THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

Volume 59 July–August 1980 Number 6

Congestion in Blocking Systems— A Simple Approximation Technique

By A. A. FREDERICKS

(Manuscript received November 1, 1979)

In the study of congestion in complex stochastic server systems, it is often desirable to have simple techniques available for obtaining approximations to important quantities of interest. This is particularly true in the early stages of the systems analysis and in cases where the only "solution" will be via a simulation. In this paper, we present an extremely simple, but surprisingly useful, technique for the approximate analysis of some such systems. Beginning with an approximation for the blocking of overflow traffic that was originally proposed by W. S. Hayward, we develop a natural extension to the approximation of blocking in a more general system as well as the determination of other (than blocking) quantities of interest. For the special, but important, case of renewal input to exponential servers, we give an explicit asymptotic (for heavy traffic) representation of the error introduced by this approximation.

I. INTRODUCTION

The study of many important stochastic server systems often leads to models that are extremely difficult to treat via exact analysis. An important example in telephony is the study of a (secondary) trunk group which is offered the superposition of several overflow streams [blocked calls from other (primary) trunk groups]. While the traffic characteristics of this pooled stream are rather complex (the stream is not even renewal), a useful characterization has been via its peakedness, defined as the variance-to-mean ratio of the number of busy

servers on an infinite trunk group offered this traffic. This peakedness concept is the heart of the "equivalent random" method introduced by Wilkinson[1] as an approximation technique for this trunking problem.† Although peakedness is generally not a complete characterization of traffic, it has been found to be quite useful in many applications besides the analysis of trunk groups offered overflow traffic.

Peakedness was used by W. S. Hayward (c. 1959) as the basis for an especially simple but surprisingly accurate approximation to the blocking experienced by the overflow traffic on the secondary trunk group in this trunking problem. If we are given an overflow stream with mean arrival rate λ offered to N (exponential) servers with rate μ, then Hayward's approximation for the resulting blocking probability (in telephony terminology, call congestion), B_c, is obtained in the following manner. First compute the peakedness z of the given overflow stream relative to an infinite group of (exponential, rate μ) servers, and then approximate B_c via

$$B_c \cong B_e\left(\frac{N}{z}, \frac{a}{z}\right), \tag{H}$$

where $a = \lambda/\mu$ is the offered load in erlangs and $B_e(N, a)$ is the Erlang loss function, which, for integral N, is defined by the well-known relation (e.g., see Ref. 2):

$$B_e(N, a) = \frac{a^N/N!}{\sum_{i=0}^{N} (a^i/i!)}.$$

(For the nonintegral number of servers, N/z, generally needed to apply (H), one can use interpolation—we discuss this further in Section 2.3.) One main purpose of this paper is to make this approximation more widely known (and appreciated) and to stimulate interest in the study of a general class of approximations of this type.

After some preliminaries in Section II, in Section III we consider a finite system of servers with given service time distribution and an arrival process for which the peakedness concept is defined and then we develop a new class of systems in each of which the blocking probability is the same as for the original system. We refer to this as the equivalent congestion model. Our investigation of this model leads naturally to Hayward's original approximation as well as to extensions to more general systems. The implications on the applicability of these approximations as provided by this model are explored via several numerical examples.

† Note that, for Poisson traffic, $z = 1$, while, for the overflow traffic in this trunking problem, $z > 1$; hence, the term "peaked" traffic, i.e., more variable than Poisson (see Refs. 2 and 3).

In Section IV we show that this equivalent congestion model can be used to obtain other (than blocking) quantities of interest. In particular, we obtain an approximation for the variance of the number of busy servers (carried process) in a G/G/N blocking system as well as for the peakedness of the resulting overflow process, i.e., attempts offered to this system which are blocked (lost). For the classical case where we are considering a (secondary) server group offered overflow traffic from a (primary) server group offered Poisson traffic, we find that the *expression* for the variance of the busy servers is exact, with the only error resulting from the appearance of the approximate call congestion for the exact value. For this case, we also find that, if we fix the peakedness z_A of the arrival process to the secondary server group while letting the offered load, a, tend to zero, then the peakedness of the resulting overflow process from the secondary group, z_0, tends to z_A. This disproves a well-known conjecture due to Wilkinson (c. 1967) that

$$z_0 \underset{\substack{z_A \text{ fixed}}}{\underset{a \to 0}{\longrightarrow}} 1$$

In Section V, we derive an explicit asymptotic (for large offered load) expression for the error introduced by Hayward's approximation for the special (but important) case of renewal input to exponential servers. This expression exhibits, in a more quantitative manner, many qualitative statements that result from our equivalent congestion model. In particular, it shows that the approximation will tend to overestimate blocking for "smooth" traffic ($z < 1$) and underestimate blocking for "peaked" traffic ($z > 1$). It also shows that the quantity (a/z) is an important accuracy parameter.

II. PRELIMINARIES

In this section, we briefly discuss some concepts concerning peakedness and its use that will be relevant to our discussion.

2.1 Definition of a simple overflow process

In our discussion, we will often refer to a class of renewal processes that are of particular interest in telephony, namely, the overflow process from a (finite) server group offered Poisson traffic. We will refer to this as a simple overflow process, or SOP. Thus, Wilkinson's equivalent random method consists of approximating the superposition of several (independent) SOPs by a single SOP. Note that while an SOP is a renewal process, the superposition of (independent) SOPs is not.

2.2 Generalizations and limitations of the peakedness concept

For a system in which the input stream is renewal and the service times are exponential, it is well known (e.g., see Ref. 4) that the peakedness, z, is given by*

$$z(\mu; A) = \frac{1}{(1 - \phi_A(\mu))} - \frac{\lambda}{\mu}, \tag{1}$$

where μ is the (exponential) service rate, A is the interarrival time distribution, ϕ_A is the Laplace Stieltjes (L.S.) transform of A, and λ is the mean arrival rate. Thus, for this class of systems, characterization of a stream by its peakedness is equivalent to characterization by the value of the L.S. transform of its interarrival time distribution function evaluated at the mean service rate of its servers. While this dependence on the value of the mean service rate, μ, is well known, as noted in Ref. 4 $z(0+)$ is often used as a convenient approximation. We show later that this can not only lead to large errors, but in some cases to meaningless results (negative probabilities). An important fact to note here is that, even for this class of systems, the resulting blocking can vary (relatively) greatly for a fixed value of peakedness,[6] a fact one should keep in mind when assessing the accuracy of any approximation that only uses a stream's mean arrival rate and peakedness for characterization.

The peakedness concept has been used to characterize state-dependent processes (e.g., see Ref. 7) as well as to analyze systems with constant service times.[8] In Ref. 4, an expression is given for the peakedness of a batched renewal process offered to exponential servers. Although early applications of peakedness were to systems with "peaked" traffic, i.e., $z > 1$, "smooth" traffic, i.e., $z < 1$, has also been considered.[9,10] Eckberg (unpublished work) has generalized these concepts to systems where the input is any stationary point process with a finite, second-order characterization and the service time is an arbitrary, finite, mean distribution. He has given explicit representations for the resulting peakedness functional $z(B; A_P)$ which depends on the service time distribution B for a given arrival process A_P. Similar results have also been obtained in studies of the infinite server queuing system. For example, in Ref. 11, Franken obtains the binomial moments for the same class of systems as studied by Eckberg. For other related work on infinite server queues, see also Refs. 12 and 13.

In this paper, we consider systems with input processes A_P and service time distributions B for which the peakedness concept is

---

* This formula can also be readily obtained from the discussion of the G/M/∞ system in Ref. 5.

defined. Hence we require the existence of an equilibrium distribution for the number of busy servers on the infinite server group with finite first and second moments. In particular, we may assume that our input process A_P is any of those noted above.

2.3 Nonintegral number of servers

Direct application of the equivalent random method (as well as Hayward's approximation) leads to the computation of the Erlang loss function for a nonintegral number of servers. In his calculation of blocking tables for peaked traffic,[14] Wilkinson made further approximations to avoid this apparent difficulty. Since then, it has been shown[15] that analytic continuation via the integral representation of the Erlang loss function provides an excellent interpolation procedure for nonintegral servers. (Another method of interpolation is given in Ref. 16.) In the discussion that follows, we tacitly assume integer values whenever it seems necessary, it being understood that the final results obtained are well defined for nonintegral values via analytic continuation or other interpolation procedures.

2.4 Hayward's approximation: An exact solution for an approximate problem

This section provides some motivation for the development of our basic equivalent congestion model as well as the approximations the model leads to. We consider a triple $(A_P; N, B)$, where A_P denotes the arrival process, N the number of servers, and $B(t)$ their service time distribution. We denote the mean arrival rate (mean number of arrivals per unit time) of the arrival processes by λ and the mean service time by $1/\mu$.

Note that, by Little's law,[17] the mean number of servers that would be busy on an infinite trunk group [with service distribution $B(t)$] is just given *gy* $a = \lambda/\mu$, i.e., the offered load in erlangs. The basic problem we address is that of finding an approximation to the blocking probability (fraction of lost attempts) B_c for the system $(A_P; N, B)$. For this purpose, we consider arrival processes that are (partially) characterized by the first two moments of the distribution of the number busy on an infinite server group offered this traffic, i.e., $A_P \cong (a, z)$ and hence $(A_P; N, B) \cong (a, z; N, B)$. Thus, in analogy with (H) of Section I, we can "write down" a generalized Hayward type approximation for this system as

$$B_c(A_P; N, B) \cong B_H(a, z; N, B) = B_e\left(\frac{N}{z}, \frac{a}{z}\right). \tag{2}$$

It is possible to construct other systems $(A'_P; N, B')$ for which $(a', z') = (a, z)$, i.e., which have the same peakedness characterization *and*

for which the application of Hayward's approximation to these systems is exact. One such system can be obtained by using a batched Poisson input with constant batch size $k = z$ and a constant service time for the trunk group (both distributions have the same means as the respective original distributions). To see this, note that, if we divide an infinite trunk group (with constant service times) into k subgroups and offer one arrival from each batch to each subgroup, then the input to each subgroup is Poisson. Since the stochastic processes recording the number of busy servers in each subgroup are identical (with probability 1), i.e., the subgroups are "perfectly" correlated, the variance-to-mean ratio of the total number busy is simply k (which we have taken to be z). Thus, if we choose $k = z$, the peakedness of our original system, then for N a multiple of z, a new system consisting of this batched Poisson input into constant holding time servers will have the same peakedness characterization. Moreover, the calls lost from the server group in the second system will be given exactly by Hayward's "approximation," i.e., eq. (2). The problem with this development is that it gives little insight into the applicability of Hayward's approximation in other, more interesting situations.

III. AN EQUIVALENT CONGESTION MODEL FOR BLOCKING SYSTEMS (A DEVELOPMENT OF HAYWARD'S APPROXIMATION)

The basic problem we are concerned with is studying the congestion in a system (partially) characterized by $(a, z; N, B)$. We begin by considering the case of "peaked" traffic ($z > 1$). A system of m subgroups is constructed which is equivalent to the original system in that each of these subgroups (as well as the total system) has the same blocking probability as the original system. We then show how an approximation to this blocking probability leads naturally to Hayward's approximation and to a similar Hayward-type approximation for "smooth" traffic ($z < 1$). Finally, we note that we can thus apply a Hayward-type approximation to a rather wide variety of systems.

3.1 An equivalent congestion system for peaked traffic ($z > 1$)

We begin by constructing an equivalent congestion model, which is essentially a generalization of the concepts introduced in Section 2.4. We assume we are given a server group of N servers offered traffic with peakedness $z > 1$. This server group is divided into m subgroups, each with N/m servers.† The underlying traffic is first offered to a *distributor*, which will allocate the arrivals to one of the m groups

† Recall from Section 2.3 that we assume integer values where needed in our development, the final results applying for noninteger values via analytic continuation.

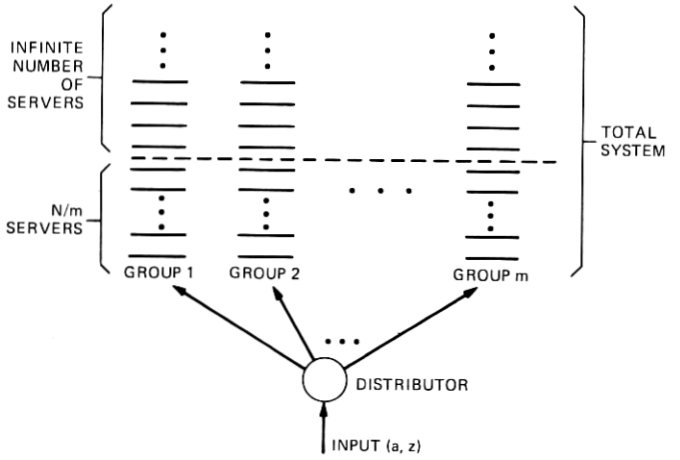

Fig. 1—Equivalent congestion model.

according to rules to be given. We also consider each of the m groups of N/m servers to be embedded in its own set of infinite servers, i.e., we have m groups of infinite servers each containing N/m servers from the original group of N servers (see Fig. 1).

We specify shortly exactly how the traffic might be allocated to each of the m groups, but for now we make three basic requirements for the class of rules we will consider.

(i) The rule must result in the groups being stochastically equivalent (in a sense that will be clear, shortly).

(ii) No arrival is offered to a group with the first N/m servers busy unless *all* groups have the first N/m servers busy.

(iii) If an arrival is offered to a group with the first N/m servers busy, it is put on the infinite set of servers associated with this group.

Clearly, (ii) implies that the arrivals lost by the m groups of N/m first servers will be exactly those that would have been lost in the original system. Moreover, because of (i) the blocking probability on any group will be exactly what the original server group of N servers would have experienced if offered (a, z).†

Now the mean number of servers busy in this system (including the infinite overflow groups) must be a, the offered load. Moreover, because of (i) the mean number of servers busy on the ith group, a_i, will all be the same and equal to $\hat{a} = a/m$. If we compute the variance of

† If the lost load in the original (and hence equivalent) system is l, then each group will share equally in the loss, i.e., have loss l/m. Since each group is offered a/m, the blocking probability on any group will be l/a, i.e., that of the original system.

the number of servers busy on this total system, we have

$$v = \sum_{i=1}^{m} v_i + \sum_{\substack{i \neq j \\ i=1,m \\ j=1,m}} \rho_{ij} \sqrt{v_i} \sqrt{v_j},$$

which must equal az, where here v_i is the variance of the number of servers busy on group i and ρ_{ij} is the correlation coefficient between the i and j server groups. Now because of (i) all the v_i's must be equal to, say, $\hat{v}$ and also all the ρ_{ij} must be equal to, say, ρ. Thus we have

$$m\hat{v} + m(m-1)\hat{v}\rho = az$$

or

$$\hat{v} = \frac{az}{(m + m(m-1)\rho)}. \tag{3}$$

And, finally, the peakedness of the traffic offered to each of these groups is given by

$$\hat{z} = \frac{z}{(1 + (m-1)\rho)}. \tag{4}$$

Thus, an individual group in this equivalent congestion model is characterized by

$$(\hat{a}, \hat{z}; \hat{n}, B) = \left(\frac{a}{m}, \frac{z}{(1 + (m-1)\rho)}, \frac{N}{m}, B\right) \tag{5}$$

and the blocking this group sees will be the desired blocking for any distribution rule employing rules (i) to (iii). While an exact result, this does not help us much as it stands since, for one thing, computation of ρ can be extremely difficult for a given distribution rule and, moreover, we still need to compute the blocking for peaked traffic when we are done. We could, in principle, eliminate the second problem by choosing m so that $\hat{z} = 1$, i.e., take

$$m = \frac{z - 1 + \rho}{\rho} \tag{6}$$

and then make the (admittedly wrong) assumption that $\hat{z} = 1$ implies Poisson traffic and hence use the Erlang loss function (as an approximation). Even with this approximation, we would still need ρ. However, motivated by Section 3.4, we note that, if we choose $m = z$ in (4) and approximate ρ by 1, we get $\hat{z} = 1$, i.e., $(\hat{a}, \hat{z}; \hat{n}, B) \cong (a/z, 1; N/z, B)$ and hence we are led to Hayward's approximation, i.e., eq. (2), for the desired blocking. To obtain some insight into what is involved in this approximation, we consider a specific distribution rule satisfying (i) to (iii). Of all the groups where one of the first N/m servers has at least

Table I—Blocking for simple overflow processes

Exact* from Ref. 14	Hayward's Approximation	Improved Approximation	a/z
0.005	0.002	0.004	0.905
0.01	0.005	0.009	1.160
0.02	0.014	0.018	1.485
0.03	0.022	0.027	1.715
0.05	0.041	0.047	2.080
0.07	0.061	0.066	2.375
0.10	0.090	0.096	2.765
0.20	0.192	0.196	3.920
0.40	0.396	0.398	6.535
	$z = 2, N = 10$		

* See footnote on this page.

one idle server, we offer the arrival to the group with the least number of busy servers on its entire infinite server group—ties are broken by random selection. In this case, the call is said to be carried. If no such group exists, we offer the arrival to the group with the least number of busy servers and consider it lost to the finite (N/m) subgroup embedded in this infinite group. It is clear that this will tend to maximize ρ (make ρ near 1)—within the constraint of equivalent congestion. Thus, it is not unreasonable to approximate ρ by 1, an upper bound. We emphasize that two approximations were made in this development of Hayward's approximation. First, for $m = z$, the traffic offered to each subsystem, $(\hat{a}, \hat{z})$, is still peaked, i.e., $\hat{z} > 1$. Intuitively, this leads one to suspect that the resulting Hayward approximation would tend to underestimate blocking. Moreover, intuitively one would expect that if the load per subgroup, a/z, were small, it would be difficult to maintain a high correlation with the given distribution rules, i.e., keep ρ near 1. Second and perhaps more important, even if $\hat{z} = 1$, as noted earlier, the blocking still may not be given by the Erlang loss function. Indeed, the stream $(\hat{a}, \hat{z})$ is a rather odd process where the arrivals to a subsystem depend not only on the state of the subsystem, but the state of other subsystems.

Section V develops some explicit quantification of the behavior of Hayward's approximation for an important class of systems. Here we give some numerical results which demonstrate the surprising accuracy of Hayward's approximation and also support the intuitive statements made above. For this purpose, we first compare Hayward's approximation with exact results for a simple overflow stream.†

Table I gives blocking values for $z = 2$, $N = 10$, and various load

† The "exact" results are from Ref. 14 and hence contain the approximations noted earlier. The impact on the accuracy is at most $a \pm 1$ for the next digit (not shown) for values greater than 0.01, but can be greater for smaller blocking values.

Table II—Blocking for simple overflow processes

Exact* from Ref. 14	Hayward's Approximation	a/z
0.001	0.0013	13.245
0.005	0.0057	15.205
0.01	0.011	16.295
0.02	0.021	17.640
0.03	0.031	18.600
0.05	0.051	20.075
0.07	0.071	21.285
0.10	0.101	22.885
0.20	0.200	27.740
0.40	0.400	39.435

$z = 2, N = 50$

* See footnote on p. 813.

levels. (The column marked "Improved Approximation" is discussed in Section IV.) As anticipated, for a/z small we see somewhat poor agreement and, in fact, an underestimation of blocking by Hayward's approximation, while the agreement becomes quite good as a/z increases. Table II shows data for $z = 2$, $N = 50$ and increasing load. Here we see excellent agreement throughout, but unlike case 1, a/z is (relatively) large throughout. As a last case, we consider an extremely large value of peakedness, $z = 10$. Table III shows data for this case but in a different way. The exact blocking is fixed at 0.40, and the offered load is increased with a corresponding increase in N to maintain this blocking level. We again see improved accuracy with increasing a/z and, moreover, a comparison of Tables I and III also shows that accuracy appears to be more related to a/z than simply to a.

We note before proceeding that if one is truly analyzing a simple overflow stream, then the above comparisons show relative errors for

Table III—Blocking for simple overflow processes

N	Hayward's Approximation	a/z
20	0.358	1.754
30	0.382	3.308
40	0.390	4.892
50	0.394	6.495
60	0.396	8.110
70	0.397	9.733
80	0.398	11.364

$z = 10, B = 0.40$*

* Exact from Ref. 14; see footnote on p. 813.

Table IV—Blocking for batched Poisson (batch size = 3)

Exact	Hayward's Approximation	$P(N)$ (Time Congestion)	a/z
0.0244	0.0184	0.0094	5
0.2264	0.2146	0.1124	10
0.4186	0.4103	0.2366	15
0.5437	0.5380	0.3339	20
0.6265	0.6224	0.4083	25

$z = 2$, $N = 20$

Hayward's approximation. However, if one is using this type of traffic to approximate a more complicated system, e.g., the superposition of several such streams, then it is not clear that the equivalent random method is preferable to Hayward's approximation—particularly in light of the latter's ease of use.

As a second numerical example, we consider a nonrenewal input, namely, batched Poisson arrivals with constant batch size (i.e., the input process discussed in Section 2.4) offered to an (exponential) server group. Note for this process, the peakedness z is given by[4] $z = (k + 1)/2$. Table IV shows blocking values for $z = 2$ ($k = 3$), $N = 20$ servers, and the load ranging from 5 to 50 erlangs, while Table V is for $z = 5$ ($k = 9$), $N = 20$ servers and the load ranging from 1 to 10 erlangs. We have also included $P(N)$, the (exact) probability that all servers are busy at an arbitrary *time* point (time congestion), for comparison. Again we see improved accuracy for larger values of a/z.

3.2 Hayward's approximation for smooth traffic ($z < 1$)

The above development can readily be inverted to derive a Hayward-type approximation for systems with smooth traffic, ($z < 1$). What we need to do is assume the system $(a, z; N, B)$, $z < 1$, is the result of splitting a larger system $(ma, z_m; mN, B)z_m > z$, into m groups. We

Table V—Blocking for batched Poisson (batch size = 9)

Exact	Hayward's Approximation	$P(N)$ (Time Congestion)	a/z
0.03991	0.0154	0.0034	1
0.1424	0.0952	0.0189	2
0.3542	0.3107	0.0691	4
0.5003	0.4696	0.1208	6
0.5966	0.5746	0.1670	8
0.6630	0.6467	0.2075	10

$z = 5$, $N = 20$

then have [analogous to (3)]:

$$V_m = maz_m = mV + m(m-1)V\rho, \tag{7}$$

and hence

$$z_m = z + (m-1)z\rho. \tag{8}$$

Thus, again assuming $\rho = 1$, we see from (8) that for $z_m = 1$ we need $m = 1/z$. The blocking probability we seek is then given (approximately) by

$$B_c(A_P; N, B) \cong B_H(a, z; N, B) \triangleq B_e\left(\frac{N}{z}, \frac{a}{z}\right), \tag{9}$$

i.e., we again have a Hayward-type approximation. Now, however, we see from (8) that, for $m = 1/z$, $z_m < 1$, and hence we would expect this type of approximation to overestimate blocking for smooth traffic. We briefly discuss an example that illustrates this and will also allow us to introduce some concepts that we will build on in Section V.

For a single (exponential) server with renewal input, the blocking probability, B_c, is given by $\phi_A(\mu)$, the L.S. transform of the interarrival time distribution evaluated at the mean service rate (e.g., see Ref. 5). With the aid of eq. (1), this can be written as

$$B_c(a; \mu) = (a + z(\mu) - 1)/(a + z(\mu)), \tag{10}$$

where $a = \lambda/\mu$. We have written $z(\mu)$ for $z(\mu; A)$ supressing the dependence on the interarrival time distribution—which, for this discussion, we assume is fixed.

Equation (10) reveals an interesting fact. As noted in Section II, although peakedness is known to depend on the mean service rate (μ), $z(0+)$ is often used as an approximation since it is generally much easier to obtain. However, we note from (10) that since B_c must be nonnegative, we must have $a > 1 - z(\mu)$. Hence, the use of $z(0+)$ not only can lead to erroneous results but, in particular, for smooth traffic, can lead to meaningless results. For example, the peakedness z_D of a system with deterministic interarrival times and exponential service times is given by [from eq. (1)]

$$z_D(\mu) = \frac{1}{1 - \phi_D(\mu)} - a = \frac{1}{1 - e^{-1/a}} - a.$$

However, for this system the peakedness is often taken to be ½, the value of $z(0^+)$ (note

$$z_D(\mu) \underset{\substack{\mu \to 0+ \\ (a \to \infty)}}{\to} \tfrac{1}{2}),$$

Table VI—Blocking for D/M/1

Offered Load, A	Exact Blocking	Hayward's Approximation*
0.2	0.0067	0.054
0.5	0.1353	0.200
1.	0.3678	0.400
2.	0.607	0.615
5.	0.819	0.819

* Uses $z(0+) = \frac{1}{2}$.

while we see that clearly

$$z_D(\mu) \underset{\substack{\mu\to\infty \\ (a\to 0+)}}{\to} 1.\dagger$$

Thus extreme caution should be taken in the use of $z(0+)$ to approximate $z(\mu)$. Keeping this in mind, we assume for the moment that $z(\mu) \equiv \frac{1}{2}$. This gives, for our Hayward approximation to the blocking in a D/M/1 system,

$$B_c(A_P; N, B) \cong B_H\left(a, \frac{1}{2}; 1, B\right) = B_e(2, 2a) = \frac{a^2}{\left(\frac{1}{2} + a + a^2\right)}, \quad (11)$$

while the true blocking is given by

$$B_c(A_P; N, B) = B_c(a) = \exp(-1/a). \quad (12)$$

It is easy to see that, for a sufficiently large,

$$B_H(a) = B_c(a) + \frac{1}{6a^3} + O\left(\frac{1}{a^4}\right). \quad (13)$$

Thus, as anticipated, this Hayward-type approximation improves with increasing load and over estimates blocking. What about "light" loads? Table VI compares $B_H(a)$ and $B_c(a)$ for various load levels. We see that the accuracy degrades quite rapidly as $a \to 0$ and, moreover, we note that this is not due to using $z(0+)$ rather than $z(\mu)$, since this improved z would, in fact, make things worse. Indeed, since $z(\mu) \to 1$ as $\mu \to \infty$ $(a \to 0)$, a Hayward-type approximation using $z(\mu)$ would result in

$$B_H(a) = B_e\left(\frac{N}{z(\mu)}, \frac{a}{z(\mu)}\right) \underset{a\to 0}{\to} \frac{a}{1 + a}. \quad (14)$$

† It is easy to show from eq. (1) that $z(\mu) \underset{\mu\to\infty}{\to} 1$ for any orderly renewal process.

Comparison of (14) and (12) shows that the (relative) error is of exponential order. The main conclusion to be drawn here is that peakedness may not be a good characterization of blocking systems for extremely light loads—a fact that is clearly demonstrated in (6).

IV. OTHER APPLICATIONS OF THE EQUIVALENT CONGESTION MODEL FOR LOSS SYSTEMS

In this section, we show that the equivalent congestion model used above to develop and extend Hayward's approximation can be used to obtain similar approximations to other quantities of interest in traffic systems as well as to improve the approximations already obtained.

4.1 Carried calls process

In many systems, it is important to determine the variance of the number of busy servers. For example, these busy servers may serve as sources to other systems. The equivalent congestion model used in the above development of Hayward's approximation can readily be used to estimate this quantity. Indeed, if we assume that $\hat{z} = 1$ ($\rho = 1$) for each subsystem in the model (and that this input traffic is Poisson), then the variance of the number of busy servers, v_b, on this finite subgroup of N/z trunks is given by (Ref. 2, p. 97)

$$v_b = \hat{a}'(1 - llt(\hat{a}, \hat{n})), \tag{15}$$

where $\hat{a}'$ is the load carried (mean number of busy servers) by the finite subgroup of $\hat{n} = N/z$ servers offered $\hat{a} = a/z$ erlangs of (Poisson) traffic and $llt(\hat{a}, \hat{n})$ is the load carried by the last server in this group if it is considered as an ordered hunt group, i.e.,

$$llt(\hat{a}, \hat{n}) = \hat{a}(B_e(\hat{n} - 1, \hat{a}) - B_e(\hat{n}, \hat{a})).$$

Using the assumed value of 1 for the correlation, ρ, between groups we find that the variance of the number busy on all z groups is given by

$$V_b = z^2 v_b = z^2 \hat{a}'(1 - llt(\hat{a}, \hat{n})). \tag{16}$$

Using the fact that

$$B_e^{-1}(\hat{n}, \hat{a}) = \frac{\hat{n} B_e^{-1}(\hat{n} - 1, \hat{a})}{\hat{a}} + 1,$$

we find that (16) can be written as

$$V_b = za'\left(1 - aB_e(\hat{n}, \hat{a})\frac{(N - a')}{za'}\right), \tag{17}$$

where $a' = z\hat{a}'$ is the total carried load, thus providing us with an approximation for the variance of the carried load. Now for the case where the input stream (a, z) is a simple overflow process, by using

the results of Brockmeyer[18] for the joint state probabilities for the total system (primary plus secondary group), one can readily derive the following expression for the exact variance, V_b^*, of the number of busy servers on the secondary group[19]

$$V_b^* = za\left(1 - a'B_c(a, z; N)\frac{(N - a')}{za'}\right),$$

where all parameters are as above and $B_c(a, z; N)$ is the *exact* blocking probability. Thus, comparing this result with (17) we see that the only error in the use of (17) to approximate V_b^* is the appearance of Hayward's approximation B_H for the true blocking probability, B_c. Besides giving more credence to Hayward's approximation for blocking, this consistency also indicates that this approach might prove useful in developing approximations for other quantities of interest in traffic systems.

4.2 Overflow process

Our equivalent congestion model can also be used to obtain an approximation to the peakedness, z_0, of the calls *lost* from the N-server group which is offered a erlangs with peakedness z. This quantity is of interest in designing hierarchical networks, since it then characterizes the traffic offered to the next level in the hierarchy. Under the assumption that $\hat{z} = 1$, we can use the well-known result (e.g., see Ref. 2) that the peakedness of the overflow from the $\hat{n}$ server subgroup is given by

$$\hat{z}_0 = 1 + \hat{a}' - \hat{a} + \frac{\hat{a}}{\hat{n} + 1 - \hat{a}'}, \tag{18}$$

where $\hat{a}'$ is the load carried by this group. Again, with perfect correlation ($\rho = 1$), the peakedness of the total overflow process is found to be $z\hat{z}_0$. Expressing the variables in (18) in terms of the original (a, z; N), we obtain the approximation

$$z_0 = z\left(1 - \frac{a}{z}B_e\left(\frac{N}{z}, \frac{a}{z}\right) + \frac{a}{N + z - a'}\right), \tag{19}$$

where a' is the total carried load. Now for large a, B_e has the asymptotic behavior [e.g., see Section V, eq. (26) and following discussion]:

$$B_e^{-1}(N, a) = 1 + \frac{N}{a} + \frac{N(N-1)}{a^2} + O\left(\frac{1}{a^3}\right),$$

from which we find that (19) readily yields the limit

$$z_0 \underset{a\to\infty}{\longrightarrow} z,$$

as might be expected. However, we see from (19) that

$$z_0 \underset{a \to 0}{\longrightarrow} z,$$

a fact that seems to contradict a well-known conjecture by Wilkinson (c. 1967) that, for a simple overflow process (SOP) offered to exponential servers,

$$z_0 \underset{a \to 0}{\longrightarrow} 1.$$

We now show that for such a system (SOP offered to exponential servers) the correct limit is

$$z_0 \underset{a \to 0}{\longrightarrow} z.$$

If a stream (a, z) is an SOP, then A_e, N_e exists such that (see Fig. 2) (for a detailed discussion, see Ref. 2, Section 4.3 as well as Section 4.7)

$$a = A_e B_e(N_e, A_e) \tag{20}$$

$$z = 1 - a + \frac{A_e}{(N_e + 1 + a - A_e)}. \tag{21}$$

We assume z is fixed and consider A_e, N_e as determined by eqs. (20) and (21) to be functions of a. Solving (21) for A_e yields

$$A_e = \frac{(z + a - 1)(N_e + A + 1)}{(z + a)}. \tag{22}$$

Now if this stream (a, z) is offered to N (exponential) servers, then the exact peakedness of the overflow, z_0^e, is given from (18) as

$$z_0^e = 1 + A_e' - A_e + \frac{A_e}{(N_e + N + 1 - A_e')}, \tag{23}$$

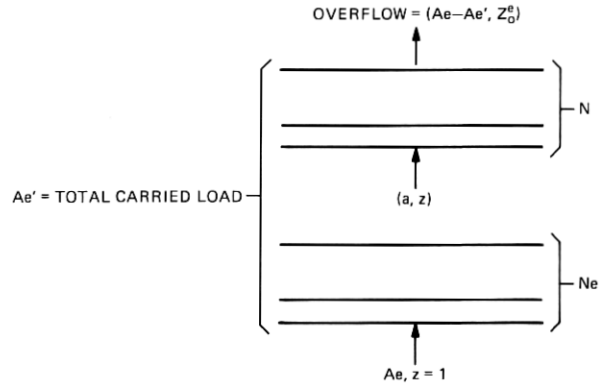

Fig. 2—Equivalent random representation for simple overflow streams.

where A'_e is the load carried by the combined $N_e + N$ server group, i.e.,

$$A'_e = A_e(1 - B_e(N_e + N, A_e)).$$

Substituting (22) in (23), we obtain

$$z_0^e = 1 - \frac{(z + a - 1)(N_e + a + 1)}{(z + a)} \cdot B_e\left(N_e + N, \frac{(z + a - 1)}{(z + a)}(N_e + a + 1)\right) + (z + a - 1)(N_e + a + 1) \Bigg/ \left[(z + a)\left(N_e + N + 1 - \frac{(z + a - 1)}{(z + a)}(N_e + a + 1) \cdot \left(1 - B_e\left(N_e + N, \frac{(z + a - 1)}{(z + a)}(N_e + a + 1)\right)\right)\right)\right]. \quad (24)$$

Substituting (22) in (20) yields the following implicit function for N_e:

$$a = \frac{(z + a - 1)}{(z + a)}(N_e + a + 1)B_e \cdot \left(N_e, \frac{(z + a - 1)}{(z + a)}(N_e + a + 1)\right). \quad (25)$$

Now for $a \to 0$, the right-hand side of (25) will tend to a finite (nonzero) limit for fixed N_e. On the other hand, for fixed a, the right-hand side can be written in the form $(C_1N_e + C_2)B_e(N_e, C_1N_e + C_2)$ where C_1 and C_2 are constants *and* $C_1 < 1$. Thus, for fixed a, the right-hand side of (25) tends to 0 as $N_e \to \infty$.[15] Thus the solution, N_e, of (25) must satisfy

$$N_e(a) \underset{a \to 0}{\to} \infty$$

$$(z > 1, \quad \text{fixed})$$

Using this asymptotic result in (24) yields the desired result, i.e.,

$$z_0^e \underset{a \to 0}{\to} z$$

$$(N_e \to \infty).$$

It should be pointed out that this behavior was not detected in the numerical studies of Wilkinson since it requires (for the cases considered by him) extremely small values of a. Figure 3 reproduces one of

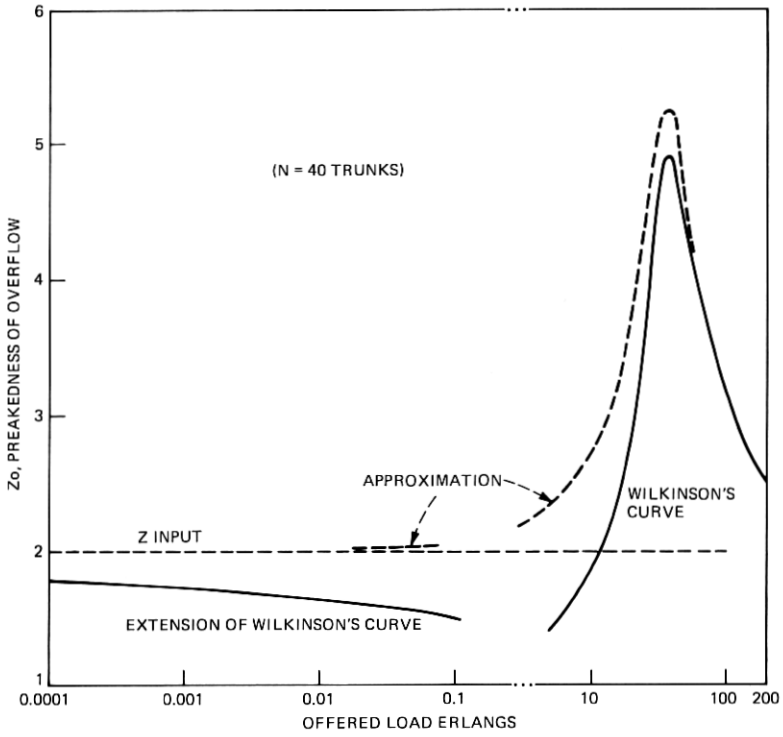

Fig. 3—Peakedness of overflow.

Wilkinson's curves. Our approximation is also shown for comparison. (Note that the true peakedness of the overflow is still only 1.75 for $a = 0.0001$.) To provide some intuitive understanding of this phenomenon, we note that fixing z means fixing the *variance*-to-mean ratio of the number busy on an infinite server group offered this traffic and hence, as $a \to 0$, the coefficient of variation of this process (number busy) will tend to infinity. That is, by fixing z and letting $a \to 0$, we are changing the "structure" of the process to one that is "infinitely bunchy."

4.3 Improved approximations

One method to improve the approximations given here would be to obtain a better estimate for ρ, i.e., better than 1. We give a simple illustration of how this might be done. Assume we have an input stream characterized by $(a, z) = (a, 2)$. Assuming $\rho = 1$ yields $m = z = 2$ from eq. (6). Given that we have two subgroups, we now attempt to estimate the actual correlation between the two busy processes.

Let n_i represent the number busy on the ith infinite subgroup ($i =$

1, 2) and let P_n be the probability that $n_1 + n_2 = n$. Note that assuming $\rho = 1$ is equivalent to assuming $n_1 = n_2$ for all n. We can then obtain an improved upper bound on ρ by assuming that, when n is even, $n_1 = n_2$ and that, when n is odd, $|n_1 - n_2| = 1$.

We then have†

$$E(n_1) = E(n_2) = \frac{E(n)}{2} = \frac{\bar{n}}{2} = \frac{a}{2},$$

$$E(n_1^2) = E(n_2^2) = \sum_{n=0}^{\infty} \begin{cases} \left(\frac{n}{2}\right)^2, & n \quad \text{even} \\ \frac{1}{2}\left(\frac{n+1}{2}\right)^2 + \frac{1}{2}\left(\frac{n-1}{2}\right)^2, & n \quad \text{odd} \end{cases} P_n$$

$$= \frac{1}{4}\sum_{n=0}^{\infty} n^2 P_n + \frac{1}{4}\sum_{n-\text{odd}} P_n,$$

$$E(n_1 n_2) = \sum_{n=0}^{\infty} \begin{cases} \left(\frac{n}{2}\right)\left(\frac{n}{2}\right), & n \quad \text{even} \\ \frac{(n-1)}{2}\frac{(n+1)}{2}, & n \quad \text{odd} \end{cases} P_n$$

$$= \frac{1}{4}\sum_{n=0}^{\infty} n^2 P_n - \frac{1}{4}\sum_{n-\text{odd}} P_n.$$

Thus

$$\text{Cov}(n_1, n_2) = E(n_1 - \bar{n}_1)(n_2 - \bar{n}_2) = \frac{1}{4}\bar{n}^2 - \frac{1}{4}\sum_{n-\text{odd}} P_n - \frac{1}{4}\bar{n}^2$$

and

$$\text{var}(n_i) = E(n_i - \bar{n}_i)^2 = \frac{1}{4}\bar{n}^2 + \frac{1}{4}\sum_{n-\text{odd}} P_n - \frac{1}{4}\bar{n}^2$$

Hence

$$\rho = \frac{\text{cov}(n_1, n_2)}{\text{var}(n_i)} = \frac{\sigma_n^2 - \sum_0}{\sigma_n^2 + \sum_0},$$

where $\sum_0 = \sum_{n-\text{odd}} P_n$ and σ_n^2 is the variance of n, which is just given by az. Thus

$$\rho = \frac{az - \sum_0}{az + \sum_0}.$$

† In what follows, $E(x)$ denotes the expected value of x.

Hence, given an estimate of $\sum_0$, we can use the resulting ρ in eq. (6) to obtain an improved value for m, the number of subgroups to be formed in the equivalent congestion model of Section III. As a very rough estimate, we could assume that $\sum_0$ equals ½ (at least if $P_0 < 0.5$). This already gives a reasonable improvement as indicated on Table I (data marked "Improved Approximation"). This method can readily be extended to $z > 2$ and results in an estimate of ρ that also depends only on a, z, and P_n. Thus, with some estimates of P_n (e.g., Ref. 7 or other techniques) and a reasonable interpolation procedure for non-integral m, we can determine an improved relation for m via (6).

V. ASYMPTOTIC BEHAVIOR OF THE GENERALIZED HAYWARD'S APPROXIMATION

In Section IV we saw that, for the D/M/1 blocking system, the error made in using the generalized Hayward's approximation was asymptotic (for $a \to \infty$) to $1/6a^3$ [eq. (13)]. Note that, if B_e is the Erlang loss function, then we have analogous to (13) for a D/M/1 loss system

$$B_e(1, a) = B_c(1, a) + \frac{1}{2a^2} + O\left(\frac{1}{a^3}\right)(a \to \infty).$$

That is, the generalized Hayward's approximation picks up one more term in the asymptotic behavior of the true blocking than does Erlang's loss function. This provides an interesting view of this type of approximation, and we are naturally led to ask if this is a more general property. We show that this is the case.

For renewal input to N exponential servers, the blocking is given by (e.g., see Ref. 5, p. 179):

$$B_c^{-1}(N, a) = \sum_{i=0}^{N} \binom{N}{i} \prod_{j=1}^{i} \frac{1 - \phi(j\mu)}{\phi(j\mu)}, \tag{26}$$

where ϕ is the L.S. transform of the interarrival time distribution, μ is the mean (exponential) service rate, $a = \lambda/\mu$ is the offered load, and $1/\lambda$ is the mean interarrival time (the empty product is taken to be 1). Using (1) to replace ϕ by z in (26), we have

$$B_c^{-1}(N, a) = \sum_{i=0}^{N} \binom{N}{i} \prod_{j=1}^{i} \frac{j\mu}{\lambda + j\mu(z(j\mu) - 1)}$$

or

$$B_c^{-1}(N, a) = \sum_{i=0}^{N} c_i$$

where $c_0 = 1$ and, denoting $z(j\mu)$ by z_j,

$$c_i = c_{i-1} \frac{(N - i + 1)}{i(z_i - 1) + a}.$$

Now for large a (small μ) we also have the asymptotic expansion[4]

$$z(\mu+) = z_0 + k\mu + O(\mu^2) \tag{27}$$

where

$$z_0 = z(0+) = \frac{(\sigma^2\lambda^2 + 1)}{2},$$

$$k = \frac{z(0+)^2}{\lambda} - \lambda^2 \frac{\alpha_3}{6},$$

σ^2 is the variance of the interarrival time, and α_3 is its third moment. Thus, for a large (μ small), we have the asymptotic expansions

$$c_1 = \frac{N}{a}\left(1 - \frac{(z_1 - 1)}{a} + \frac{(z_1 - 1)^2}{a^2}\right) + O\left(\frac{1}{a^4}\right)$$

$$c_2 = c_1 \frac{(N-1)}{a}\left(1 - \frac{2}{a}(z_2 - 1)\right) + O\left(\frac{1}{a^4}\right)$$

$$c_3 = c_2 \frac{(N-2)}{a} + O\left(\frac{1}{a^4}\right)$$

$$c_4 = O\left(\frac{1}{a^4}\right)$$

$$c_i = o\left(\frac{1}{a^4}\right), \qquad i > 4.$$

Thus, for $a \to \infty$,

$$B_c^{-1}(N, a) = 1 + \frac{N}{a} + \frac{1}{a^2}(N(N-1) - N(z_1 - 1))$$

$$+ \frac{1}{a^3}(N(z_1 - 1)^2 - N(N-1)(z_1 - 1) - 2N(N-1)(z_2 - 1)$$

$$+ N(N-1)(N-z)) + O\left(\frac{1}{a^4}\right). \tag{28}$$

Now

$$B_H^{-1}(N, a) = B_e^{-1}\left(\frac{N}{z_1}, \frac{a}{z_1}\right) = \sum_{i=0}^{N} d_i$$

where

$$d_0 = 1$$

and

$$d_i = d_{i-1}\frac{((N/z_1) - i + 1)}{a/z_1} = \frac{(N - iz_1 + z_1)}{a}.$$

Hence, for $a \to \infty$,

$$B_H^{-1}(N, a) = 1 + \frac{N}{a} + \frac{N(N - z_1)}{a^2} + \frac{N(N - z_1)(N - 2z_1)}{a^3} + O\left(\frac{1}{a^4}\right). \quad (29)$$

Subtracting (28) from (29), we obtain

$$B_H^{-1}(N, a) - B_c^{-1}(N, a) = \frac{N(z_1^2 - 2z_2 + z_1 + 2N(z_2 - z_1))}{a^3} + O\left(\frac{1}{a^4}\right). \quad (30)$$

Now, since $z_i = z_0 + O(1/a)$, we see from (30) that

$$B_H^{-1}(N, a) - B_c^{-1}(N, a) = \frac{Nz_0(z_0 - 1)}{a^3} + O\left(\frac{1}{a^4}\right)$$

or

$$B_c(N, a) = \frac{B_H(N, a)}{1 - [B_H(N, a)Nz_0(z_0 - 1)/a^3]} + O\left(\frac{1}{a^4}\right). \quad (31)$$

In addition to providing an analytic representation of the asymptotic error, eq. (31) shows explicitly some of the qualitative properties that evolved from our somewhat heuristic development of Hayward's approximation in Section IV. Specifically, we see that for peaked traffic ($z > 1$) Hayward's approximation will underestimate blocking while for smooth traffic ($z < 1$) it will tend to overestimate blocking—at least in the asymptotic region for large a. Moreover, we note that a/z is, indeed, a reasonable accuracy parameter.

VI. CONCLUDING REMARKS

Hayward-type approximations have proved extremely useful in obtaining simple (but often accurate) approximations to various quantities of interest in a multitude of applications. We have given a qualitative development of the original approximation of Hayward for the blocking seen by overflow traffic and shown how the model used for this purpose could be used to obtain extensions and enhancements. In particular, we have shown how to apply Hayward's approximation to more general systems, determine similar types of approximations for other quantities of interest in stochastic server systems, and improve the accuracy of these approximations. For the special case of renewal input to exponential servers, we have given an explicit asymptotic expression for the error introduced by the generalized Hayward approximation which exhibits, in a quantitative manner, the qualitative statements that evolved in our development.

It is hoped that this paper will generate research into this apparently fruitful area of approximation techniques.

VII. ACKNOWLEDGMENTS

I would like to thank A. E. Eckberg, L. J. Forys, H. Heffes, D. L. Jagerman, and J. S. Kaufman for many helpful discussions on the concept of peakedness and its applications. I would also like to thank J. M. Holtzman and W. Whitt for many helpful comments and suggestions.

REFERENCES

1. R. I. Wilkinson, "Theories of Toll Traffic Engineering in the U.S.A.," B.S.T.J., *35*, No. 2 (March 1956), pp. 421–514.
2. R. B. Cooper, *Introduction to Queueing Theory*, New York: MacMillan, 1972.
3. L. Kosten, *Stochastic Theory of Service Systems*, Oxford, England, Pergamon, 1973.
4. J. M. Holtzman and D. L. Jagerman, "Estimating Peakedness from Arrival Counts," 9th International Teletraffic Congress, 1979, Paper 134N.
5. L. Takacs, *Introduction to the Theory of Queues*, New York: Oxford, 1962.
6. J. M. Holtzman, "The Accuracy of the Equivalent Random Method with Renewal Inputs," B.S.T.J., *52*, No. 9 (November 1973), pp. 1673–1679.
7. R. R. Mina, "Some Practical Applications of Teletraffic Theory," 5th International Teletraffic Congress, 1967, pp. 428–434 (appendix by R. Syski).
8. P. J. Burke, "The Overflow Distribution for Constant Holding Times," B.S.T.J., *50*, No. 10 (December 1971), pp. 3195–3210.
9. G. Bretschneider, "Extensions of the Equivalent Random Method to Smooth Traffic," 7th International Teletraffic Congress, 1973, paper 411.
10. R. M. Potter, "The Equivalent Non-Random Method and Restrictions Imposed on Renewal Overflow Systems by the Specification of a Finite Number of Overflow Traffic Moments," 9th International Teletraffic Congress, 1979, Paper 247R.
11. P. Franken and J. Kerstan, "Queueing Systems with an Infinite Number of Server Devices," Operations Forsch Math Statist, *1* (1968), pp. 17–76.
12. M. F. Neuts and S.-Z. Chen, "The Infinite Server Queue with Semi-Markovian Arrivals and Negative Exponential Services," J. Appl. Prob., *9* (1972), pp. 178–184.
13. W. Whitt, "Heavy Traffic Limit Theorems for Queues: A Survey," *Mathematical Methods in Queueing Theory*, Lecture Notes in Economics and Mathematical Systems 98, A. B. Clarke (ed.), Berlin-Heidelberg-New York: Springer-Verlag, 1974.
14. R. I. Wilkinson, *Nonrandom Traffic Curves and Tables*, Traffic Study Center, Bell Laboratories, 1970.
15. D. L. Jagerman, "Some Properties of the Erlang Loss Function," B.S.T.J., *53*, No. 3 (March 1974), pp. 525–551.
16. L. Y. Rapp, "Planning of Junction Networks in a Multi-Exchange Area—Part 1," Ericson Technics, Vol. 20, pp. 77–130, 1964.
17. J. D. C. Little, "A Proof of the Queueing Formula $L = \lambda W$," Operations Research, *9* (1961), pp. 383–387.
18. E. Brockmeyer, "Det Simple Overflowproblem I Telefontrafikteorien," Teleknik, 1954.
19. H. Heffes, unpublished work.