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

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

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

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

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

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

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 rth 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 rth 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

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

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

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

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

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

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,

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

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

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

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 $\operatorname{Re} s \ge 0$,

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

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

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

Define

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

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

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

from which we can recover the transform

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

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

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

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 qth 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:

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

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

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

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 nth 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

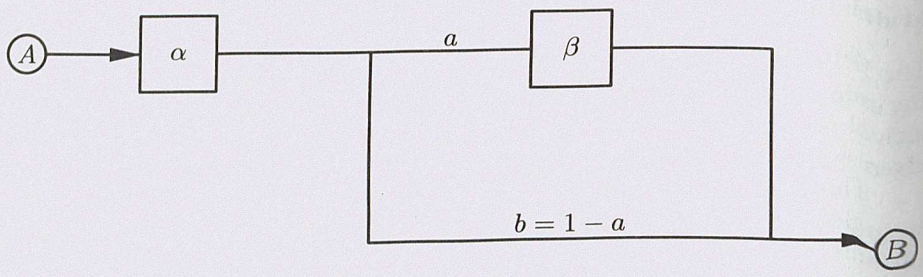

Figure 3.1 Equivalent System for Cox Model

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 $\Phi_N^1(s)$. Note that $\Phi_N^2(s)$ 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 $\hat{\Phi}_i^1(s)$ 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 $\hat{\Phi}_i^2(s)$ 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

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

This is a linear system in the unknown $\hat{\Phi}_i^1(s)$ and $\hat{\Phi}_i^2(s)$. 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 $\hat{\Phi}_i^1(s)$ and $\hat{\Phi}_i^2(s)$, and $A(s)$ is the matrix of coefficients. It is possible to verify that $dA(s)/ds = I$, the identity matrix. Thus, taking the nth 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 matrix, obtaining all the moments of the overflow distribution. These expressions are not particularly simple, for which reason we now end our discussion of the two-stage Cox model.

This model is a simple case of a more general class studied by Cox [16,3]. The n-stage Cox system is a straightforward extension of Fig. 3.1, where the n stages are placed in tandem, each having its own service time rate and probability of exit. Such a system can produce an arrival process having an arbitrary rational Laplace transform, subject to some mild constraints on the parameters. This generality in producing traffic streams is not particularly useful if the model has a large number of stages, in which case a large amount of information is required to specify the traffic streams present in a network. This can be a serious handicap (see Section 3.4).

Stochastic Models and Flow Conservation. As shown in the preceding discussion, analyzing the $GI/M/N/N$ queue is quite complex, especially when considering the description of the secondary system even when the input is Poisson (Eqs. 3.12 and 3.18). We would like a simplified view that somehow captures the essence of overflow and that can lead to simple numerical procedures. Such a view is provided by the notion of network flows as found in other areas of engineering, such as transportation or electrical flows. The most important feature of these models is that some form of conservation equation must be satisfied at the nodes of the network. The "thing" that is conserved at the nodes, and that thus flows in the network, is the mean of the arrival, carried, and overflow processes. This concept can be seen by considering the Kosten system, where some input enters the primary group, part of it accepted and part of it diverted to the secondary group. If we identify the mean of the three stochastic process in question — offered, carried, and overflow — with the magnitude of the flow, then it is not hard to see from Eqs. (3.49) and (3.53) that the standard conservation equation $A = \hat{A} + \overline{A}$ is verified. The notion of offered traffic is crucial for this model since offered traffic represents the magnitude of the entering flow, though it has no significance in terms of actual calls in the network.

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

separation coefficients for two systems with overflow, such as for the Brockmeyer system, obey a combination law of the form $E_{N \otimes L} = E(A, N \times L)$, 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.

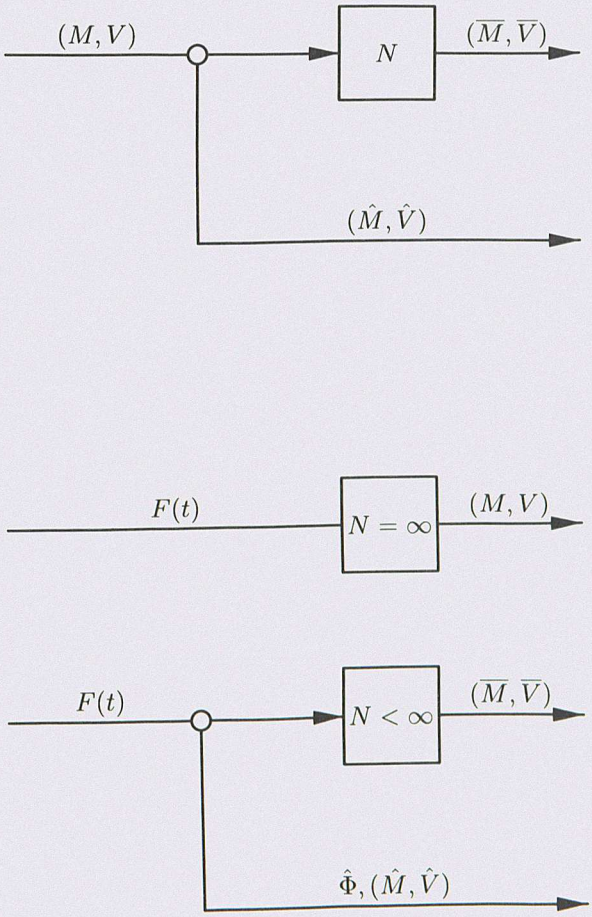

Figure 3.2 Moment-Matching Technique

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

Equivalent Random Theory

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

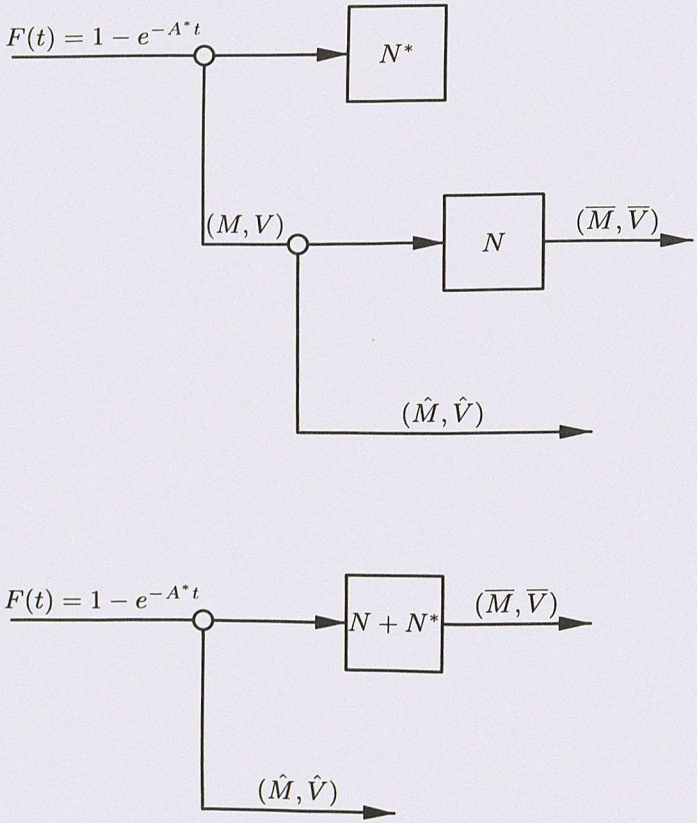

Figure 3.3 Equivalent Random Theory System

method to the analysis of overflow systems. As we have seen, the traffic overflowing a group offered Poisson traffic is always peaked. Thus it is reasonable to use the overflow model, where a primary group of N^* equivalent circuits offered some equivalent Poisson traffic A^*, as a generator of the (M, V) traffic whenever $V > M$. A^* and N^*, is chosen such that the parameters of its overflow are precisely M and V (see Fig. 3.3). From the Kosten model, we know that A^* and N^* must satisfy

$$M = A^* E(A^*, N^*) \tag{3.87}$$

$$V = M\left(1 - M + \frac{A^*}{N^* + 1 + M - A^*}\right) \tag{3.88}$$

Note that, in general, N^* is not an integer, which requires the use of a generalized Erlang B function. A fundamental question about the system defined by Eqs. (3.87) and (3.88) is whether there exist solutions $A^*, N^* \geq 0$ for an arbitrary pair M, V, with $V \geq M$. Although no formal proof seems to exist, this seems to be the case and the solution seems to be unique. Approximate values have been given by Rapp [2], which can either be used as such or as good starting values in the numerical solution:

$$A^* \approx V + 3Z(Z-1) \tag{3.89}$$

$$N^* \approx \frac{A^*(M+Z)}{M+Z-1} - M - 1. \tag{3.90}$$

Having computed the parameters of the equivalent system, we can compute the mean and variance of the overflow traffic using the Brockmeyer model; these values are given by

$$\hat{M} = A^* E(A^*, N + N^*) \tag{3.91}$$

$$\hat{V} = \hat{M}\left(1 - \hat{M} + \frac{A^*}{N + N^* + 1 - A^* + \hat{M}}\right). \tag{3.92}$$

The ERT method is iterative since it can be used to compute the blocking of the overflow traffic $(\hat{M}, \hat{V})$, which can also be peaked. Note, however, that the method is not iterative for the carried traffic, which can be smooth, and that it works only for peaked offered traffic. Thus the ERT method must be extended in order to be used for computing the overflow and carried parameters of carried traffic.

ERT for Smooth Traffic

In the case where $Z < 1$, the equations that define the equivalent system (3.87) and (3.88) generally have a negative solution for N^*. Possible extensions to standard ERT can be made along three directions by relaxing various parameters of the model: (1) use an equivalent traffic source that is smoother than Poisson, and match the moments of its overflow with the real parameters, (2) extend the Erlang B formula to the full domain of real N, or (3) generate the real smooth traffic by a mechanism other than overflow.

The first technique, although not widely used, falls within the framework of equivalent renewal processes and is dealt with below. The other methods have been proposed and used for network algorithms, and thus are described briefly here.

The second technique — the possibility of using an extended Erlang B function to construct an equivalent system for smooth traffic — was proposed first by Bretschneider [17]. The idea was extended by Nightingale [18] under the name of ERT-N, based on the analytic continuation of the Erlang B function

in the complete (A, N) plane, as shown in [5]. For this case, only the extension to negative N is used; the extension to negative A has not been used in any practical case. The application of the method is straightforward. Given the parameters of the real traffic, the equivalent system is computed by solving the standard ERT equations (3.87) and (3.88); the parameters of the overflow are computed via Eqs. (3.91) and (3.92), where now N^* is negative if the traffic is smooth. Note that the equivalent system must still have nonnegative A. This, however, is not always possible. The derivation of the conditions in which the equivalent traffic is positive is left as an exercise (Problem 3.18).

The third way to extend ERT to smooth traffic is to assume that the mechanism producing the real traffic is something other than overflow. Such a model has been proposed by Katz [19] in which the real offered traffic is produced by the portion of the equivalent traffic that is *carried* in the equivalent group. The Laplace transform of the distribution of the interarrival time is given by Eq. (3.55). We know, however, that this process is not renewal and that its parameters cannot be computed using the general renewal method. The parameters of the equivalent system are given implicitly by the system:

$$M = A^* \left[1 - E(A^*, N^*)\right] \tag{3.93}$$

$$V = M - AB(N^* - M) \tag{3.94}$$

Having determined the value of the equivalent system, the parameters of the traffic overflowing the real group are computed as follows:

1. If $N^* \leq N$, then all calls carried in the equivalent system are also carried in the real group. Thus, we get $\hat{M} = \hat{V} = 0$; that is, there is no overflow from the real group.

2. If $N^* > N$, then some of the calls carried in the equivalent group are blocked in the real group. Assuming that the traffic offered to the real group is effectively A^*, we have for the parameters of the traffic carried in the real system

$$\overline{M} = A^* \left[1 - E(A^*, N)\right]$$
$$\overline{V} = \overline{M} + (\overline{M} - A^*)(N - \overline{M})$$

The parameters of the overflow traffic are computed by

$$\hat{M} = A^* - \overline{M}$$
$$\hat{V} = V - \overline{V}$$

Note that computing the variance in this way neglects the correlation that exists between the carried and overflow traffic on the real system.

It is claimed that in practice the accuracy of the method is comparable to the accuracy of the equivalent random method for peaked traffic and thus is adequate to evaluate network performance.

Generalized ERT

As mentioned in the preceding section, ERT (now standing for equivalent *renewal* theory) can be generalized to an arbitrary renewal input process by replacing the Poisson input by some arbitrary renewal stream. The moment matching is done between the traffic parameters available, on the one hand, and the moments of the renewal stream and the size of the equivalent group N^*, on the other. The blocking probabilities and overflow and carried traffic parameters are then computed for the equivalent renewal stream offered to a group of size $N + N^*$.

Such a generalized ERT was proposed by Potter [20], who suggested using input streams with various Erlang$_k$ interarrival distributions as input to the primary group for modeling smooth traffic. The equivalent streams have an interarrival time distribution of the E_n type, where $n = 1, 2, 3, 6, 10, \infty$. These streams have some interesting properties. As k increases, the peakedness of the overflow traffic decreases, with its smallest value attained for the deterministic stream. In fact, it has been shown [13] that $\hat{Z} \geq 0.5$ for any k. Thus these input processes cannot be used to model traffic whose peakedness is below 0.5. For a given phase k, the equivalent system has only two parameters: (1) the intensity of the Erlang stream and (2) the size of the equivalent group. Thus the model can be used directly for a two-moment representation of the traffic. Tables have been produced that relate the parameters of the overflow traffic of the equivalent group to these two parameters, allowing the equivalent system to be sized. No practical use of the technique, however, seems to exist in real network applications.

Equivalent IPP Model for Peaked Traffic

Consider again a peaked traffic (M, V). Such a traffic could be generated by an interrupted Poisson process [9,21], suitably chosen as to reproduce the values of M and V. Note, however, that the IPP is a three-parameter model, which means that the traffic should be characterized by a third parameter in addition to the mean and variance.

Let us first consider a somewhat simpler problem. Assume for the moment that the IPP is in fact produced by the overflow of some Poisson traffic A^* from a finite group of size N^*. In this case, the parameters of the IPP can be expressed as a function of the parameters of the equivalent system as

$$\lambda = A^* \frac{\delta_2(\delta_1 - \delta_0) - \delta_0(\delta_2 - \delta_1)}{(\delta_1 - \delta_0) - (\delta_2 - \delta_1)} \tag{3.95}$$

$$\omega = \frac{\delta_0}{\lambda} \left(\frac{\lambda - A^* \delta_1}{\delta_1 - \delta_0} \right) \tag{3.96}$$

$$\gamma = \frac{\omega}{A^*} \left(\frac{\lambda - A^* \delta_0}{\delta_0} \right) \tag{3.97}$$

where

$$\delta_n = \frac{1}{A^*}\frac{M_{n+1}}{M_n}, \tag{3.98}$$

where the factorial moments M_n can be computed by the Kosten model of Section 3.1. Given an arbitrary value for A^* and N^*, there is no proof that the corresponding λ, ω, γ are not negative (although this seems to be the case in practice). The same question can be posed for an arbitrary set of three moments with $Z \geq 1$.

This relation between the moments of the IPP process and the overflow of a Poisson stream is useful when the real traffic is described by its first two moments only. The difficulty is that the IPP is a three-parameter process, and that there are only two quantities with which to match it. In this case, the IPP model can be used by making the following approximations. First, given M and V, compute $A^*$and N^* by the standard ERT. Then use these values to obtain the parameters of the IPP system by Eqs. (3.95–3.97), and use Eq. (3.72) to obtain the mean and variance of the overflow. Another possibility is to use the equivalent Poisson traffic computed from ERT as the value for λ. In this case, only two parameters are left to compute, given by

$$\omega = \frac{M}{A^*}\left(\frac{A^* - M}{Z - 1} - 1\right) \tag{3.99}$$

$$\gamma = \omega\left(\frac{A^*}{M} - 1\right)$$

In general, when a two-moment match is desired, one parameter of the IPP must be chosen arbitrarily. In practice, this choice is constrained by the requirement that the parameters of the equivalent IPP system should be non-negative, which imposes the requirement that $\lambda > M$. Finally, the method is recursive, since overflow traffic is always peaked, and the IPP method can be used again on this traffic. This is not the case for the carried traffic, which is generally smooth; some other method must be used to calculate its overflow.

The Two-Stage Cox Model

The Cox model of Section 3.3 can be used to represent a stream of traffic whose first three moments are known, provided the system of Eqs. (3.81) and (3.82) can be inverted. This question is discussed in detail by Guérineau and Labetoulle in [22], in which the following conditions are given. Let $P = \alpha\beta$ and $S = \alpha + \beta$. These can be written in terms of the coefficients δ_0, δ_1 and δ_2 as

$$P = \frac{2\mu^2\delta_0(\delta_2 - \delta_1)}{2\delta_1 - \delta_0 - \delta_2}$$

$$S = \frac{\mu\,[2(\delta_1 - \delta_0 + 1)(\delta_2 - \delta_1) + \delta_1(2\delta_1 - \delta_0 - \delta_2)]}{2\delta_1 - \delta_0 - \delta_2}$$

and of course we have

$$\alpha = \frac{S + \sqrt{S^2 - 4P}}{2}$$

$$\beta = \frac{S - \sqrt{S^2 - 4P}}{2}$$

$$b = \frac{P}{\mu\delta_0\alpha}$$

Obviously, there exist solutions if and only if

$$P > 0,\ S > 0,\ S^2 - 4P > 0,\ a > 0.$$

For peaky traffic, a necessary and sufficient condition is that $\delta_0 < \delta_1 < \delta_2 < 2\delta_1 - \delta_0$. In the case of smooth traffic, the condition $2\delta_1 - \delta_0 < \delta_2 < \delta_1 < \delta_0$ is necessary but not sufficient.

Because the parameters of the real offered traffic are computed from the superposition of various overflow and carried traffic streams, quite possibly the set of three moments does not meet the required conditions. In such a case, some further approximation is required. If the traffic is peaky, the first two moments can always be matched. Since the equivalent system is a three-parameter model, one parameter is arbitrary. Guérineau and Labetoulle suggest choosing $\mu\delta_1 < S < \infty$, with a recommended value of $S = 2\mu\delta_1$. Having chosen S, the parameters for the equivalent Cox generator are given by

$$\alpha = \frac{S + \sqrt{S^2 - 4P}}{2}$$

$$\beta = \frac{S - \sqrt{S^2 - 4P}}{2}$$

$$b = \frac{1}{\alpha}(S - P/\mu\delta_0)$$

where

$$P = \frac{\mu\delta_0}{\delta_1 - \delta_0 + 1}(S - \mu\delta_1).$$

In the case of smooth traffic, one first tries the same values as for peaky traffic. If the parameters of the generator are positive, then a solution has been obtained. If not, only the first two moments are matched. In order to do this, α and β are kept arbitrary, and a is fixed to 1. The parameters are then

$$\alpha = \frac{S + \sqrt{S^2 - 4P}}{2}$$

$$\beta = \frac{S - \sqrt{S^2 - 4P}}{2}$$

which is valid if and only if

$$\delta_1 \geq \frac{4\delta_0^2}{1+4\delta_0}.$$

This means that smooth traffic can be modeled by the Cox system if it is no smoother than Erlang$_2$. When this is not so, the authors suggest replacing it with an Erlang$_2$ traffic and using the Cox model. In this case, we have $\alpha = \beta = 2\mu\delta_0$.

After constructing the equivalent system, we must calculate the parameters of the carried and overflow traffics. This can be done by means of the transform of the overflow process from the primary group, using either the method proposed in [22] or the formulas derived in Section 3.3. It is not clear which of these two methods is preferable from the point of view of numerical stability and computation time.

The Bernouilli-Poisson-Pascal Process

As is well known [6], the overflow traffic offered to a large group has a busy-circuit distribution that is very close to a Pascal distribution. Given the mean and variance of a peaked traffic, the parameters of the corresponding Pascal distribution can be computed analytically from Eqs. (3.27) and (3.28). From the values of the parameters, the overflow could be analyzed just as for the Kosten system, with the arrival rate corresponding to a Pascal process instead of the constant rate of the Poisson process. Note, however, that the analysis would become quite complex — probably the reason why it has not been carried out.

Instead, the Bernouilli-Poisson-Pascal (BPP) method can be used to approximate the overflow system by making a second assumption [23]. Given an offered traffic with arbitrary mean and variance, *the busy-circuit distribution in a group of size N is given by the same distribution used to describe the offered traffic*, that is, the Pascal distribution (or Bernouilli distribution for smooth traffic). For a group of finite size N, the distribution (3.25) must be truncated; with $\mu = 1$, we have the truncated Pascal distribution, denoted by q_j:

$$q_j = \frac{p_j}{\sum_{i=0}^{N} p_i}, \quad j = 0, \ldots N$$

The time congestion for a group of size N is given by

$$E_N \triangleq q_N = \frac{\binom{-r}{N}(-\beta)^N}{\sum_{j=0}^{N}\binom{-r}{j}(-\beta)^j} \tag{3.100}$$

and the call congestion by

$$B = E\left[1 + \frac{N}{M}(Z-1)\right]. \tag{3.101}$$

Equation (3.101) should be compared with the general relation (3.48) between time and call congestion in the case of a general renewal input. The time congestion obeys the interesting recurrence

$$E_k = \frac{q_{r+k-1}E_{k-1}}{k + q_{r+k-1}E_{k-1}} \quad k = 1, \ldots \tag{3.102}$$
$$E_0 = 1$$

which is a generalization of the well-known recurrence for the Erlang B function. Similarly, under this approximation, the mean and variance of the traffic carried in the group are given by

$$\overline{M} = \frac{nq - (n+N)qE}{1-q} \tag{3.103}$$

$$\overline{V} = \frac{nq + \overline{M}(n+1)q - (N+1)(n+N)qE}{1-q} - \overline{M}^2 \tag{3.104}$$

From these values, we can get $\hat{M} = M - \overline{M}$ and, neglecting the correlation, $\hat{V} = V - \overline{V}$. It is easy to show that these quantities reduce to the values already calculated for the Kosten system, that is, when $q \to 0$ and $n \to \infty$ while maintaining $Z = 1$.

As indicated in Problem 3.4, the BPP distribution is inadequate to describe smooth offered traffic because the probabilities given by the binomial distribution could become negative. The situation is better in the case of a finite group since a sufficient condition exists to guarantee that the probabilities do not become negative in the range of interest. That is, for $j \leq N$, $N \leq M/(1-Z)$, a condition generally met in practical situations unless Z is very low. Although the method cannot represent all potential situations, the advantages of simplicity and uniformity in representing both peaked and smooth traffics make it attractive for modeling traffic in network studies.

Finally, let us comment on the accuracy of the method. The accuracy was checked in [23] by comparing the parameters of the carried traffic for a hyperexponential and a gamma distribution of the interarrival times. It was shown numerically that the values were similar for the two cases and were quite good. A strong assumption is needed for the BPP model, stating that the busy-circuit distribution is BPP for any arbitrary arrival process. In fact, the situation is even worse from a theoretical point of view, since it can be shown that there is *no* renewal input process such that its overflow has the Pascal distribution [13]. These two facts would seem to make the BPP model a very poor candidate indeed for the analysis of overflow. The accuracy of this model, however, is at least as good as that of the other models. Let us explain this apparent anomaly in the more general context of renewal arrival processes and moment-matching methods.

General Renewal Input

In general, when a traffic is represented by a renewal process, specifying a number of moments of the real traffic constrains the input process and to some degree determines the blocking experienced by this process. Moment-matching methods would be expected to work well if specifying the moments of the real traffic limits the equivalent renewal process to a small range. In this case, we would expect the moment-matching technique to be "forced" to choose the right equivalent process, which would explain the good accuracy obtained by moment matching.

From Eqs. (3.53) and (3.54), the values of M and V uniquely determine $\Phi'(0)$ and $\Phi(\mu)$ for *any* renewal input. In fact, these values constrain all of the $\Phi(\mu)$ since they must satisfy Eq. (3.52), where $\Phi(\mu)$ appears in all the terms through h_n. Bounds have been computed in [24] for $\Phi(x), x \geq 1$ as

$$[\Phi(\mu)]^x \leq \Phi(x) \leq \Phi_m(x) \triangleq p_1 + p_2 e^{-bx},$$

where b is defined implicitly by the equation

$$b = \frac{A(1-e^{-b})}{1-\Phi(\mu)}$$

and

$$p_2 = \frac{A}{b}$$
$$p_1 = 1 - p_2$$

Furthermore, from Eq. (3.46), we see that $\Phi(\mu)$ is the leading term in the expression for the call congestion B although B also depends on all the $\Phi(\mu)$s. Using the bounds for Φ, it is possible to write bounds for the blocking experienced by an arbitrary renewal stream $B_{min} \leq B \leq B_{max}$ where

$$B_{min} = \left\{1 + \binom{N}{1}\frac{1-\Phi(\mu)}{\Phi(\mu)} + \ldots + \binom{N}{N}\frac{[1-\Phi(\mu)]\ldots[1-\Phi^N(\mu)]}{[\Phi(\mu)]^{N(N+1)/2}}\right\}^{-1}$$

$$B_{max} = \left\{1 + \binom{N}{1}\frac{1-\Phi_m(\mu)}{\Phi_m(\mu)} + \ldots + \binom{N}{N}\frac{[1-\Phi_m(\mu)]\ldots[1-\Phi_m(N\mu)]}{\Phi_m(\mu)\ldots\Phi_m(N\mu)}\right\}^{-1}$$

The point is that these bounds are sharp, and it is possible to identify renewal processes that reach them. The lower bound is given by a renewal input with constant interarrival times m', given by $\exp\{-m'\} = \Phi(\mu)$. Similarly, a renewal process with a step at time $t \to 0$ and another at $t \to \infty$ has a blocking probability arbitrarily close to 1. The upper bound can be viewed as such a process, but constrained by V to have the second step at finite t. The point is that the two bounds can be widely separated, as shown in [24]. This means that the blind selection of a renewal process to represent a traffic specified by only two moments cannot be expected to yield an accurate value for the blocking.

The excellent accuracy of moment matching has been explained in the case of the BPP method [23]. Assume that the system is divided into K subsystems of infinite size, and let p_i be the probability that an incoming call is served by the ith subsystem. Let X_i be the number of calls in subsystem i, and M_i and V_i its mean and variance. We then have

$$X = \sum_i X_i, \; M_i = Mp_i$$
$$V = M_i p_i (1 + p_i(Z-1))$$
$$\text{cov}(X_i, X - X_i) = Mp_i(1-p_i)(Z-1)$$

If A_i is the interarrival time distribution at subsystem i, its Laplace-Stieltjes transform is given by

$$a_i(s) = \frac{a(s)\,(1-(1-p_i))}{1-(1-p_i)a(s)}.$$

Choosing $p_i = 1/K$, we can choose K such that M_i/V_i is as close to 1 as desired. We also note that $a_i(s)$ tends to a negative exponential distribution as $p_i \to 0$. The mean carried traffic can then be viewed as the sum of a large number of infinitesimal quasi-regular traffics; this property is independent of the precise nature of the arrival renewal process. Because this microscopic structure of the stream is the important factor in the particular value of the mean, we can expect the BPP model to adequately describe the carried traffic in an infinite group, irrespective of the particular nature of the arrival renewal process. This is indeed the case.

More generally, equivalent methods are found to work well when used to evaluate overflow systems. The reason for this success is found not in the constraints imposed on the equivalent process by the prescribed values for the mean and variance, but in the additional information used in selecting the equivalent process. In other words, equivalent theory works because, when it is applied to overflow systems, the fact that the M, V traffic is produced by an overflow mechanism is used to select the equivalent renewal process. This is obvious when considering that the most successful equivalent systems for peaked traffic are the Poisson overflow and the IPP. Similarly, the success of the BPP method is also due to the choice of the correct distribution to represent carried traffic; it would probably be not as successful if an arbitrary renewal input stream with the correct M, V parameters were used. This is only a conjecture, however, and for now we only note that, for the problems arising in circuit-switched studies, the renewal equivalent method is generally sufficient.

Summary

Moment-matching techniques are widely used in analyzing networks. In fact, it is fair to say that these techniques are the only tools available that are rapid

enough to be used in this context. Even so, the situation is not completely satisfactory, at least for certain problems. Network analysis really requires a technique that can represent *any* kind of traffic, smooth or peaked, within the same model, and for *arbitrary* values of the parameters. None of the methods we have discussed here meets these requirements; all are designed for a particular type of traffic, smooth or peaked — or, if they apply to both, do so by using different models for different ranges of peakedness.

As a case in point, consider the BPP model. Although probably the most attractive from the point of view of generality, it suffers from some defects. It cannot represents arbitrarily smooth traffic, and it represents smooth and peaky traffics by two separate models. Suppose that we want to use this model for network analysis or synthesis within a computer program. It is true that, in a real network, traffic will not be so smooth that the model cannot apply. This, however, is not the case in the design process, during which a search algorithm may produce some extreme traffic patterns when it is far from its optimal solution. Some care must be taken that these values do not cause a fatal error in the program. Similarly, the nature of the traffic offered to a given link may change, say from peaky to smooth, during this search procedure. Because the BPP model uses two different distributions for these two cases, the parameters of the overflow and carried traffic may not change continuously when this transition occurs — or, more likely, their derivatives may not be smooth. If the search algorithm uses gradients, these discontinuities in the derivative may cause the procedure to fail.

To avoid these difficulties, given that there exists no unified model that is guaranteed to be smooth and robust in terms of the parameters of the offered traffic, caution should be exercised when using moment-matching models in optimization procedures.

3.5 Heterogeneous Blocking

The traffic models described in Sections 3.1 to 3.4 apply to a single stream of offered traffic. Unfortunately, real networks are somewhat more complex since the traffic offered to a group is generally a mixture of streams, each with its own set of moments, and is not of the renewal type. Because of their different statistical characteristics, the streams effectively experience different blocking probabilities.

The heterogeneous blocking problem can thus be formulated as follows. Given n distinct streams of traffic offered to a group of size N, each with a known set of moments, compute the blocking probability (*parcel* blocking), as well as the carried and overflow traffic moments, for each stream. In this general form, the problem has no known solution, even under the assumption that the input streams are independent; only approximate methods are available. The

problem with two streams has been solved exactly when one of the streams is Poisson and the other is renewal. The case of two renewal streams has also been analyzed, but simplifications are required to obtain a solution. For this reason, we concentrate here on practical methods that can yield sufficiently accurate results with relatively little computation.

First, we discuss the aggregation of the input streams and the assumption of independence. Then, we describe one case where an exact solution is known, mostly to illustrate the difficulty of the problem. Finally, heuristics usable in network problems are examined, starting with empirical formulas and moving to progressively more complex models based on those of Section 3.4.

Aggregation of Input Streams

The heterogeneous blocking problem arises because calls offered to a group in a network have been either blocked or accepted for service on some other group, but still must be connected through some intermediate group in order to establish a connection to their destination. Depending on the structure of the network and the routing method used, these processes cannot be expected to be independent. The dependence between various streams of traffic offered to a single group was investigated in [25] for simple arrangements of overflowing groups. Such dependence effects are quite difficult to compute; in practice, it is assumed that the streams to be merged are independent. This is a good approximation in large networks, where many different streams overflow on a given link (see Chapter 2).

Assuming this independence, and also assuming that stream i has k^{th} moment α_k^i, the kth moment of the aggregate stream is given by

$$\alpha_k = \sum_{k_1+k_2+\ldots k_m=1} \frac{k!}{k_1!k_2!\ldots k_m!}\alpha_{k_1}^1\alpha_{k_2}^2\ldots\alpha_{k_m}^m. \tag{3.105}$$

In the case of $m = 2$, we have

$$\begin{aligned}
\alpha_1 &= \alpha_1^1 + \alpha_1^2 \\
\alpha_2 &= \alpha_2^1 + 2\alpha_1^1\alpha_1^2 + \alpha_2^2 \\
\alpha_3 &= \alpha_3^1 + 3\alpha_2^1\alpha_1^2 + 3\alpha_1^1\alpha_2^2 + \alpha_3^2
\end{aligned}$$

Given the moments of the aggregate stream, its blocking probability and the parameters of the carried and overflow traffic can be computed by the methods of Section 3.4. The difficult part is determining how much of this blocking is actually experienced by each stream.

Heterogeneous Blocking for Poisson Inputs

The heterogeneous blocking problem has been solved in the case where all the input streams are Poisson. In this case, all streams experience the same

blocking, and it is possible to characterize the individual streams overflowing in a common infinite group as follows. Let a^i, $i = 1, \ldots n$ be the parameters of each Poisson stream offered to a common group of size $N < \infty$, and let p_i be the probability that a given overflow call originates from stream i. Because all streams experience the same blocking probability, $p_i = a^i/a$, where a is the mean of the aggregate stream and is given by $a = \sum_i a^i$. Assume that, at a given instant, there are m calls in progress in the overflow group. The joint conditional probability of having $k_1, \ldots k_n$ calls of stream $1, \ldots n$, given that there are m calls in progress, is given by

$$p(k_1, \ldots k_n \mid m) = \frac{m!}{k_1! \ldots k_n!} p_1^{k_1} \ldots p_n^{k_n}, \tag{3.106}$$

where

$$m = k_1 + k_2 + \ldots k_n.$$

The jth conditional moment of the ith overflow stream α_j^i is given by

$$\alpha_j^i = \sum_{k_1, \ldots k_m, m} k_i^j p(k_1, \ldots k_n \mid m) p_m,$$

where of course $p(m)$ is the probability of having m calls in progress. This completely characterizes the overflow traffic. The first few moments are given by

$$\alpha_1^i = p_i \alpha_1 \tag{3.107}$$

$$\alpha_2^i = p_i^2 \alpha_2 + p_i(1 - p_i)\alpha_1 \tag{3.108}$$

$$\alpha_3^i = p_i^3 \alpha_3 + 3p_i^2(1 - p_i)\alpha_2 + (1 - 2p_i)p_i(1 - p_i)\alpha_1 \tag{3.109}$$

Note that these overflow traffics are not Poisson, as expected. Thus *their* blocking cannot be computed with this method; a more general technique is required.

Heterogeneous Blocking for Two Streams

Unfortunately, the traffic that overflows a primary group is not Poisson; since it becomes mixed with other streams, the real situation requires a more detailed analysis. In the simplest case, one overflow stream is combined with a number of background streams, all of which are Poisson. All the background streams can be aggregated into a single, equivalent Poisson stream, reducing the problem to a two-stream problem with one overflow stream and one Poisson stream. This problem dates back to the work of Wilkinson [6], where the magnitude of the heterogeneous blocking is estimated for the Brockmeyer system, that is, when the renewal stream is a Poisson overflow process. Assuming that the secondary group of the system is offered a Poisson traffic of given magnitude in addition to

the overflow traffic, we can compute the blocking experienced by this Poisson stream B_1 and compare it with B_2, the overall blocking experienced by the aggregate traffic. B_1 can be computed from Eq. (3.21) since the call congestion for a Poisson traffic is just the marginal time congestion of the secondary group. B_2, on the other hand, is just the overall blocking probability of the system. Depending on the system parameters, the ratio between these two probabilities can be quite high, sometimes more than 10. This effect seems to be more pronounced for small groups, typically five circuits, but is still not negligible at large traffic values. For instance, for a traffic of 30 Erlangs in 15 circuits, the ratio is still almost 2.5, which should not be neglected.

An exact solution to the heterogeneous blocking problem was given by Kuczura [26] for the case of one Poisson and one general renewal streams offered to the same group. For the renewal stream, the call congestion is given by the busy-server distribution of the embedded Markov chain, which we shall now compute. The state probabilities of the system are obtained by solving the Kolmogorov equations, using the arrival times of the renewal calls as regeneration points. The Kolmogorov equations for the transition function $P_{i,j}(t)$ are

$$\begin{aligned} P'_{i,0}(t) &= -\lambda P_{i,0}(t) + \mu P_{i,1}(t) \\ P'_{i,j}(t) &= -\lambda P_{i,j-1}(t) - (\lambda + j\mu)P_{i,j}(t) + (j+1)\mu P_{i,j+1}(t) \\ P'_{i,N}(t) &= -\lambda P_{i,N-1}(t) - N\mu P_{i,N}(t) \\ P_{i,j}(0) &= \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{3.110}$$

The system of equations has the unique solution [27]

$$P_{i,j}(t) = a_j + \sum_{k=1}^{N} a_k(i,j)e^{\lambda_k \mu t},$$

where

$$\begin{aligned} \rho &= \frac{\lambda}{\mu} \\ a_j &= \frac{\rho^j}{j!}\frac{1}{\sum_{k=0}^{N}\frac{\rho^k}{k!}} \\ a_k(i,j) &= \frac{N!}{j!}\frac{\rho^{N-i}D_i(\lambda_k)D_j(\lambda_k)}{\lambda_k D_N(\lambda_k)D'_N(\lambda_k+1)} \\ D_i(x) &= \sum_{k=0}^{i}\binom{i}{k}\rho^{i-k}(x)^k \\ xD'_N(x+1) &= \sum_{k=0}^{N}\rho^N(N)_k\frac{D^2_{N-k}(x)}{D_N(x)} \\ D_0(x) &= 1 \end{aligned}$$

and where λ_k are the N roots of the polynomial $D_N(x)$. Of course, the knowledge of $P_{i,j}$ determines the embedded Markov process completely. More important, however, is knowledge of the stationary distribution of calls in the group and the congestion probabilities. As usual, let p_j be the stationary distribution just before an arrival. The one-step transition probability $r_{i,j}$ is then

$$r_{i,j} = \int_0^\infty P_{i+1,j}(t)dF(t), \quad 0 \le i < N, \quad 0 \le j < N$$

$$r_{N,j} = \int_0^\infty P_{N,j}(t)dF(t), \quad 0 \le j < N$$

Replacing the value of $P_{i,j}$ from Eq. (3.110), we obtain

$$r_{i,j} = a_j + \sum_{k=1}^{N} a_k(i+1,j)\phi(-\lambda_k\mu) \quad 0 \le i < N, \quad 0 \le j < N$$

$$r_{N,j} = a_j + \sum_{k=1}^{N} a_k(N,j)\phi(-\lambda_k\mu), \quad 0 \le j < N$$

The stationary distribution can be found as usual by solving the linear system

$$p_j = \sum_{i=0}^{N} p_i r_{i,j}$$

$$\sum_j p_j = 1$$

The call blocking seen by the renewal stream is given by P_N since the busy-circuit distribution as seen by this stream is the one for the Markov chain.

The call blocking for the Poisson stream can be obtained by using the memoryless property of the Poisson process. In this case, the busy-circuit distribution as seen by this process is the distribution at an arbitrary point in time q_j. It can be obtained by writing the equilibrium equation for the rates in and out of the state, and is expressed as usual in terms of the stationary probabilities for the embedded chain

$$(j+1)\mu q_{j+1} = \nu p_j + \lambda q_j,$$

which can be solved as

$$q_j = q_0 \frac{\rho^j}{j!} + \sigma \sum_{i=0}^{j-1} \left(\frac{i!}{j!}\right) \rho^{j-i-1} p_i,$$

where

$$\sigma = \frac{\nu}{\mu}$$

and

$$q_0 = \left[1 - \left(\frac{\sigma}{\rho}\right) \sum_{j=1}^{N} \left(\frac{\rho^j}{j!}\right) \sum_{i=0}^{j-1} \left(\frac{i!}{\rho^i}\right) p_i\right] \Big/ \sum_{j=0}^{N} \left(\frac{\rho^j}{j!}\right).$$

The time congestion is then given by q_N, which is also the call congestion for the Poisson stream.

At this point, let us consider the calculation of the overflow and carried-traffic parameters. From a practical point of view, that is, for use in network applications, this calculation is probably not very useful, considering the computations required by the technique just to evaluate blocking. All the roots of a Nth degree polynomial must be calculated in order to arrive at the transition matrix of the embedded Markov chain. After the transition matrix is computed, the stationary probabilities are computed by solving the linear system of order N. Care must be exercised in determining the roots of the polynomial in order to avoid numerical instability, as discussed in [26] in some depth. Similarly, the numerical solution of Markov chains is not always easy, especially for large systems, where great care must be taken to avoid propagating numerical errors while solving the linear system. For this reason, the calculation of the mean and variance of the carried and overflow traffics would probably be too difficult to use in a network context.

The heterogeneous blocking problem has been solved by Manfield and Downs, albeit approximately, in the case of two arbitrary renewal input streams [28]. Formulas are proposed for the moments of the overflow and carried streams, but these formulas are even more complex than in the case where one stream is Poisson; this complexity seems to preclude the use of these formulas in network algorithms. The formulas have been used to suggest approximate formulas based on the analysis of the two-stream case.

Let $F_1(t)$ and $F_2(t)$ be the interarrival time distribution of the two streams, let $\phi_i(s)$, $i = 1, 2$ be the respective Laplace-Stieltjes transforms of the corresponding density functions, and let λ_i equal the corresponding intensity. Also let $p_i = \lambda_i/(\lambda_1 + \lambda_2)$. The expected number of blocked calls during a period of congestion is given by

$$E^1 = \frac{p_1\phi_1(N\mu)}{1 - \phi_1(N\mu)} + \frac{p_2\lambda_1}{N\mu} \tag{3.111}$$

$$E^2 = \frac{p_2\phi_2(N\mu)}{1 - \phi_2(N\mu)} + \frac{p_1\lambda_2}{N\mu} \tag{3.112}$$

The distribution of congestion periods is quite complex, and Manfield and Downs make the simplification that the probability that a congestion period starts with an arrival of type i is given by p_i. They also assume that the interarrival time distribution is produced by an IPP process. Thus we have

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

and the corresponding transforms,

$$\phi_i(s) = \frac{k_i^1 r_i}{s + r_i} + \frac{k_i^2 r_i^2}{s + r_i}, \quad i = 1, 2.$$

From this, we get the congestion probability for the two streams by replacing in Eqs. (3.111) and (3.112). This can be extended to more than two streams by taking

$$E^i = \frac{p_i \phi_i(N\mu)}{1 - \phi_i(N\mu)} + \frac{(1 - p_i)\lambda_i}{N\mu}, \tag{3.113}$$

where

$$p_i = \frac{\lambda_i}{\sum_j \lambda_j}, \tag{3.114}$$

which is an approximation for the individual stream-congestion functions.

Approximate Methods

The work on two-stream blocking clearly indicates the need to partition the blocking probability between the various streams in a reasonable manner, especially when designing and analyzing networks. Although the Brockmeyer system can be used for Poisson stream blocking, in general, one would like to partition the blocking among an arbitrary number of non-Poisson streams, either peaked or smooth. Furthermore, as shown by the work of Kuczura and Manfield and Downs, exact methods based on the solution of the Kolmogorov equations call for too much computation to be usable for network problems. For these reasons, approximate methods are required.

The simplest approximations are based on a partition of the blocking derived from the values of the first two moments of the traffic streams. First we examine some simple empirical formulas based on numerical studies. Then we describe some partitioning models based on the techniques described in Section 3.4, with special emphasis on the generalization of the BPP and Cox models to the heterogeneous case.

Empirical Formulas. These techniques assume that the aggregate blocking for all the streams has been computed, for instance, by one of the methods of Section 3.4. This probability is then partitioned between the streams using some reasonable formula, with the condition that the sum of the parcel overflows is equal to the aggregated overflow.

Let B be the aggregate blocking probability and B^k be the blocking for stream k. An early method to estimate the stream blocking was proposed by Katz [19] in the case of peaked traffic:

$$B^k = B\left[1 + K\left(\frac{V^k}{M^k} - \frac{V}{M}\right)\right]. \tag{3.115}$$

Although this method is adequate for a given link, the constant K is traffic dependent and must be adjusted for each link. The actual form of K is given by

$$K \approx \frac{M}{V}\left[4.1 \times 10^{-4}\left(\frac{M}{V}\right)\left(\frac{1}{B}\right)^{2} - 0.156\left(\frac{M}{V}\right)\left(\frac{1}{B}\right)^{1/2}\right.$$
$$\left. -0.0747\left(\frac{1}{B}\right) + 0.69\left(\frac{1}{B}\right)^{1/2}\right] - 0.183, \tag{3.116}$$

where the coefficients were found by regression over simulation results of trunk groups in networks. The method was somewhat improved by Butto, Colombo, and Tonietti [29], and apparently independently by Deschamps [30], who gave an analytic approximation for the constant K when one of the streams is Poisson. The estimate of K is made by using Eq. (3.115) for the Poisson stream, for which we know that the variance is equal to the mean. For this parcel, B is given by the group time congestion of the corresponding Brockmeyer system as given by Eq. (3.21). The only remaining unknown in Eq. (3.115) is K, which can easily be computed. In cases where the resulting parcel blocking is not in the interval $(0, 1)$, the value is set at either 0 or 1, depending on whether the result was larger than 1 or smaller than 0.

Another empirical formula based on numerical experiments was proposed by Kuczura [21], who suggested

$$B^k = B\frac{M^k Z^k}{\sum_j M^j Z^j}, \tag{3.117}$$

where M^k is the mean of the kth input stream. This formula partitions the flow in proportion to the variances instead of the means. The author claims that it is more accurate than Katz's formula, based again on numerical results.

Finally, we should also mention yet another formula proposed by Olsson as described in [13], where

$$B^i = B\left(\frac{M}{M^i}\right)\frac{V^i + \dfrac{{M^i}^2}{V^i}}{\displaystyle\sum_j V^j + \frac{{M^j}^2}{V^j}}. \tag{3.118}$$

Clearly, it is possible to produce a rather large number of similar formulas, based on both intuition and numerical studies. Needed instead, however, is a more systematic way to generate such formulas, ideally in relation to the single-stream models of Section 3.4. Recent work on heterogeneous blocking takes this direction, suggesting models based on the BPP technique, the IPP model, and the Cox generator.

The BPP Approximation. A systematic procedure for computing the stream-carried traffic, based on an application of the BPP technique to the individual streams, has been proposed by Delbrouck [23]. From the joint distribution of busy circuits, a convolution operation is performed over $n-1$ streams to get the marginal distribution for a given stream, from which the moments of the carried traffic can be computed. Let X_i be the number of calls of stream i present in the group. The basic assumption of the method is that this distribution is represented approximately by the truncated BPP distribution that satisfies the recurrence

$$p_i(x_i+1) = \frac{\alpha_i + \beta_i x_i}{x_i} p_i(x_i), \tag{3.119}$$

which is obtained from Eq. (3.25) when applied to stream i, and where α_i and β_i are computed from the stream mean and peakedness by Eqs. (3.27) and (3.28). Next we assume that the joint occupation probability has the product form

$$p(x_1, x_2 \ldots x_n) = K \prod_{i=1}^{N} p_i(x_i), \tag{3.120}$$

where K is the appropriate normalization constant

$$K^{-1} = \sum_{x_1+x_2\ldots x_n \le N} p_1(x_1)p_2(x_2)\ldots p_n(x_n).$$

Given this joint distribution, we want to compute the moments of the carried traffic for each stream. This can be done in several ways consistent with Eqs. (3.119) and (3.120), which in turn determine which overflow parameters are available, as well as the computational requirements.

Convolution Method. This method uses the marginal distribution of each stream, obtained by taking the convolution of the joint distribution over all but the ith stream. The marginal distribution is given by

$$\theta_i(x) = \sum p(x_1, \ldots x, \ldots x_n), \tag{3.121}$$

from which we can get the carried traffic and variance

$$\overline{M}^i = \sum_{x=1}^{N} x\theta_i(x) \tag{3.122}$$

$$\overline{V}^i = \sum_{x=1}^{N} x^2\theta_i(x_i)$$

where the sum is carried over $x_1 + \ldots x_{i-1} + x_{i+1} + \ldots x_n \le N - x$, that is, over all streams but stream i. The call congestion is computed directly from the difference between offered and carried traffics:

$$B^i = \frac{M^i - \overline{M}^i}{M^i}. \tag{3.123}$$

From the marginal distribution (3.121), we can get all the moments of the carried traffic of each stream. Note, however, that the technique implies a large amount of computation since each distribution is the result of a convolution over all the other streams in the group.

A simpler way to arrive at the individual congestion is to apply the call-congestion formula of Eq. (3.101) to each stream, replacing Z and M by Z^i and M^i. This would give

$$B^i = E\left(1 + \frac{N}{M}(Z^i - 1)\right). \tag{3.124}$$

This partition formula, proposed by Delbrouck [31], is called the *BPP2 formula*. This formulation has a number of advantages over the convolution method. The sum of the individual blocked traffic is identical to the value of the overall blocked traffic as computed from Eq. (3.101). This method is also much faster than the convolutions, but does not give the second moments of either the carried or the overflow traffic. Also, the formula overestimates the blocking for parcels having an individual peakedness greater than the total peakedness ($Z^i \geq Z$), and underestimates the blocking for the other parcels.

Another formula proposed to correct this, called the *BPP3 formula*, is written

$$C_i \propto E\left[1 + \frac{N}{M}\left(\left(\frac{V^i}{M^i}\right)^\gamma - 1\right)\right], \tag{3.125}$$

where γ is a constant that must be determined. The equation is only a proportionality since the sum of the parcel losses on the right-hand side does not add up to the total loss computed from the total blocking probability. The formula can be renormalized, so that the partition formula becomes

$$B^i = CE\left[1 + \frac{N}{M}\left(\left(\frac{V^i}{M^i}\right)^\gamma - 1\right)\right], \tag{3.126}$$

where

$$C^{-1} = \sum_i C_i,$$

which now satisfies

$$BM = \sum_i B_i M_i.$$

In practice, the value of the parameter γ is not important, and values in the range $0.4 \leq \gamma \leq 0.5$ appear to give satisfactory accuracy.

Recursive Method. The defects of the convolution method have been corrected in [31], where a recursive procedure is proposed for computing the time congestion of each stream and the first two moments of the carried traffic. The recursion is over N, the number of circuits in the group. Let

$m_i^j(N)$ = The ith moment of the distribution of calls from stream j that are present in the group. This is the carried traffic if $N < \infty$, and the offered traffic if $N = \infty$.

$m_1^{j*}(N)$ = The product of the time congestion with the average of X_j given that all N trunks are busy. In a sense, this measures the amount of traffic that could have been carried during the time when the group was busy. This is another measure of blocked traffic, but from the point of view of time congestion.

We also define

$$\begin{aligned} M^j &= m_1^j(\infty) \\ v^j(N) &= m_2^j(N) - (m_1^j(N))^2 \\ V^j &= v^j(\infty) \end{aligned}$$

M^j and V^j have the usual interpretation of the mean and variance of the offered traffic of stream j, while m_1 and v have the same meaning for the carried traffic. E_N is the time congestion in the group of size N. The moments can be calculated recursively for groups of increasing size n:

$$\begin{aligned} E_n &= \frac{\sum_{i=1}^k Z_i^{-1}\hat{a}^i(n-1)}{n + \sum_{k=1}^k Z_i^{-1}\hat{a}^i(n-1)} \\ m_1^j(n) &= (1-E_n)\left[\alpha_j + \beta_j m_1^j(n-1)\right] \\ m_1^{j*}(n) &= (1-E_n)\left[\alpha_j E_{n-1} + \beta_j m_1^{j*}(n-1)\right] \\ m_2^j(n) &= (1-E_n)\left[(\alpha_j+\beta_j)m_1^j(n-1) + \beta_j m_2^j(n-1) + \alpha_j\right] \\ \hat{a}^j(n) &= M^j - m_1^j(n-1) \end{aligned}$$

with the initial values

$$E_0 = 1, \quad m_i^j(0) = m_1^j(0) = 0, \quad \hat{a}^j(0) = M_j$$

The recurrence, solved from $n = 1$ to N, gives the time congestion and the first two moments of the carried traffic for stream i.

Overflow Parameters. In addition to the mean of the overflow traffic $\hat{a}^j(N)$, we need the call congestion for each offered stream and the variance of the overflow streams. Although not directly available from the recurrence, these parameters can be estimated using the standard ERT method for each stream (see Fig. 3.4). Assume that there exists a fictitious primary group of size N^* to which a Poisson stream of intensity A^j is offered. Calls blocked on the primary group are then offered to the real group, which we will call the secondary group. Because we know that the traffic offered to this secondary

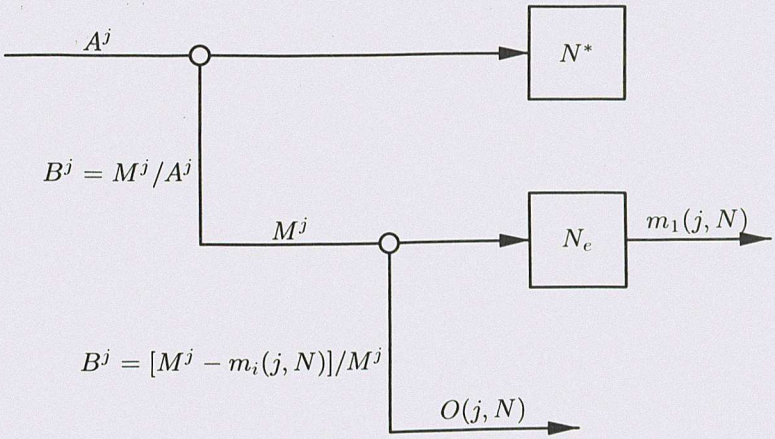

Figure 3.4 ERT Model for Stream Call Congestion

group has parameters M^j and V^j, the value of the equivalent Poisson traffic can be computed by Rapp's approximation,

$$A^j = M^j Z^j + 3Z^j(Z^j - 1),$$

and the call congestion in the primary group is given by M^j/A^j. Suppose now that this traffic is offered to a fictitious secondary group of size N_e that *carries exclusively calls from stream* j. N_e is chosen such that the traffic it carries is precisely $m_1^j(N)$. Obviously, $N_e \neq N$. Calls offered to this fictitious secondary group are blocked with a probability $\hat{a}^j(N)/M^j$. Hence the call congestion for stream j on the real group is approximated by

$$B^j = \frac{M^j - m_1^j(N)}{M^j}.$$

The overflow variance $\hat{v}^j$ for stream j can be computed with this equivalent system by noting that $\hat{a}^j(N)$ is the overflow of a Poisson traffic A^j in a group of size $N_e + N^*$. These sizes need not be computed since the overflow traffic is already known. The value of the offered and overflow traffics being known, Rapp's approximation prescribes the value of the peakedness of the overflow $\hat{Z}^j$ through the equation

$$A^j = \hat{a}^j(N)\hat{Z}^j + 3\hat{Z}^j(\hat{Z}^j - 1),$$

which in turn yields the variance

$$\hat{v}^j = \hat{Z}^j \hat{a}^j(N),$$

which is the required expression for the variance of the overflow. It has also been noted by Delbrouck [31] that the individual stream call congestion can be rewritten

$$B^j = E_N \left(1 + \frac{m_1^{j\,*}(N) E_n^{-1}}{M^j}(Z^j - 1)\right).$$

Compare with Eq. (3.124). This in turn suggests the approximation

$$\frac{E_N M^j}{\sum_l M_l} = \frac{m_1^{j\,*}(N)}{N}, \tag{3.127}$$

which has a straightforward interpretation. As pointed out by Delbrouck, this equation can easily be incorrect for particular values of traffic. Thus, although it has a strong intuitive appeal and seems well verified in most cases, it is not clear under what conditions Eq. (3.127) should be considered correct.

Akimaru's Approximation

Another partition formula, proposed by Akimaru and Takahashi [32], is based on the moment match of the overflow traffic to the IPP model. First consider Eq. (3.48), which links the time and call congestion for a renewal input stream offered to a group of finite size. Assume that this renewal stream is made up of the superposition of a Poisson stream and an IPP stream, and also that the parameters of the IPP stream are matched to the parameters of some overflow traffic. In this case, Eq. (3.48) becomes

$$\frac{B}{E} = 1 + (Z - 1)\frac{N}{M}\frac{M + 3Z}{N - 1 + M + 3Z}, \tag{3.128}$$

where M and Z are the mean and peakedness of the aggregate stream. Let B^1 and B^2 be the call congestion of the Poisson and IPP streams, respectively. Because the Poisson stream is memoryless, its call congestion is identical to the overall call congestion B, as noted by Wilkinson [6]. Using this fact, we can rewrite Eq. (3.128) as

$$B^1 = \left[1 + (Z - 1)\frac{N}{M}\frac{M + 3Z}{N - 1 + M + 3Z}\right]^{-1} B. \tag{3.129}$$

We can eliminate B from this equation by using the conservation equation $a^2 B^2 + a^1 B^1 = aB$, obtaining

$$\frac{B^2}{B^1} = 1 + (Z^2 - 1)\left(\frac{N}{M}\right)\frac{M + 3Z}{N - 1 + M + 3Z}. \tag{3.130}$$

We generalize to an arbitrary number of streams by

$$B_k = \frac{1 + (Z_k - 1) f(N, A, Z)}{1 + (Z - 1) f(N, A, Z)} B, \tag{3.131}$$

where

$$f(N, A, Z) = \frac{N(3Z + A)}{A(N - 1 + A + 3Z)}.$$

The total call congestion is computed by any one of the methods of Section 3.4, and then partitioned according to Eq. (3.128). Note that this method does not yield information on the second moment of the carried or overflow traffic.

The Cox Model for Heterogeneous Blocking

Finally, we briefly discuss the Cox model as a method for computing parcel blocking [15]. First consider the case where two renewal streams are offered to a common trunk group. To be able to use the Cox generator we represent the input streams by their first three moments, from which the parameters of the two generators are computed. Define a busy period as an interval during which all the circuits of the group are busy. Parcel blocking is calculated on the system where the sources are the equivalent Cox generator.

Let E_i, $i = 1, 2$ be the expected number of calls from stream i that are blocked during a busy period, and p_i the probability that a blocked call belongs to stream i. Obviously we have $p_i = E_i/(E_1 + E_2)$. We can calculate the parcel-blocking probabilities as follows. First we distinguish busy periods as being of type 1 or type 2, depending on whether the call that has caused the saturation belongs to stream 1 or stream 2. Let $E_{i,j}$ be the expected number of blocked calls of stream i, given that stream i is in state j. Recall that the state of the stream is given by the state of the associated Cox generator since we have replaced the actual renewal process by the generator via the appropriate moment match.

Then, for $i = 1$, we have

$$E_{1,1} = \frac{\alpha_1 b_1}{\alpha_1 + N\mu}(1 + E_{1,1}) + \frac{\alpha_1 a_1}{\alpha_1 + N\mu} E_{1,2}$$

$$E_{1,2} = \frac{\beta_1}{\beta_1 + N\mu}(1 + E_{1,1})$$

and similar equations for stream 2. These equations are derived by considering the possible states of the generator and the probabilities of being in these states. Suppose that the generator is in the first stage. In this case, the system has $N + 1$ exponential servers, one of rate α_1 (the customer in the first stage), and N servers at rate μ. The Cox server may end service first, with probability $\alpha_1/(\alpha_1 + N\mu)$; a new call is then generated with probability b_1, or the generator stays in state 1, which produces the term $(1 + E_{1,1})$. Or the generator may enter state 2, with probability $a_1\alpha_1/(\alpha_1 + N\mu)$, which gives the term $E_{1,2}$, or the busy period may end, with probability $N\mu/(\alpha_1 + N\mu)$. The combination of these events yields the expression for $E_{1,1}$, and a similar argument yields the

corresponding equation for $E_{1.2}$. The solution to the system of equations is

$$E_{1.1} = \frac{\alpha_1(\beta_1 + b_1 N\mu)}{N\mu(\alpha_1 a_1 + \beta_1 + N\mu)}$$
$$E_{1.2} = \frac{\beta_1(\alpha_1 + N\mu)}{N\mu(\alpha_1 a_1 + \beta_1 + N\mu)}$$

We must now consider whether the current busy period is of type i, that is, whether it is caused by a call of stream i. Let r_i be the probability that it is of type i. We can assume that this probability is proportional to the arrival rate of the two streams; we get

$$r_1 = \frac{\alpha_1\beta_1}{\beta_1 + a_1\alpha_1} \Bigg/ \left[\frac{\alpha_1\beta_1}{\beta_1 + a_1\alpha_1} + \frac{\alpha_2\beta_2}{\beta_2 + a_2\alpha_2}\right]$$
$$r_2 = \frac{\alpha_2\beta_2}{\beta_2 + a_2\alpha_2} \Bigg/ \left[\frac{\alpha_1\beta_1}{\beta_1 + a_1\alpha_1} + \frac{\alpha_2\beta_2}{\beta_2 + a_2\alpha_2}\right]$$

Also assume that the system is in a stationary state when a busy period begins, and let q_1 be the probability that stream 1 is in state 1 when a busy period of type 2 starts. Then,

$$q_1 = \frac{\beta_1}{\beta_1 + a_1\alpha_1}.$$

From this,

$$E_1 = r_1 E_{1,1} + r_2 \left[q_1 E_{1,1} + (1 - q_1)E_{1,2}\right].$$

The corresponding equation for E_2 can be obtained in the same manner.

The solution of an arbitrary number of streams m is computed by solving m two-stream problems. One stream is singled out for calculation, and the remaining $m - 1$ streams are aggregated into a single stream. The assumption here is that this aggregated stream is also of the renewal type. We calculate the overflow traffic for the stream of interest by the methods just described, repeating the process m times to get the mean of the overflow traffic for all the input streams.

It is also possible to calculate the transform of the individual overflow processes by generalizing the results of Section 3.3, although the results cannot be used to evaluate network performance because of the computational requirements. Thus we give only a brief outline of the procedure, which is not fundamentally different from the one used in calculating the overflow and carried parameters for a single stream.

First we must define the transforms for four types of intervals. Assume that the group is in state i at the current time. We are interested in the interval until the next overflow from stream 1. There are four conditions under which this can happen, coming from the product of the current state, denoted by $k = 1, 2$

of generator $m = 1, 2$. We then have four transforms $\hat{\Phi}_i^{(k,m)}(s)$ to compute. We write the Kolmogorov equations for these functions, using an argument similar to that of Section 3.3, obtaining a system of order $4(N+1)$. Here, too, taking the rth derivative and evaluating it at zero gives the moments of all the distributions of the overflow. Because we are concentrating on stream 1, we are interested in the overflow distribution $\hat{\phi}(s)$ caused by this stream. Letting P_2 be the probability that the generator for stream 2 is in state 1 immediately after an overflow from stream 1, we have

$$\hat{\Phi}(s) = \hat{\Phi}^{(1,1)} P_2 + \hat{\Phi}^{(1,2)}(1 - P_2),$$

where

$$P_2 = \frac{\alpha_1 b_1 p_N(1,1) + \beta_1 p_N(2,1)}{\alpha_1 b_1 p_N(1,1) + \beta_1 p_N(2,1) + \alpha_1 b_1 p_N(1,2) + \beta_1 p_N(2,2)},$$

where $p_k(i,j)$ is the probability distribution corresponding to $\hat{\Phi}^{(i,j)}(s)$. Clearly the amount of computation can become quite large for a given link, and the total computation time required for the analysis of a network precludes the use of this method except in the simplest cases. Nevertheless, such a model can be useful to check the accuracy of the approximations that must be used in network analysis.

3.6 Computational Considerations and Accuracy

The traffic models discussed in the previous sections are those mentioned most frequently in the literature on circuit-switched networks. These models represent only a small part of the work on the subject; the number of actual and potential models that can be generated from the birth-and-death and renewal techniques is quite large. A natural question is which of these models should be used in network problems. This is a complex question to which there is no definite answer. It is possible, however, to state some considerations in selecting a traffic model to implement on a computer and use in the automated design and analysis of networks. Computation time, robustness, regularity, and accuracy are the more important factors, in that order, and a few words of advice are offered to the unwary on each of these topics. (Manual design using charts is a different matter, and is not discussed here, since traffic models suitable for one method may or may not be suitable for the other.)

Computation Time

The time required to evaluate the blocking and the parameters of the traffic streams is of primary importance in network problems. Unfortunately, this topic is not discussed often in the teletraffic literature. This omission is understandable when the sole purpose of the exercise is to compute the performance

of a single isolated system, perhaps to study the effect of one parameter, or to evaluate the accuracy of an approximate method on a small system.

The picture changes dramatically if we envisage using computer implementations of these methods for network study. In such a case, we must evaluate blocking probabilities hundreds, if not thousands of times for a single network evaluation. Any method, however accurate, that requires long computation times, must be rejected, since the total computation time would quickly become prohibitive for a network of any significant size. Similar remarks hold for the amount of storage required by a model. It may be useful to evaluate a group using a five-moment match, as has been proposed, but the amount of data required in the case of a moderately large network certainly makes the method impractical.

Robustness

Robustness, also an important aspect, especially for computer implementation, is related to the finite size of numbers that can be represented on computers and to the consequently limited accuracy in their representation. A robust implementation must meet at least three requirements: (1) It should yield meaningful results even when its input takes extreme values, (2) the computation algorithm must not be subject to fatal errors in evaluating intermediate results, and (3) no significant errors must be introduced by the calculation process.

Input may take on extreme values when the traffic algorithms are used within other design or evaluation procedures that may be searching for solutions well outside the expected range of parameter values. For instance, an Erlang B program, when asked to compute $E(A = 1, N = 200)$, should return a value of zero, not some cryptic message about overflow or underflow. Although any human designer would immediately know the correct answer, the computer implementation must be carefully designed to avoid meaningless error messages or fatal errors at execution time. This same remark holds for moment-matching methods, where two methods that are equally fast may turn out to be quite different from the point of view of robustness. A good example is the calculation of equivalent systems for smooth traffic, which may not exist for certain values of the parameters. A robust implementation should be able to recognize this fact and arrive at a sensible solution, instead of blindly going into a loop and terminating with an answer of dubious value. For this reason, equivalent systems that are known to exist for all values of the parameters are superior to those that are known not to exist in some ranges (e.g., ERT-N) — or worse, to those where it is not known under what conditions the equivalent system exists.

The second aspect of robustness concerns the actual steps to compute a value. This is better explained by way of a trivial example. Suppose we

want to compute $p_j = e^{-A}A^j/j!$ for the Poisson model. To make this more precise, let us take $A = 100$ and $j = 100$. Note that the numerical value of this probability is quite reasonable, about 0.513296. A naive implementation would compute each term separately and then take the product. This, of course, would not work, since the evaluation of 100! would probably exceed the largest number that can be represented on the computer, causing an overflow error. A more astute implementation would recognize the fact that p_j can be computed recursively, since

$$p_j = p_{j-1}A/j \tag{3.132}$$

The proper normalization would be done at the end by summing all terms and dividing the ps by this sum. Of course, in the present case, we know what the normalization constant should be. In general, we do not have such an expression; we can only guess what the value of p_0 could be and then renormalize after the calculation is completed. This approach would be somewhat better but would not meet the requirement for robustness. This is because of the form $A^k/k!$; for large-enough A, the numerator will grow initially much faster than the factorial, and the iteration term may become too large before the denominator has time to catch up.

Nevertheless, it is possible to get a very robust implementation in one of two ways. For the particular case of the Poisson formula, we first compute the $\log p_j = j \log A - \log \Gamma(j+1)$, using a specialized algorithm to evaluate the logarithm of the Gamma function directly without the computation of the factorial itself. If the logarithm is larger than the logarithm of the smallest number available, the value of p_j is given by $p_j = e^{\log p_j}$. Otherwise, a value of zero is returned. Another robust implementation, which can be applied in other contexts, is based on the fact that the sequence p_j typically increases as j increases, goes by a maximum when j is nearest A, and then decreases monotonically to zero. The recurrence (3.132) is then computed in two directions. Choose the starting value $m = [A]$, the integer that is closest to A. The general step is

$$\begin{aligned} p_m &= 1 \\ p_k &= p_{k-1}\frac{A}{k} \quad k \geq m \\ p_{k-1} &= p_k\frac{k}{A} \quad k < m \end{aligned}$$

In other words, solve the recurrence in two separate loops — for increasing and decreasing values of the index, respectively — in such a way that the terms are monotone decreasing on each side of the initial point. In this way, all the significant terms are included, and no underflow error can occur if each recurrence is stopped when the term becomes sufficiently small. The correct values are then obtained by normalization as usual.

This small example shows how the numerical calculation of traffic models can be very different from their analytic formulation. In other cases, an even more drastic reformulation must be made in order to prevent round-off errors and ensure the stability of the computation, which is the third aspect of robustness mentioned previously.

Consider again the traffic model with throttled arrivals, which is defined for integer values of N and m. This model is used in network-optimization algorithms, which are often evaluated over the reals. For this reason, we want to extend the model to fractional values for both N and m, much in the same way that the Erlang B function can be extended to fractional values. A simple extension is first presented and shown to be numerically unstable over some values of the parameters. A recursive approach is then explained that is both robust and efficient for an integer number of reserved trunks and arbitrary parameters [33]. The model is finally extended to arbitrary trunk numbers, and an efficient and robust algorithm is proposed.

Recall the equations for the model:

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

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

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

We want to extend Eqs.(3.29–3.31) to real m and N. Recall the definition of the incomplete exponential function,

$$e_m(x) = \sum_{j=0}^{m} \frac{x^j}{j!}, \tag{3.133}$$

which is used to express Eqs. (3.31) and (3.30):

$$p_0^{-1} = e_m(a) + (1-r)^{-m}\left[e_N(\overline{a}) - e_m(\overline{a})\right] \tag{3.134}$$

$$B = p_0 \frac{a^N}{N!}(1-r)^{N-m} \tag{3.135}$$

$$Q' = p_0 e_{m-1}(a) \tag{3.136}$$

We obtain the required generalization by noting that

$$E(a,m) = \frac{a^m/m!}{e_m(a)}, \tag{3.137}$$

where $E(\cdot,\cdot)$ is the Erlang B function, for which we have a generalization to nonnegative reals, and $N!$ is extended to the reals by the Gamma function. In

what follows, $N!$ should be understood to mean $\Gamma(N+1)$ whenever N is not integer. We can express Eqs. (3.134–3.136) for fractional values as

$$p_0^{-1} = \frac{a^m/m!}{E(a,m)} + (1-r)^{-m}\left[\frac{\overline{a}^N/N!}{E(\overline{a},N)} - \frac{\overline{a}^m/m!}{E(\overline{a},m)}\right] \tag{3.138}$$

$$B = p_0 \frac{a^N}{N!}(1-r)^{N-m} \tag{3.139}$$

$$Q' = p_0 \frac{a^{m-1}/(m-1)!}{E(a,m-1)} \tag{3.140}$$

For $0 < m < 1$, Eq. (3.140) is written as

$$Q' = p_0 \frac{a^m}{m!}\left(\frac{1}{E(a,m)} - 1\right).$$

Eqs. (3.138–3.140) are the required generalizations for arbitrary nonnegative m and N. Mathematically, they give the correct value of the probabilities for any argument. In practice, however, the form of the equations is such that severe numerical difficulties can arise if we want to carry out the calculation as indicated. This is best understood by looking at Eq. (3.138). This quantity requires the computation of

$$\left[\frac{\overline{a}^N/N!}{E(\overline{a},N)} - \frac{\overline{a}^m/m!}{E(\overline{a},m)}\right], \tag{3.141}$$

where both terms may be large but close to each other; since we must compute their differences, large round-off errors may occur that would make the final answer virtually useless.

A general solution could be to use a program for computing the Erlang B function with very high accuracy. This would have the effect of slowing down the computation in all cases, even those where it is not required. As we shall now show, we can rearrange the calculation to avoid numerical instabilities in all cases of interest. From a mathematical point of view, Eqs. (3.138–3.140) are perfectly adequate. The massive reformulation described below is required only because computers can do arithmetic only to a limited accuracy.

Recursive Formulation. First we give an expression for the blocking probabilities for integer $N-m$ but real N and m. In itself, this expression can be used when the number of reserved trunks is a fixed integer. The algorithm proposed is a recursion on the number of reserved trunks. We can write

$$\frac{1}{B(a,r,m,N+1)} = \frac{\displaystyle\sum_{k=0}^{m}\frac{a^k}{k!} + \sum_{k=m+1}^{N}\frac{a^k(1-r)^{k-m}}{k!} + \frac{a^{N+1}(1-r)^{N+1-m}}{(N+1)!}}{\displaystyle\frac{a(1-r)}{(N+1)}\frac{a^N(1-r)^{N-m}}{N!}}$$

$$= \frac{(N+1)}{\overline{a}}\frac{1}{B(a,r,m,N)} + 1, \tag{3.142}$$

and also for Q',

$$\frac{1}{Q'(a,r,m,N+1)} = \frac{\displaystyle\sum_{k=0}^{m}\frac{a^k}{k!} + \sum_{k=m+1}^{N}\frac{a^k(1-r)^{k-m}}{k!} + \frac{a^{N+1}(1-r)^{N+1-m}}{(N+1)!}}{\displaystyle\sum_{j=0}^{m-1}\frac{a^j}{j!}}$$

$$= \frac{1}{Q'(a,r,m,N)} + \frac{a^{N+1}}{(N+1)!}\frac{(1-r)^{N+1-m}}{\displaystyle\sum_{j=0}^{m-1}\frac{a^j}{j!}}. \tag{3.143}$$

Using the well-known recursion for the Erlang function,

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

we obtain the following expression for the last term:

$$\frac{a^{N+1}}{(N+1)!}\frac{(1-r)^{N+1-m}}{\displaystyle\sum_{j=0}^{m-1}\frac{a^j}{j!}} = \frac{a^{N+1-m}(1-r)^{N+1-m}}{(N+1)N\ldots(m+1)}\frac{E(a,m)}{1-E(a,m)}.$$

Replacing in Eq. (3.143), we get

$$\frac{1}{Q'(a,r,m,N+1)} = \frac{1}{Q'(a,r,m,N)} + \frac{a^{N+1-m}(1-r)^{N+1-m}}{(N+1)N\ldots(m+1)}\frac{E(a,m)}{1-E(a,m)}, \tag{3.144}$$

which is the required recursion. The initial values are given by

$$B(a,r,m,m) = E(a,m) \tag{3.145}$$

$$Q'(a,r,m,m) = 1 - E(a,m) \tag{3.146}$$

This technique is particularly efficient when the number of reserved trunks is small, as is often the case. The total amount of work required is then approximately equal to that required to compute the Erlang B function for Eqs. (3.145) and (3.146), with a few additional iterations.

Fractional Number of Reserved Trunks. The recursion with integer $N-m$ is not adequate for network optimization, where both N and m may vary. Thus an extension is also required for noninteger $N-m$.

A robust method has been proposed in [33] for this case, based on the model developed for integer $N-m$. We extend recursions (3.142) and (3.144) to a fraction $0 < \delta < 1$ above N. Using Eqs. (3.138–3.140) and (3.137),

we have

$$\frac{1}{B(a,r,m,N+\delta)} = \frac{e_m(a) + (1-r)^{-m}\left[e_{N+\delta}(\overline{a}) - e_m(\overline{a})\right]}{\dfrac{\overline{a}^{N+\delta}}{(N+\delta)!}(1-r)^{-m}}$$

$$= \frac{(N+\delta)!}{N!}\frac{1}{\overline{a}^{\delta}}\frac{1}{B(a,r,m,N)}$$

$$+ \frac{(N+\delta)!}{\overline{a}^{N+\delta}}\left[e_{N+\delta}(\overline{a}) - e_N(\overline{a})\right] \tag{3.147}$$

and also

$$\frac{1}{Q'(a,r,m,N+\delta)} = \frac{e_m(a)}{e_{m-1}(a)} + \frac{(1-r)^{-m}}{e_{m-1}(a)}\left[e_{N+\delta}(\overline{a}) - e_m(\overline{a})\right]$$

$$= \frac{e_m(a)}{e_{m-1}(a)} + \frac{(1-r)^{-m}}{e_{m-1}(a)}\left[e_{N+\delta}(\overline{a}) - e_N(\overline{a}) + e_N(\overline{a}) - e_m(\overline{a})\right]$$

$$= \frac{1}{Q'(a,r,m,N)} + \frac{(1-r)^{-m}}{e_{m-1}(a)}\left[e_{N+\delta}(\overline{a}) - e_N(\overline{a})\right]. \tag{3.148}$$

Let

$$S_N(\overline{a}) = \frac{(N+\delta)!}{\overline{a}^{N+\delta}}\left[e_{N+\delta}(\overline{a}) - e_N(\overline{a})\right]. \tag{3.149}$$

As we see, we can compute Eqs. (3.147) and (3.148) by Eqs. (3.142) and (3.144) to the nearest $N-m$ integer, expressing the correction term of Eqs. (3.147) and (3.148) in terms of Eq. (3.149). It remains only to obtain Eq. (3.149) for arbitrary values of a, r, m, and N. Incidentally, we can easily check that, if $0 < \delta < 1$, then $0 < S_N < 1$. In particular, if $\delta = 1$, we get back Eqs. (3.142) and (3.144) for an integer number of reserved trunks, as we should.

Computation of the Correction Term. We now give an expression for Eq. (3.149) that is robust for all values of traffic and trunk numbers. Using Eq. (3.137), we get

$$S_N(a) = \frac{1}{E(a,N+\delta)} - \frac{(N+\delta)!}{N!}\frac{1}{a^{\delta}}\frac{1}{E(a,N)}. \tag{3.150}$$

We encounter the same difficulty as with Eq. (3.141). If a/N is small, we must compute the difference of two large quantities that differ by a small amount, causing large numerical errors.

Thus we must use a technique less sensitive to these errors. First we consider the case $a \le N$, since in the other case we can compute Eq. (3.150) directly with good accuracy. From Jagerman [5], we know that

$$E(a,N)^{-1} = \psi(a,N)^{-1} - \sum_{j=1}^{\infty}\frac{a^j}{(N+1)(N+2)\ldots(N+j)} \tag{3.151}$$

where

$$\psi(a, N) = e^{-a} \frac{a^N}{\Gamma(N+1)}$$

We use Eq. (3.151) to compute both terms of Eq. (3.150) simultaneously. We get

$$S_N(a) = \sum_{j=1}^{\infty} \frac{a^j}{(N+\delta+1)(N+\delta+2)\dots(N+\delta+j)} \times \left[\frac{1}{a^\delta}\frac{(N+\delta+j)!}{(N+j)!} - 1\right], \tag{3.152}$$

a convergent series with positive terms, which can be computed with excellent accuracy whenever $a \leq N$, and also when $a \leq 1$.

Summary. The point of this discussion is that although we may have a mathematically sufficient formulation of a traffic model, in the present case, expressed by Eqs. (3.139) and (3.140), the numerical evaluation may require some rather profound transformations to be numerically robust. We can summarize the computation as follows. Define

$$N_r = N - m$$
$$K = \lceil N_r \rceil$$
$$\delta = N_r - K$$

1. Compute $B(a, r, m, N-\delta)$ and $q(a, r, m, N-\delta)$ by Eqs. (3.142) and (3.144). If $\delta = 0$ we have an integer number of reserved trunks, and we stop.

2. If $\delta > 0$, compute $S_{N-\delta}(\overline{a})$

   a. If $\overline{a} \leq N$ or $\overline{a} \leq 1$, use Eq. (3.152).

   b. Otherwise use Eq. (3.150).

3. Correct the values obtained in (1) by Eqs. (3.147–3.149).

Unfortunately, the numerical problems that can be encountered depend to a large extent on the particular model being calculated; for this reason, it is not possible to give general rules that would guarantee a robust implementation. Most of the difficulties come from the evaluation of the normalization constant, but, as we have seen here, other problems can be associated with the numerical calculation. Each case must be analyzed separately, and ad hoc transformations must be made that depend on the structure of the model.

Continuity

Finally, an equally important requirement is continuity in the numerical implementation of an algorithm. In many cases, we must compute the derivative of a probability with respect to the parameters of the system. Typical examples that occur quite frequently are $\partial B/\partial N$ and $\partial B/\partial a$. These derivations are often computed by finite differences, with the resulting necessity of computing the blocking probability of a link at two slightly different values of traffic. A potential problem can occur if the traffic model crosses a significant boundary between these two points. For instance, the traffic may be smooth for the initial value and peaked for the incremented value, or N may be an integer for the initial value and a fraction for the incremented one. If the traffic model used is not smooth at this boundary, very inaccurate values can ensue for the derivative, with potentially disastrous results for the algorithm that requires them.

Smoothness of the model has two very distinct meanings. The first is the smoothness of the analytic representation when two models are used to represent the same function on different parts of its domain. An example is the ERT method extended to smooth traffic by Katz's method. In this case, there are two potential points of trouble: $Z = 1$ and the value of Z where the overflow is set to zero. At these points, the first derivative is not necessarily continuous — which may or may not be important, depending on the application. The potential discontinuity comes from the fact that the models for peaked and smooth traffics are different, with no guarantee of continuity.

The second meaning of smoothness is related to the algorithmic implementation when a smooth function is evaluated by different numerical techniques over different parts of its domain — for example, the computation of the Erlang B function by the method proposed in [4], where different formulas are used for N integer and N real. The point is that, unless suitable precautions are taken, the numerical values computed by the algorithm may show a discontinuity at these points in spite of the fact that the function is continuous in the first sense.

Accuracy

Finally, the methods should be accurate, in the sense that the values computed for blocking and traffic should closely approximate the values measured in a real system. This requirement, however, is the least important, at least within reasonable limits. The reason is that network design and evaluation methods have simplifying assumptions of their own, generally unavoidable, which limit the accuracy of the results. In other words, it does not make much sense to compute link-blocking probabilities with five-digit accuracy if the network-blocking algorithm that uses these procedures is accurate only to two digits.

In practice, one should use the simplest model that gives sufficient accuracy in the network parameters being evaluated. Standard ERT has a long history

in the analysis of hierarchical networks, and IPP and BPP have been used successfully for some types of nonhierarchical ones. The latter models may not be required for certain types of networks where traffic is nearly Poisson, or where high accuracy in the blocking is not required — for example, in some dimensioning problems, a straightforward single-moment model may be sufficient. In any case, experimentation is nearly always required, and a network-simulation program is always needed to check the accuracy of the traffic model.

Problems

3.1. Derive the value of the mean and variance of the carried traffic for the Poisson arrival process in a group of size N Eqs. (3.15) and (3.16).

3.2. Demonstrate the recursion relation (3.4) for the Erlang B function.

3.3. Show that the integral representation of the Erlang B function satisfies Eq. (3.4).

3.4. This problem is related to the use of the binomial distribution as a means of representing smooth traffic with arbitrary parameters M, $V < M$. First, show that a finite-population model yields an ordinary binomial distribution. More precisely, let n be the finite number of sources of traffic, each arriving into a group of servers of infinite size. The BD coefficients are given by $\lambda_k = \lambda(n-k)$ and $\mu_k = k\mu$. Show that

$$p_k = \binom{n}{k}\left(\frac{\lambda}{\lambda+\mu}\right)^k\left(\frac{\mu}{\lambda+\mu}\right)^{n-k}.$$

Discuss the use of this model for representing smooth traffic.

3.5. Write the Kolmogorov equations for the Brockmeyer system. From these, derive the joint state probability distribution Eq. (3.17).

3.6. Draw the transition diagram that corresponds to Eqs. (3.32–3.34) for the IPP.

3.7. Using Eqs. (3.81) and (3.82), show that the traffic generated by the two-stage Cox process can have arbitrary peakedness.

3.8. Derive Eq. (A.20) for the residual life of a renewal process.

3.9. Consider a renewal process where the distribution of lifetimes for the intervals is given by $F(t)$. Show that the probability that the interval lasts up to time s given that it has already lasted up to $t-s$ is given by $(1-F(s))/(1-F(s-t))$.

3.10. Consider a renewal arrival process with negative exponential distribution for the intervals between the arrivals with mean $1/\lambda$. The renewal paradox is as follows. We can argue that if we start observing the arrivals at a random point in time, the average time before the next arrival should be $1/2\lambda$, that is, half the average duration between arrivals. Using Eq. (A.24), show that in fact this average value is given by $1/\lambda$. Explain the apparent contradiction.

3.11. Derive Eq. (3.44) for the binomial moments of the renewal traffic at an arbitrary point in time.

3.12. Show that the time and call congestion for a renewal arrival process are equal if and only if the arrival stream is Poisson. The "only if" part can be proved by considering a fixed interarrival distribution $U(t)$. If the ratio E/B is equal to 1, then $\Phi(N)$ must satisfy a particular relation. The proof is by showing that this relation does not hold except when $U(t)$ corresponds to the Poisson process.

3.13. Derive Eq. (3.78) for the IPP process.

3.14. For an arbitrary IPP, verify that $r_1, r_2 > 0$.

3.15. Consider an arrival process produced by the k-stage Erlang generator. Show that the traffic thus produced is always smooth.

3.16. Starting from Eqs. (3.53) and (3.54) for the renewal arrival process, show that the first two parameters of the overflow traffic are indeed those found by the Kosten model when the input is Poisson.

3.17. Demonstrate Eq. (3.59) for the overflow traffic of a renewal input stream.

3.18. Suppose that we want to use the extended ERT method to represent smooth traffic. Compute the condition under which the equivalent system exists.

3.19. Consider the equivalent carried model of Eqs. (3.93) and (3.94) for generating smooth traffic. Show that the equivalent system always exists if $Z < 1$.

3.20. Consider an offered traffic represented by a negative binomial distribution. Show that no renewal input can have this distribution. As a consequence, the representation of overflow traffic by a Pascal distribution is incorrect, at least theoretically. This is done by considering the ratio of β^*_{n+1}/β^*_n in the limit $n \to \infty$, both for the general renewal

input and for the Pascal distribution, and showing that the limits are different.

3.21. Derive Eqs. (3.107–3.109).

3.22. Consider the use of the BPP approximation in the computation of stream blocking. A requirement is that the partitioning formulas for the stream blocking must be such that the sum of the overflow traffics are identical to the aggregate overflow. Verify that this is indeed the case for Eq. (3.124). Is it also the case for Eq. (3.123)?

3.23. Consider the calculation of stream blocking using the BPP model. Using the recursive formulation, show that

$$\frac{m_1^j(N)}{M^j} = \frac{m_1^i(N)}{M^i} \quad \text{if } Z^i = Z^j$$

$$\frac{M^j - m_1^j(N)}{M^j} = E_N \quad \text{if } Z^j = 1$$

Interpret these equalities.

3.24. Interpret Eq. (3.127).

References

[1] Syski, R., *Introduction to congestion theory in telephone systems*, North-Holland, 1986.

[2] Wallström, B., "Congestion studies in telephone systems with overflow facilities," *Ericsson Technics*, vol. 22, pp. 187–345, 1966.

[3] Kleinrock, L., *Queuing Systems* , Wiley, 1975.

[4] Farmer, R.F., and Kaufman, L., "On the numerical evaluation of some basic traffic formulae," *Networks*, vol. 8, pp. 153–186, 1978.

[5] Jagerman, D.L., "Some properties of the Erlang loss function," *Bell System Technical Journal*, vol. 53, pp. 525–551, 1974.

[6] Wilkinson, R.I., "Theories for toll traffic engineering in the USA," *Bell System Technical Journal*, vol. 35, pp. 421–514, 1956.

[7] Brockmeyer, E., "The simple overflow problem in the theory of telephone traffic," (in Danish), *Teleteknik*, vol. 5, pp. 361–374, 1954.

[8] Akinpelu, J.M., "The overload performance of engineered networks with nonhierarchical and hierarchical routing," *International Teletraffic Congress*, vol. 10, pp. 3.2.4.1–3.2.4.7, June 8–15, 1983.

[9] Kuczura, A., "The interrupted poisson process as an overflow process," *Bell System Technical Journal*, vol. 52, pp. 437–448, 1973.

[10] Kosten, L., "On the blocking probability of graded multiples," (in German), *Elektr. Nachr.-Techn.*, vol. 14, pp. 5–12, 1937.

[11] Heffes, H., and Holtzman, J.M., "Peakedness of traffic carried by a finite trunk group with renewal input," *Bell System Technical Journal*, vol. 52, pp. 1617–1642, 1973.

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

[13] Pearce, C.E.M., and Potter, R.M., "Some formulae old and new for overflow traffic in telephony," *International Teletraffic Congress*, vol. 8, 1976.

[14] Potter, R.M., "Explicit formulae for all overflow traffic moments of the Kosten and Brockmeyer systems with renewal input," *Australian Telecommunication Research*, vol. 13, pp. 39–49, 1980.

[15] Guérineau, J.P., Labetoulle, J., and Lebourges, M., "A three moment method for traffic in a network based on a coxian input process," *International Teletraffic Congress*, vol. 10, pp. 4.4b.1–4.4b.7, 1983.

[16] Cox, D.R., "A use of complex probabilities in the theory of stochastic processes," *Proc. Cambridge Phil. Soc.*, vol. 51, pp. 313–319, 1955.

[17] Bretschneider, G., "Extension of the equivalent random method to smooth traffics," *International Teletraffic Congress*, vol. 7, pp. 411/1–411/9, 1973.

[18] Nightingale, D.T., "Computation with smooth traffics and the Wormald chart," *International Teletraffic Congress*, vol. 8, pp. 145.1–145.7, 1977.

[19] Katz, S., "Statistical performance analysis of a switched communication network," *International Teletraffic Congress*, vol. 5, 1967.

[20] Potter, R.M., "The equivalent non random method and restrictions imposed on renewal overflow systems by the specification of a finite number of overflow traffic moments," *International Teletraffic Congress*, vol. 9, pp. potter-1–potter-6, 1979.

[21] Kuczura, A., and Bajaj, D., "A method of moments for the analysis of a switched communication network's performance," *IEEE Transactions on Communications*, vol. COM–27, pp. 185–193, 1977.

[22] Guérineau, J.P., and Labetoulle, J., "End to end blocking in telephone networks: a new algorithm," *International Teletraffic Congress*, vol. 11, pp. 2.4.A.1:1–1.4.A.1:7, 1985.

[23] Delbrouck, L.E.N., "A unified approximate evaluation of congestion functions for smooth and peaky traffics," *IEEE Transactions on Communications*, vol. COM–29, pp. 85–91, 1981.

[24] Holtzman, J.M., "The accuracy of the equivalent random method with renewal inputs," *Bell System Technical Journal*, vol. 52, pp. 1673–1679, 1973.

[25] Wallström, B., and Reneby, L., "On individual losses in overflow systems," *International Teletraffic Congress*, vol. 9, pp. wallstrom-1–wallstrom-8, 1979.

[26] Kuczura, A., "Loss systems with mixed renewal and Poisson inputs," *Operations Research*, vol. 21, pp. 787–795, 1972.

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

[28] Manfield, D.R., and Downs, T., "Decomposition of traffic in loss systems with renewal input," *IEEE Transactions on Communications*, vol. COM–25, pp. 44–58, 1977.

[29] Butto, M., Colombo, G., and Tonietti, A., "On point to point losses in communication networks," *International Teletraffic Congress*, vol. 8, 1976.

[30] Deschamps, P.J., "Analytic approximation of blocking probabilities in circuit switched communication networks," *IEEE Transactions on Communications*, vol. COM–27, pp. 603–606, 1979.

[31] Delbrouck, L.E.N., "Use of Kosten's system in the provisioning of alternate trunk groups carrying heterogeneous traffic," *IEEE Transactions on Communications*, vol. COM-31, pp. 741–749, 1983.

[32] Akimaru, H., and Takahashi, H., "An approximate formula for individual call losses in overflow systems," *IEEE Transactions on Communications*, vol. COM-31, pp. 808–811, 1983.

[33] Girard, A., "Blocking probability of noninteger groups with trunk reservation," *IEEE Transactions on Communications*, vol. COM–33, pp. 113–120, 1985.

CHAPTER 4

Link-Decomposition Method

Performance measures for circuit-switched networks can take many forms, depending on the time scales considered and the ultimate objective of the planner. The most common — and probably most important — measure of the quality of a network is its ability to carry calls under normal circumstances. This qualitative measure is captured by the notion of total or average (over all origin-destination pairs) loss probability of the network. This global performance index, used to compare networks or routing algorithms, is also the main constraint in dimensioning.

Because it averages all the customers, the average blocking index gives no indication of the individual grade of services actually seen by customers. These grades of service are represented by $L^{i,j}$, the probability that a first-offered call presenting itself at node i, with node j as its destination, cannot be carried by the network. The grade of service can be expressed by the $L^{i,j}$s themselves or by the average

$$\overline{L} = \frac{\sum L^{i,j} A^{i,j}}{\sum A^{i,j}}. \tag{4.1}$$

These objectives are adequate for long-term planning, where decisions are made on an annual basis or longer, and where we can assume that the network parameters, such as the offered traffic, are stationary. In the shorter term, however, networks are not stationary. Traffic demands exhibit seasonal, daily, even hourly variations. This kind of consideration leads to other measures of the grade of service, such as the robustness of the network under traffic fluctuations — in particular, under generalized and focused overload. Similarly, network components are subject to failure. In this context, another measure of quality is the robustness of the network under failures, measured as the fraction of the nominal traffic that can be carried under specified failures. These questions are not specific to circuit-switched networks, but arise for all kinds of communication networks and indeed in other contexts. Because these problems are generally dealt with in the circuit-routing stage of network planning, they will not be discussed here any further.