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 $n$. – n servers are busy

$\overline{n}$ = 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 $A$.

$\alpha_q$, $\mu_q$, $M_q$, $\beta_q$ = The $q^{\text{th}}$ 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, $\beta_q^*$ denotes the $q^{\text{th}}$ 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 $\lambda$ 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 $\tau$ 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

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

$$= \frac{A^k/k!}{e_N(A)},$$

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,

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

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

$$\begin{aligned} M(n,z) &= \sum_{k=0}^{\infty} M_k(n)\frac{z^k}{k!} \\ &= \sum_{k=0}^{\infty} (1+z)^k p_{n,k} \end{aligned}$$

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

Secondary

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

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

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:

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

primary

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

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

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

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

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

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

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

$$\binom{n}{k} p^n q \longrightarrow E(x) = \frac{nq}{p}, \quad var(x) = \frac{nq}{p^2}$$

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 $\beta$ is independent of the state of the system. Assuming a Poisson arrival of rate $\lambda$, the BD model can be used with an arrival rate

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

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

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 $\lambda$ 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

$\lambda$ = 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

$$\begin{aligned} G(z) &= \sum_{k=0}^{\infty} p_k z^k \\ &= G_1(z) + G_2(z), \end{aligned}$$

where

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

and

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

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

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

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