s) - (\lambda + \gamma)\pi_{k1}(s) + \lambda\pi_{k-1,1}(s), \quad k = 1, \ldots \tag{3.76}$$
$$s\pi_{k0}(s) = -\omega\pi_{k0}(s) + \gamma\pi_{k1}(s) \tag{3.77}$$

where

$$\pi_{kj}(s) \triangleq \int_0^\infty e^{-st}p_{kj}(t)dt.$$

This set of difference equations can be solved readily, yielding

$$\pi_k(s) = \frac{s + \omega + \gamma}{g(s)}\left[\frac{\lambda(s + \omega)}{g(s)}\right]^k, \quad k = 1, \ldots \tag{3.78}$$
$$g(s) \triangleq s^2 + (\lambda + \gamma + \omega)s + \lambda\omega$$

Replacing in Eq. (3.74), we have

$$h_n(s) = \left[\frac{\lambda(s + \omega)}{g(s)}\right]^n, \quad n = 1, \ldots$$

and taking the inverse transform, we get the required distribution

$$F(t) = k_1(1 - e^{-r_1 t}) + k_2(1 - e^{-r_2 t}), \tag{3.79}$$

where

$$r_1 = \frac{1}{2}\left[\lambda + \omega + \gamma + \sqrt{(\lambda + \omega + \gamma)^2 - 4\lambda\omega}\right]$$
$$r_2 = \frac{1}{2}\left[\lambda + \omega + \gamma - \sqrt{(\lambda + \omega + \gamma)^2 - 4\lambda\omega}\right]$$
$$k_1 = \frac{\lambda - r_2}{r_1 - r_2}$$
$$k_2 = 1 - k_1$$

It is left as an exercise (Problem 3.14) to show that these parameters are nonnegative for an arbitrary IPP. Given the interarrival distribution, all the parameters of interest concerning the carried and overflow processes can be computed from the general model.

Multistage Systems

The IPP generator produces an arrival process with interarrival times given by the sum of two exponentials with different weights. This method of producing arrival processes can be extended to produce almost any type of arrival to any given accuracy.

The simplest such system, the k-stage exponential process, is composed of k single-server queues in series with identical servers. A customer arrives at the entry of the system, passes through all the servers one after the other, and exits to produce the event that marks the arrival of a new call in the group. At each stage, the service time is drawn from the same negative exponential distribution with parameter $k\mu$. Immediately after the call arrives, another customer enters the first stage, and the process is repeated. It is evident that the total time spent by a customer in the system is the sum of k independent negative exponential variables. Thus the Laplace transform of the transit time, which is also the distribution of the interarrival time of calls in the group, is simply the convolution of the individual transit times. It is given by

$$\Phi(s) = \left(\frac{k\mu}{s + k\mu}\right)^k.$$

The interarrival time density is the Erlang$_k$ distribution, as can be seen in any table of Laplace transforms. It is given by

$$f(t) = \frac{k\mu(k\mu t)^{k-1}e^{-k\mu t}}{(k-1)!}.$$

This distribution has a straightforward interpretation as the distribution of the time required to collect k arrivals from a Poisson process of parameter μ.

It is now quite easy to compute the first two moments of the traffic generated by this process. From Eqs. (3.53) and (3.54), we get

$$M = \mu$$

$$Z = 1 - \mu + \frac{(k\mu)^k}{(1 + k\mu)^k - (k\mu)^k}$$

It is not difficult to see that the traffic generated by this process is smooth, although computing the smallest peakedness available appears quite difficult.

This technique can be extended in various ways to yield increasingly complex arrival distributions. One obvious choice is to draw the individual service times at each stage from negative exponential distributions with different means. Another possibility is illustrated by the two-stage Cox model [15]. Consider the system of Fig. 3.1, where the two stations are exponential servers with distinct rates α and β, respectively. Here, however, a customer entering the generator need not go through all stages. Instead, a customer arriving at A

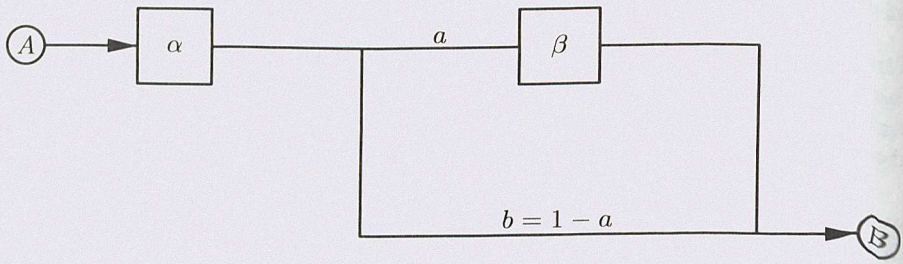

Figure 3.1 Equivalent System for Cox Model

immediately enters service at station 1. After its service is completed, the customer exits at B with probability $b = 1 - a$, or enters server 2 with probability a, Customers leaving server 2 then exit, generating a new call arrival, and a new customer immediately arrives at A. The time spent by a customer in the system is a weighted sum of exponential distributions with different holding times; its Laplace transform is given by

$$\Phi(s) = \frac{\alpha}{\alpha + s}\left[b + a\frac{\beta}{\beta + s}\right]. \tag{3.80}$$

The moments of the traffic generated by the process can be computed by Eq. (3.52); the mean and peakedness are given by

$$M = \frac{\alpha\beta}{(a\alpha + \beta)} \tag{3.81}$$

$$Z = \frac{1 + \alpha + \beta}{1 + \alpha a + \beta} - \frac{\alpha\beta}{(1 + \alpha a + \beta)(\alpha a + \beta)} \tag{3.82}$$

Note that, in this case, the value of peakedness produced by this process depends on *three* parameters, making the analysis quite complex.

Let us now compute the carried and overflow distributions. Of course, we could do this using the general formulas (3.47) and (3.63). Instead, we show how to compute the interarrival distribution of the carried traffic directly from the description of the system, demonstrating that it is often possible to get representations that are simpler than for the general case by exploiting the structure of a given arrival process.

Suppose that a call is offered to a group of finite size N whenever a customer leaves the generator. We want the Laplace transform of the interarrival time distribution for calls accepted in the group. When a call is accepted, the generator can be in either one of two states: in state 1, where the customer is being served in stage 1, or in state 2, where it is in the second server. Thus

we are led to define $X(i,j)$, the interval from an arbitrary instant to the next arrival when there are i customers in service in the group and the generator is in state j, where $j = 1, 2$. The corresponding transforms are denoted $\Phi_i^j(s)$.

Consider now the situation where the generator is in state 1. Two independent exponential service processes are running concurrently: the first server of the generator, at rate α, and the customers in service, at rate $i\mu$. The interval to the next event has a Laplace transform given by the combination of the two processes, that is, $(\alpha + i\mu)/(\alpha + i\mu + s)$. The next event can be one of three possibilities:

1. The customer leaves the server and leaves the generator, in which case a call arrival is produced. This has probability $\alpha b/(\alpha + i\mu)$. The Laplace transform of the subsequent interval is 1, since the waiting time for the next customer arrival to the first server is zero.

2. The customer leaves the server, and enters the second stage of the generator, with no consequent new-call arrival. This has probability $\alpha a/(\alpha + i\mu)$. The subsequent interval has a Laplace transform given by $\Phi_i^2(s)$.

3. The event is a call termination in the group, with probability $i\mu/(\alpha+i\mu)$. The subsequent interval has a Laplace transform given by $\Phi_{i-1}^1(s)$.

The Laplace transform of the interval distribution is then given by

$$\Phi_i^1(s) = \frac{\alpha + i\mu}{\alpha + i\mu + s}\left[\frac{\alpha b}{\alpha + i\mu} + \frac{\alpha a}{\alpha + i\mu}\Phi_i^2(s) + \frac{i\mu}{\alpha + i\mu}\Phi_{i-1}^1(s)\right], \quad i = 0, \ldots N \tag{3.83}$$

and by a similar argument for the other state:

$$\Phi_i^2(s) = \frac{\beta + i\mu}{\beta + i\mu + s}\left[\frac{\beta}{\beta + i\mu} + \frac{i\mu}{\beta + i\mu}\Phi_{i-1}^2(s)\right], \quad i = 0, \ldots N \tag{3.84}$$

$$\Phi_k^j(s) = 0 \text{ if } k < 0$$

We end up with a system of $2(N+1)$ equations in as many unknowns. Because of the particular structure of the recurrence, using the fact that $\Phi_0^1(s) = \Phi(s)$, we can get a solution in closed form:

$$\Phi_N^1(s) = \frac{\alpha N\mu\left[\alpha\beta(\alpha + s) + (\beta + N\mu + s)(\beta + bs)\right]}{(\alpha + s)(N\mu + s)\left[\alpha a + \beta + N\mu + s\right](\beta + s)}. \tag{3.85}$$

Let p_n be the probability that the state of the group is n just after an arrival. We can compute the transform of the carried-call distribution by noting that

$$\overline{\Phi}(s) = (1 - p_N)\Phi(s) + p_N\Phi_N^1(s). \tag{3.86}$$

This equation simply states that if the group is not full immediately after a call is accepted, then the distribution until the next accepted call is just the interarrival distribution. If, on the other hand, the group is full after the last call accepted, then the interval is determined by $\Phi^1_N(s)$. Note that $\Phi^2_N(s)$ is not used since we know that, immediately after a call is accepted, the generator is in state 1. The probability of having a full group immediately after a call is accepted is given by

$$\begin{aligned} p_N &= E\frac{N\mu(\beta+\alpha a)}{\alpha\beta(1-B)} \\ &= \frac{B}{1-B}\frac{N\mu(\alpha a+\beta+N\mu)}{\alpha(\beta+bN\mu)}, \end{aligned}$$

where the call congestion B can be computed from Eq. (3.46). From Eq. (3.86), the various moments of the carried traffic can be computed by standard methods, although an analytic formulation would probably be quite complex.

The final point in studying the Cox generator is to compute the transform $\hat{\Phi}(s)$ of the interarrival distribution of the overflow process. This can always be done using the general technique of Eq. (3.63). Although we can state the equations using a technique similar to the one by which we arrived at the transform of the carried traffic, there does not seem to be a closed-form solution such as in Eq. (3.85).

Let $\hat{\Phi}^1_i(s)$ be the transform of the interval between the current instant and the next overflow, given that i calls are present and that the generator is in state 1, and let $\hat{\Phi}^2_i(s)$ be the same transform for state 2. We can write the following general equation relating these transforms, using an argument along the same lines as for the carried traffic. We get

$$\begin{aligned} (\alpha+i\mu+s)\hat{\Phi}^1_i(s) &= \alpha b\hat{\Phi}^1_{i+1}(s)+\alpha a\hat{\Phi}^2_i(s)+i\mu\hat{\Phi}^1_{i-1}(s) \\ (\beta+i\mu+s)\hat{\Phi}^2_i(s) &= \beta\hat{\Phi}^1_{i+1}(s)+i\mu\hat{\Phi}^2_{i-1}(s) \end{aligned}$$

This is a linear system in the unknown $\hat{\Phi}^1_i(s)$ and $\hat{\Phi}^2_i(s)$. Although there is no easy analytic solution for the system, it can be solved numerically by noting that the moments are obtained from the derivatives of the transform taken at the origin. Write the system as

$$A(s)\mathbf{x}(s) = \mathbf{b}(s),$$

where $\mathbf{b}(s)$ is the vector of $\hat{\Phi}^1_i(s)$ and $\hat{\Phi}^2_i(s)$, and $A(s)$ is the matrix of coefficients. It is possible to verify that $dA(s)/ds = I$, the identity matrix. Thus, taking the n^{th} derivative of the linear system with respect to s and evaluating it at zero, we obtain

$$A(0)\mathbf{b}^{(n)}(0)+\mathbf{b}^{(n-1)}(0) = 0,$$

which we can then solve numerically for the value of the coefficient matr obtaining all the moments of the overflow distribution. These expressions a not particularly simple, for which reason we now end our discussion of t two-stage Cox model.

This model is a simple case of a more general class studied by Cox [16, The n-stage Cox system is a straightforward extension of Fig. 3.1, where the stages are placed in tandem, each having its own service time rate and prob bility of exit. Such a system can produce an arrival process having an arbitra rational Laplace transform, subject to some mild constraints on the paran ters. This generality in producing traffic streams is not particularly useful the model has a large number of stages, in which case a large amount of inf mation is required to specify the traffic streams present in a network. This c be a serious handicap (see Section 3.4).

Stochastic Models and Flow Conservation. As shown in the p ceding discussion, analyzing the $GI/M/N/N$ queue is quite complex, especia when considering the description of the secondary system even when the inp is Poisson (Eqs. 3.12 and 3.18). We would like a simplified view that someh captures the essence of overflow and that can lead to simple numerical p cedures. Such a view is provided by the notion of network flows as found other areas of engineering, such as transportation or electrical flows. The m important feature of these models is that some form of conservation equati must be satisfied at the nodes of the network. The "thing" that is conserv at the nodes, and that thus flows in the network, is the mean of the arriv carried, and overflow processes. This concept can be seen by considering Kosten system, where some input enters the primary group, part of it accep and part of it diverted to the secondary group. If we identify the mean of three stochastic process in question — offered, carried, and overflow — w the magnitude of the flow, then it is not hard to see from Eqs. (3.49) and (3.5 that the standard conservation equation $A = \hat{A} + \overline{A}$ is verified. The notion offered traffic is crucial for this model since offered traffic represents the mag tude of the entering flow, though it has no significance in terms of actual c in the network.

This flow representation of the means of stochastic processes works well the Kosten system, and can be extended to the Brockmeyer system. Consi for simplicity, the case of Poisson input. Here the input flow is A, which sp into two parts in the ratio $E(A, N)$. The flow $A[1 - E(A, N)]$ goes into primary group, and the part $A[E(A, N)]$ becomes the input to the second group. There is now a slight difference between the Brockmeyer system the ordinary flow model. In the flow model, we would say that this flow sp into two parts, the part $AE(A, N)\,[1 - E\,(AE(A, N), L)]$ flowing onto the s ondary group, and the part $AE(A, N)\,[E\,(AE(A, N), L)]$ being rejected. T approach, however, does not coincide with the results (3.19) and (3.23) for Brockmeyer system. A better fit with the flow model is to suppose that

separation coefficients for two systems with overflow, such as for the Brockmeyer system, obey a combination law of the form $E_{N \otimes L} = E(A, N + L)$, where the notation $N \otimes L$ indicates a system where calls blocked on the primary group of size N are offered to the secondary group of size L. The flow carried on the secondary group is obtained as the difference between the offered and blocked flows. With this composition law, the Brockmeyer equations for the mean reduce to the conservation equations between offered, carried, and overflow traffic at both stages of overflow.

The analogy must not be carried too far; there are some difficulties with the flow model. Only the means of the stochastic processes can be approximated as conserved flows. Such a conservation rule does not apply to the second moment of the first overflow of the Kosten system, even in the case of Poisson input, since we have

$$V = A$$

$$\overline{V} = \overline{M} - A(N - \overline{M})E_N \tag{3.16}$$

$$\hat{V} = \hat{M}\left(1 - \hat{M} + \frac{A}{(N + 1 - A + \hat{M})}\right), \tag{3.14}$$

where it is obvious that there is no conservation of variance. Nevertheless, this conservation assumption is frequently made, if only to be able to use the conserved flow model. In these cases, one should always remember that using the flow model provides only an approximation, and that accuracy depends on the actual values of the parameters.

From a theoretical point of view, the passage from the stochastic description of the system to a flow model must be done very carefully. For another example of possible difficulties in setting up the appropriate definitions, consider the case of two links in series. We want to define the mean and variance of the traffic offered to the second link from the parameters of the process that represents the carried traffic on the first link. The usual procedure is to use the moments of the busy-circuit distribution on the first group as the moments of the offered traffic to the second group. This, however, is inconsistent with our definition of the offered traffic as the busy-server process on the infinite group, but with holding times that are chosen *independently* of the holding times of the calls in service in the finite group. This means that there are *two distinct* processes associated with carried traffic: (1) the busy-server process (BSP), whose state is the number of calls in service in the finite group, and (2) the carried-arrival process (CAP), whose state is the number of calls in progress in the infinite group. These processes are different: At a given time, the number of calls present in the two systems need not be identical since the durations of individual calls are different.

The peakedness of the BSP can be computed from Eqs. (3.15) and (3.16). As A becomes large, this peakedness understandably tends to 0 because the

number of busy servers in the group cannot exceed N; as the arrival rate increases, the number of busy servers will tend to N while the variations in the number of busy servers will tend to zero, since most servers are busy all the time.

We can define a peakedness for the CAP if we construct an infinite group in which a call is set up every time an event occurs in the CAP — that is, whenever a call is accepted — but where the holding time is generated from a negative exponential distribution with the same mean, but independently from the holding time of the real call. As the arrival rate increases, whenever a call terminates, a new call is accepted almost immediately. Intervals between call terminations are exponentially distributed, however, and the distribution of time intervals of the accepted calls tends to a negative exponential; the peakedness will go to 1. For this reason, CAP is not equivalent to BSP; the two processes cannot be used interchangeably. Since we have defined offered traffic in terms of the infinite group, we normally should use the moments of the CAP as the moments of the offered traffic. In practice, calculating these moments is quite difficult both theoretically and numerically, and great care must be taken to ensure stability. For these reasons, the CAP is not used in network-analysis algorithms. We refer the interested reader to [11].

We must make a final remark concerning these flow representations. In other areas of engineering, network flows entering a node can emerge in arbitrary proportions, the only condition being that the conservation equation must be satisfied. An example is transportation networks, where the flow pattern can be selected arbitrarily subject to the capacity constraints of the arcs. In the case of stochastic flows in the Kosten and Brockmeyer systems, this is not possible. The amount of carried traffic is determined by a number of factors, none of which is directly under the user's control. The first factor, of course, is the size of the two groups. The second is the stochastic description of the entering flow: If the arrival process is a renewal process, we need the full knowledge of $F(t)$; if it is not, the situation is even more complex since we need the transition matrix of a semi-Markov process. The fact that the links are selected in a precise order is another element that constrains the separation of the offered traffic into carried and overflow parts. This element is related to the routing technique used in the network; we will return to it in other chapters.

For all these reasons, a direct application of flow models to obtain an arbitrary separation of flows at the nodes is generally not very accurate as a model of alternate routing. Nevertheless, flow representation of stochastic processes is such a useful simplifying device that it is used in many situations, modified to take into account the particular nature of alternate routing. Now we see why traffic moments play such a crucial role in network analysis. The moments are the quantities that allow the transformation from a completely stochastic description of traffic, with all its complexity, to a more manageable,

but less accurate, representation in terms of deterministic vector flows, where this description is exact for the first moment, but only approximate for the others.

3.4 Moment-Matching Techniques for Overflow

Representing traffic by a small set of moments, although simpler than the full stochastic description, raises another problem. If we know only the first two moments of the traffic offered to a link, how can we compute the carried and overflow moments? One method, dating back to the introduction of alternate routing in telephone networks in the mid-1950s, is called the *moment-matching technique*. Let us now describe this class of methods and the particular problems raised by its use in network analysis.

A long-standing, much-studied problem in circuit switching is calculating traffic parameters subject to multiple overflows. This problem arises when a traffic stream, generally Poisson, is offered to some trunk group. The blocked calls are then offered to a second group, the calls blocked on this second choice are offered to a third one, and so on. The object is to compute the probability of not being able to make a connection on any of these groups, and to estimate the traffic carried on each one.

In principle, this problem is readily solved by repeated application of the techniques of Section 3.3, since the overflow process is also of the renewal type. In practice, this is not done for two reasons. Historically, the problem was studied, and partly solved, long before modern renewal theory was commonly used in teletraffic theory. Thus adequate methods already existed, making recourse to the exact renewal method somewhat redundant. Also, the renewal method is generally quite difficult to apply in practical cases because of numerical stability problems, long computation times, or both. For this reason, a discussion of the classical moment-matching systems is in order since these systems are the basis of all the techniques currently used in network analysis or synthesis.

Moment-matching techniques work as follows. The arrival process is represented by a small number of parameters, generally two or at most three. A process, called the *equivalent process*, is then selected to represent the actual arrival process. The parameters of the equivalent process are chosen in such a way that the moments of the traffic it generates are equal to the moments of the real offered traffic — hence the generic name moment-matching techniques. This equivalent process is then used to compute all the quantities of interest pertaining to the group: time and call congestion, moments of overflow and carried traffics, and so forth (see Fig. 3.2). The usefulness of this technique depends strongly on the possibility of choosing an equivalent process that yields accurate parameters for the overflow and carried traffics with reasonable computation times. Also, if the method is to be usable in network calculations, it must be reasonably easy to select the parameters of the equivalent process.

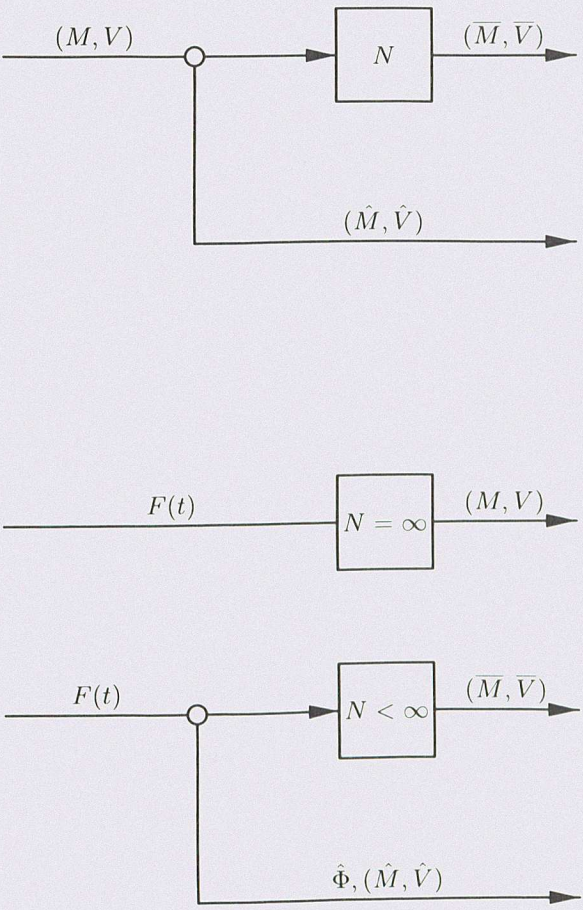

Figure 3.2 Moment-Matching Technique

Although not the only ones possible, renewal processes form an impo class of equivalent processes. The general theory of renewal processes is known, encompassing most of the equivalent methods currently used in net algorithms. For this reason, we review some of the more important proc of this type, only briefly indicating other types at the end of the discussi

Equivalent Random Theory

Equivalent random theory (ERT) is the first application of the moment-ma