THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 59, No. 6, July–August
Printed in U.S.A.

A Class of Data Traffic Processes— Covariance Function Characterization and Related Queuing Results

By H. HEFFES

(Manuscript received December 28, 1978)

While the "call" or "session" is the basic entity that is set up in many data traffic applications, the performance analysis of data network elements depends on the internal units of traffic into which calls are decomposed. In a packet-switching network, the packet represents the basic internal unit of traffic, and packets from different calls time-share facilities and contend for network resources, giving rise to queuing delays. In this paper, we consider the problem of characterizing the doubly stochastic packet process resulting from a superposition of call types, each type having a stochastically varying number of calls in progress. We obtain statistical properties of the process and use them to obtain an approximating process, based in part upon time constants associated with the packet-rate covariance function. We discuss existing queuing models dealing with this approximating class of inputs and present results showing the effect of call and packet traffic parameters on queuing performance.

I. INTRODUCTION

In this paper, our interest is in calculating delay information when packets resulting from a collection of calls in progress contend for a data network resource; e.g., transmission delays experienced by packets resulting from the collection of virtual circuits on a trunk in a packet switching network. Since the number and types of calls in progress are stochastically varying, the resulting packet process is doubly stochastic. To do an exact analysis when this class of processes is offered to a queuing system, e.g., representing a trunk or packet switch, can be quite intractable. The approach we take is to obtain queuing results by approximating the packet process by a simpler doubly stochastic

process that captures the important statistical properties and that is amenable to analysis when offered to a queuing system.

In this paper, we (*i*) characterize the overall packet arrival process in terms of the covariance function of the packet rate, which is itself a stochastic process, and its moments, (*ii*) approximate it by a simpler process which matches the above characterization, and (*iii*) analyze the performance of the queuing system whose input is the resulting packet process.

The characterization depends on the statistics of the call origination process, call durations, the rate of packet arrivals per call, and the discipline for limiting the number of calls in progress.

The approximating process is a simple doubly stochastic Poisson process[1] where the intensity of a Poisson process varies with the state of a continuous-time Markov chain. This process has been studied in the literature,[2,3] and performance results are available when this process is offered to a queuing system.

The single server with general-service time distribution is treated in Ref. 2, and the single- and multi-server cases with exponentially distributed service times are considered in Refs. 4 to 6. Special cases of the model in Ref. 2 are studied in Refs. 7 to 9, which all consider the single server with exponential service-time distribution.

The approximating process is a generalization of a superposition of a batched Poisson process and a Poisson process, which was used in Ref. 10 to study buffer behavior. Other single-server queuing models that deal with correlated arrivals appear in Ref. 11, which uses a branching Poisson process[1] model for the input and a diffusion approximation to obtain queuing performance and in Refs. 12 and 13 where discrete-time queues with correlated inputs are studied for the single-server and multiserver cases, respectively.

Section II of this paper presents the traffic model for the packet arrival process resulting from a call mix and Section III obtains the approximating process in terms of the statistical parameters of the packet-arrival process. These statistical parameters are derived in Section IV for two disciplines for limiting the number of calls in progress. In Section V some of the queuing models from the literature, which treat our approximating process as the input process, are discussed and closed-form analytic results are presented for queuing performance (e.g., delay and queue size distributions) using one of these models. Section VI compares queuing results obtained from our analysis to exact results for special cases.

In Section VII, we present numerical examples that illustrate the behavior of queue size and delay statistics (e.g., mean queue size and tail of delay distribution) as a function of traffic mix, limits placed on

the number of calls in progress, mean call duration, and the mean number of packets per call.

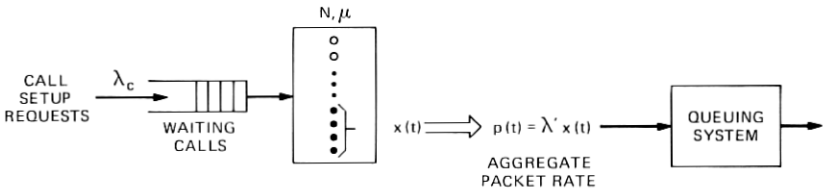
The appendices contain details of the investigations.

II. TRAFFIC MODEL

We consider requests for call setup that arrive at random and, upon arrival, a decision is made whether or not to set up the call. This decision may be based on the number of calls already in progress, and possible disposition policies for requests that are not immediately set up range from placing them in queue to clearing them. The call session duration is random and defined to be measured from the time a call is set up until its completion. During the lifetime of a call, packets arrive at random to a given network resource, and packets from different calls contend for the resource, giving rise to queuing delays. Our interest is in the packets resulting from the collection of calls in progress.

In this paper, we assume that requests for call setup arrive as a Poisson process with rate λ_c. If, upon the arrival of a call setup request, N calls are in progress, the call setup is not initiated. Two disciplines are considered for limiting the number of calls in progress. In the first discipline, the call setup request is placed in queue, i.e., these blocked calls are delayed (BCD) and, when a call session terminates, a call setup request is taken from the queue. In the second discipline, call setup requests that arrive while N calls are in progress are cleared (BCC) and assumed to be lost from the system.

If a call setup request arrives when less than N calls are in progress or if a call setup request is taken out of queue in the BCD case, then the call setup is initiated. Measured from the time of initiation of call setup, the duration of the call session is assumed to be exponentially distributed with mean value μ^{-1}. While the call is in progress, i.e.,



λ_c CALL ARRIVAL RATE
μ^{-1} MEAN CALL DURATION
N LIMIT ON CALLS IN PROGRESS
x (t) NUMBER OF CALLS IN PROGRESS AT TIME t
λ' PACKET RATE/CALL IN PROGRESS

Fig. 1—Call and packet processes (blocked calls delayed).

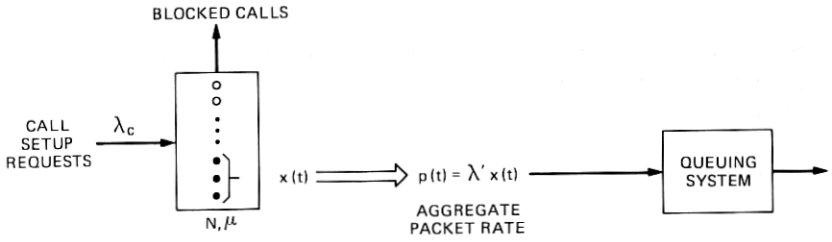


Fig. 2—Call and packet processes (blocked calls cleared).

during the session, packets arrive as a Poisson process with rate λ'. If we denote $x(t)$ as the number of calls in progress, then the packet rate corresponding to this collection of calls in progress is $p(t) = \lambda' x(t)$.

Figure 1 schematically shows the call and packet processes for the blocked-calls-delayed case. The number of call sessions in progress is modeled as the number of busy servers in an M/M/N queuing system, and the M/M/N delay distribution corresponds to the call setup delay distribution. Figure 2 shows the blocked calls cleared case where the M/M/N blocking corresponds to a call setup request being rejected. The figures show the resulting packet processes offered to a queuing system.

In general, we must deal with the packet process resulting from a mix of call types, each type with its own set of traffic parameters, call limit, and queue discipline. This is shown schematically in Fig. 3, where the overall stochastic packet rate process is

$$p(t) = \sum_{i=1}^{J} p_i(t) = \sum_{i=1}^{J} \lambda'(i) x_i(t), \tag{1}$$

where J is the number of call types, $x_i(t)$ is the number of type i calls in progress, and $\lambda'(i)$ is the rate of packet arrivals for a type i call in progress. We note that each call type also has its own calling rate $\lambda_c(i)$, mean call duration $\mu(i)^{-1}$, and call limit N_i. Throughout this paper, it is assumed that packet and call processes corresponding to different call types are statistically independent.

The process of packet arrivals is a doubly stochastic Poisson process, i.e., a Poisson process where the rate itself is a realization of a stationary, continuous-time stochastic process. The rate process (see Fig. 3) is a fairly complicated process with up to $\prod_{i=1}^{J} (N_i + 1)$ levels. Since exact analysis, when this process is offered to a queuing system, can be quite intractable, we will be interested in constructing an approximating packet process, which is also doubly stochastic but analyzable when offered to a queuing system and which matches the important statistical properties of the $p(t)$ process.

The quantities we use to characterize the process are the moments of $p(t)$ which relate to the distribution of arrival rates at an arbitrary time and the covariance function which relates to how dependent the rate at one instant of time is to the rate at another instant of time. We will be interested in the mean, variance, and third moment of $p(t)$ denoted by m, v, and μ_{3p} and the integral of the covariance function $r(t)$. For a large class of doubly stochastic Poisson processes, the integral of the arrival rate covariance function is directly related to the limiting variance-to-mean ratio of the number of arrivals during a time interval.[1] If $n(T)$ is the number of arrivals in a time interval of length T, then

$$\lim_{T\to\infty} \frac{\text{Var}[n(T)]}{E[n(T)]} = 1 + \frac{2\int_0^\infty r(t)\,dt}{m}. \tag{2}$$

Furthermore, the covariance integral is a useful way of defining a time constant for the process. In particular, the time constant defined by

$$\tau_c = \frac{1}{v}\int_0^\infty r(t)\,dt \tag{3}$$

is such that the exponential covariance function approximation

$$r_e(t) = ve^{-t/\tau_c}, \tag{4}$$

which matches $r(0)$ and the covariance integral, will be seen to be a good approximation to the time function $r(t)$ over a wide range of conditions.

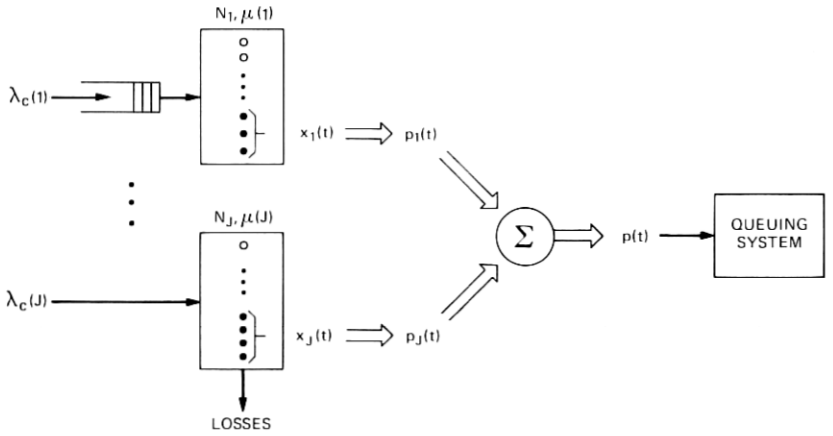


Fig. 3—Superposition of packet processes (delay and clear disciplines).

Since the packet rates corresponding to different call types are independent, τ_c is simply the convex combination

$$\tau_c = \sum_{i=1}^{J} \frac{v_i}{v} \tau_c(i) \tag{5}$$

of call type i time constants, $\tau_c(i)$, where v_i is the variance of the packet rate from type i calls. We note that long time constant call types will only affect the overall time constant if their variances are significant contributors.

Obtaining the moments and time constant of the packet rate corresponding to a given call type is deferred to Section IV. We now turn our attention to obtaining our approximating doubly stochastic Poisson process.

III. THE APPROXIMATING PROCESS

In choosing our approximating process, we are motivated by the desire to have a randomly varying, correlated packet arrival rate. This, together with the exact exponential behavior of the covariance function for no call limiting[14] and the approximate exponential behavior for finite call limiting (as is seen later), suggests using a doubly stochastic Poisson process where the rate process is determined by the state of a continuous-time Markov chain. We call this a Markov-modulated Poisson process or MMP. Because of our desire to construct a process which is amenable to analysis when offered to a queuing system and which has an exponential covariance function, we use a two-state MMP for which simple analytic or algorithmic queuing results are available.

The process is shown schematically in Fig. 4. We have a two-state continuous-time Markov chain where the rates of exiting states 1 and 2 are r_1 and r_2, respectively. When the chain is in state j ($j = 1, 2$), the arrival process is a Poisson process with rate λ_j. We note the equivalence between this process and the superposition of a Poisson and an

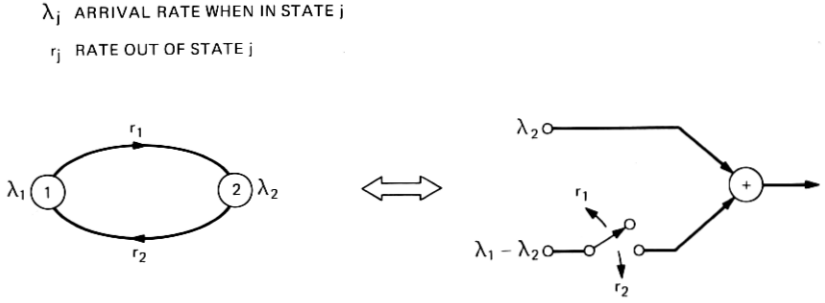


Fig. 4—The two-state MMP and equivalence to P + IPP.

interrupted Poisson process (IPP).[15] The model has four parameters which will be chosen to match m, v, μ_{3p}, and τ_c.*

If we denote

$$\Pi = [\Pi_1, \Pi_2]$$

as the row vector of equilibrium probabilities for the state of the MMP,

$$\Lambda = \begin{bmatrix} \lambda_1 & o \\ o & \lambda_2 \end{bmatrix}, \tag{6}$$

the diagonal matrix of Poisson intensities and

$$A = \begin{bmatrix} -r_1 & r_1 \\ r_2 & -r_2 \end{bmatrix}, \tag{7}$$

then it is known that

$$\Pi = [\Pi_1, \Pi_2] = \frac{1}{r_1 + r_2}[r_2, r_1], \tag{8}$$

the mean arrival rate is

$$m = \lambda_1\Pi_1 + \lambda_2\Pi_2 = \frac{\lambda_1 r_2 + \lambda_2 r_1}{r_1 + r_2}, \tag{9a}$$

and the covariance function of the arrival rate is[16]

$$r(t) = \Pi\Lambda[e^{At} - \mathbf{1}\Pi]\Lambda\mathbf{1},$$

where

$$\mathbf{1} = [1, 1]^T.$$

Evaluating $r(t)$ gives

$$r(t) = \frac{r_1 r_2(\lambda_1 - \lambda_2)^2}{(r_1 + r_2)^2} e^{-(r_1+r_2)t}$$

and

$$v = \frac{r_1 r_2(\lambda_1 - \lambda_2)^2}{(r_1 + r_2)^2}. \tag{9b}$$

The third moment is clearly

$$\mu_{3p} = \frac{\lambda_1^3 r_2 + \lambda_2^3 r_1}{r_1 + r_2}, \tag{9c}$$

* We note that these can represent the four parameters of the processes of Figs. 1 or 2, in which case they would be indexed with the call type, or the aggregated process of Fig. 3.

and the time constant for the process is

$$\tau_c = \frac{1}{v}\int_0^\infty r(t)\,dt = \frac{1}{r_1 + r_2}. \tag{9d}$$

Equations (9) can be inverted to yield the desired parameters

$$r_1 = \frac{1}{\tau_c(1+\eta)} \tag{10a}$$

$$r_2 = \frac{\eta}{\tau_c(1+\eta)} \tag{10b}$$

$$\lambda_1 = m + \sqrt{v/\eta} \tag{10c}$$

and

$$\lambda_2 = m - \sqrt{v\eta}, \tag{10d}$$

where

$$\eta = 1 + \frac{\delta}{2}\left[\delta - \sqrt{4+\delta^2}\right] \tag{11a}$$

and

$$\delta = \frac{\mu_{3p}^*}{v^{3/2}} = \frac{\mu_{3p} - 3mv - m^3}{v^{3/2}}. \tag{11b}$$

(The quantity δ corresponds to a measure of skewness defined by eq. 3.89 in Ref. 17.) It can be simply shown, using results in Ref. 18, that the parameters defined by (10) are nonnegative. Note that the equilibrium state probabilities for the two-state MMP are

$$\Pi = \frac{1}{1+\eta}[\eta, 1]. \tag{12}$$

We note that, with the above parameters, if $n(T)$ denotes the number of arrivals over a time interval of duration T then

$$\lim_{T\to\infty} \frac{\mathrm{Var}[n(T)]}{E[n(T)]} = 1 + 2\frac{v}{m}\tau_c. \tag{13}$$

Thus for a given time constant the variability of the number of arrivals is directly related to the variability of the arrival rate. Also, for a given arrival rate variability, longer time constants imply more variability in the number of arrivals.

We can write the two-state MMP as the superposition of a Poisson process with intensity $\lambda_2 \geq 0$ and an IPP with parameters

$$\lambda_s = \lambda_1 - \lambda_2 = \sqrt{v/\eta}\,(1+\eta) \tag{14a}$$

$$\omega_s = r_2 = \eta\tau_c^{-1}/(1+\eta) \tag{14b}$$

and

$$\gamma_s = r_1 = \tau_c^{-1}/(1 + \eta). \tag{14c}$$

This is depicted in Fig. 4, where λ_s represents the Poisson intensity into the switch, ω_s^{-1} is the mean off (open) time of the switch, and γ_s^{-1} is the mean on (closed) time of the switch.

With this identification, we note that we can use any queuing results available (e.g., Ref. 7) for Poisson plus IPP input. This is done in a later section.

We now consider the problem of determining the packet rate statistics from the underlying traffic parameters.

IV. STATISTICAL CHARACTERIZATION OF THE PACKET PROCESS

To determine the packet rate statistics, we must look at the rates due to each call type. For the aggregate process of Fig. 3, we have

$$m = \sum_{i=1}^{J} m_i, \qquad v = \sum_{i=1}^{J} v_i; \qquad \mu_{3p}^* = \sum_{i=1}^{J} \mu_{3p}^*(i)$$

and

$$\tau_c = \sum_{i=1}^{J} \frac{v_i}{v} \tau_c(i), \tag{15}$$

where m_i, v_i, $\mu_{3p}^*(i)$, and $\tau_c(i)$ are the mean, variance, third central moment, and time constant for the type i packet rate. These quantities are simply related to the corresponding moments and covariance function of the number of type i calls in progress, denoted by M_i, V_i, $\mu_3^*(i)$ and $R_i(t)$, by

$$m_i = \lambda'(i)M_i, \qquad v_i = (\lambda'(i))^2 V_i, \qquad \mu_{3p}^*(i) = (\lambda'(i))^3 \mu_3^*(i), \tag{16a}$$

and

$$\tau_c(i) = \frac{1}{V_i} \int_0^\infty R_i(t)\, dt. \tag{16b}$$

The quantities M_i, V_i, $\mu_3^*(i)$ and $\int_0^\infty R_i(t)\, dt$ depend on the call traffic parameters, the call limit, and the queue discipline for achieving this limit. We first consider the case where blocked call setup requests are delayed.

4.1 Packet rate statistics—blocked calls delayed

This situation is depicted in Fig. 1. If we denote the call offered load by

$$a_i = \lambda_c(i)/\mu(i),$$

then clearly

$$M_i = a_i. \tag{17a}$$

From Appendix B, we get

$$V_i = a_i(1 - C_i) \tag{17b}$$

$$\mu_3^*(i) = V_i - 2a_i(N_i - a_i)C_i \tag{17c}$$

and

$$\int_0^\infty R_i(t)\, dt = \frac{1}{\mu(i)} a_i,$$

where $C_i = C(N_i, a_i)$ is the Erlang C function. Note that this results in the time constant

$$\tau_c(i) = \frac{1}{\mu(i)(1 - C_i)}. \tag{17d}$$

When eqs. (17) are used in (16), the packet rate statistics for this BCD call type are completely specified.

We remark that the exponential function which agrees with $R_i(t)$ at $t = 0$ and has the same covariance integral is

$$R_{e_i}(t) = a_i(1 - C_i)e^{-(1-C_i)\mu(i)t}. \tag{18}*$$

In Appendix A, $R_{e_i}(t)$ is shown to be a good approximation to the exact $R_i(t)$ over a wide range of loads. One comparison from Appendix A is shown in Fig. 5a. The parameters are such that over 16 percent of arriving calls experience call setup delays ($C_i = 0.167$). Comparisons for lower and higher call setup delay probabilities, shown in Figs. 5b and 5c, are discussed in Appendix A.

4.2 Packet rate statistics—blocked calls cleared

For this case, depicted in Fig. 2, the results of interest are standard results from teletraffic theory[14] and are summarized here.†

$$M_i = a_i(1 - B_i) \tag{19a}$$

$$V_i = M_i\left[1 - a_iB_i\left(\frac{N_i}{M_i} - 1\right)\right] \tag{19b}$$

$$\mu_3(i) = (a_i^2 + 3a_i + 1)M_i - N_i(N_i + a_i + 2)(a_i - M_i)$$

* Beneš (Ref. 14) has used a single exponential to approximate the covariance function of the number of busy servers in a blocking system.

† Also see Appendix C.

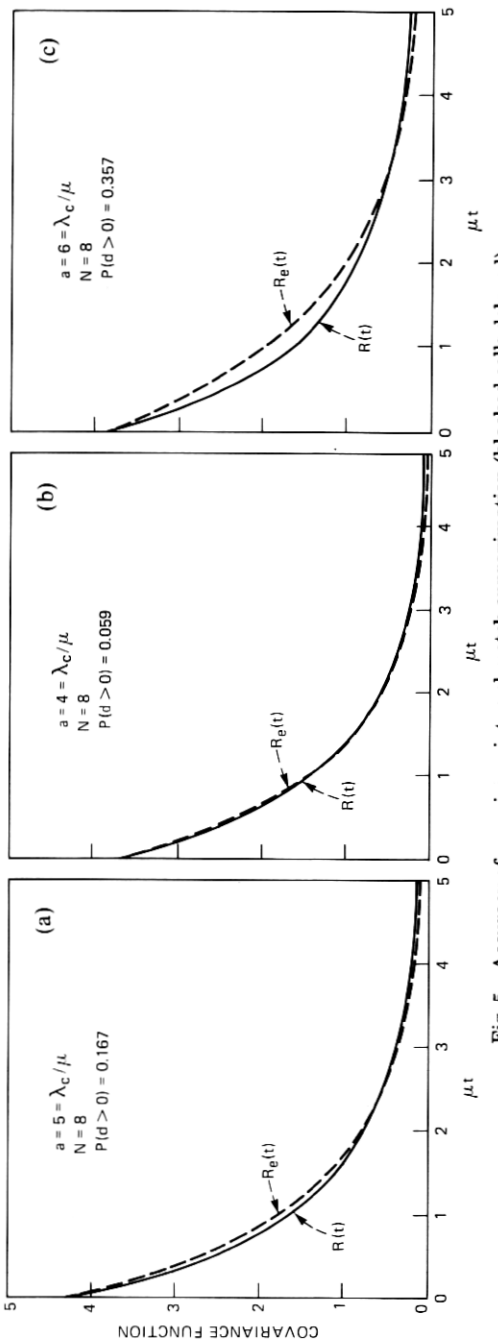


Fig. 5—Accuracy of covariance integral match approximation (blocked calls delayed).

and

$$\mu_3^*(i) = \mu_3(i) - 3M_iV_i - M_i^3, \tag{19c}$$

where $B_i = B(N_i, a_i)$ is the Erlang B function.

From Appendix C and Refs. 14 and 19, we obtain the time constant

$$\tau_c(i) = \mu(i)^{-1}\left[1 - \frac{a_iN_iB_i}{V_i}\left\{1 + \frac{M_i}{N_i} + 2\alpha_{N_i}(a_i) + \alpha'_{N_i}(a_i)\right\}\right], \tag{19d}$$

where α and α' satisfy the recursions

$$\frac{1}{\alpha_{j+1}(a_i)} = -(j + 1 + a_i)/a_i - (j/a_i)\alpha_j(a_i)$$

and

$$\alpha'_{j+1}(a_i) = [(j\alpha'_j(a_i) - 1)/a_i][\alpha_{j+1}(a_i)]^2,$$

with initial conditions

$$\alpha_1(a_i) = -a_i/(1 + a_i)$$

and

$$\alpha'_1(a_i) = -a_i/(1 + a_i)^2.$$

Equations (19), when used in (16), completely specify the packet rate statistics for this BCC call type.

We remark that the exponential function which agrees with $R_i(t)$ at $t = 0$ and has the same covariance integral is

$$R_{e_i}(t) = V_ie^{-t/\tau_c(i)}. \tag{20}*$$

In Appendix C, $R_{e_i}(t)$ is shown to be a good approximation to the exact $R_i(t)$ over a wide range of loads. One comparison from Appendix C is shown in Fig. 6a. The parameters are such that 42 percent of the call setup requests are blocked. At lower loads, the results are even closer (see Fig. 6b). Appendix C discusses some of the properties of the above exponential approximation.

4.3 Special cases

The above results completely specify the overall packet rate statistics and therefore the MMP approximation process of Section III. Before proceeding to use these results in a queuing analysis, we consider some special cases. In all these situations, we present the

* This uses a somewhat different time constant from that used in the exponential covariance function approximation of Beneš (Ref. 14). The covariance integral is studied, however, in Chapter 6, Section 6 of this reference, where the variance of time averages is investigated.

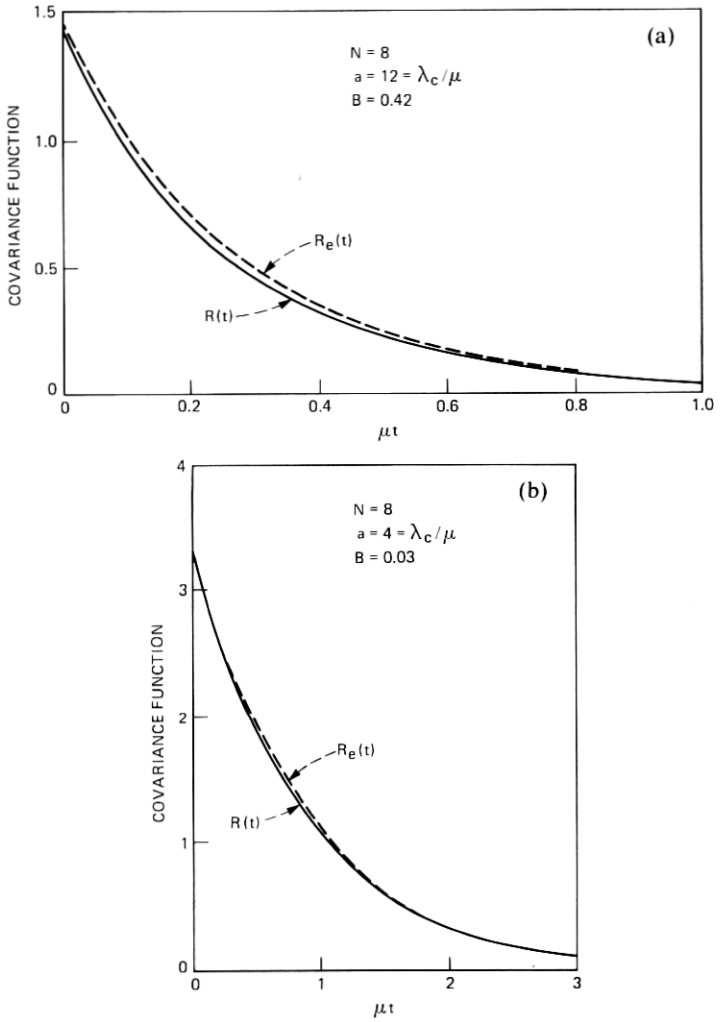


Fig. 6—Accuracy of covariance integral match approximation (blocked calls cleared).

results for an MMP approximation to the packet rate process for a given call type and thus drop the call type index (i) from the notation.

4.3.1 Case 1 – N = ∞

This case corresponds to no limit placed on the number of calls in progress. Here, $m = \lambda' a$, $v = (\lambda')^2 a$, $\mu_{3p}^* = (\lambda')^3 a$, and $\tau_c = \mu^{-1}$. From (10) and (11), the parameters for our approximating process are given

by

$$\lambda_1 = \lambda'\left(\frac{1 + 2a + \sqrt{1 + 4a}}{2}\right), \tag{21a}$$

$$\lambda_2 = \lambda'\left(\frac{1 + 2a - \sqrt{1 + 4a}}{2}\right), \tag{21b}$$

$$r_1 = \frac{\mu}{2}\left[1 + \frac{1}{\sqrt{1 + 4a}}\right], \tag{21c}$$

and

$$r_2 = \frac{\mu}{2}\left[1 - \frac{1}{\sqrt{1 + 4a}}\right]. \tag{21d}$$

We note that the geometric mean of the arrival rates for our two-state MMP approximation is the expected arrival rate, i.e.,

$$\sqrt{\lambda_1 \lambda_2} = m = \lambda' a. \tag{22}$$

4.3.2 Case 2 – N = 1; blocked calls cleared

Here the two-state MMP is exact, with parameters $\lambda_1 = \lambda'$, $\lambda_2 = 0$, $r_1 = \mu$, and $r_2 = \lambda_c$ as expected, and degenerates into an IPP.

4.3.3 Case 3 – N = 1; blocked calls delayed

Here the exact process is a two-state process with arrival rates of λ' and zero. The on time of the exact process is distributed as the busy period in an M/M/1 queue, and the off time is exponentially distributed as the call interarrival time.

The two-state MMP, which has exponentially distributed on and off times corresponding to the above process, has the same two levels of arrival rates $\lambda_1 = \lambda'$, $\lambda_2 = 0$ and the same probabilities of being in each of the states as the exact process. The transition rates for our MMP are $r_1 = \mu(1 - a)^2$, $r_2 = \lambda_c(1 - a)$. We note that r_1^{-1} exceeding the mean length of the M/M/1 busy period is required to match the time constant.

We now turn our attention to available queuing results for systems with our class of doubly stochastic Poisson approximating processes as inputs.

V. QUEUING MODELS

Several queuing models that deal with the Markov-modulated Poisson process as an input appear in the literature. A single-server queue

with general service time distribution and multilevel input has been studied by Neuts,[2] where algorithmic results are presented,* and more recently in Ref. 4, where the exponential service time case is presented. Kuczura[7] treats the superposition of a Poisson and interrupted Poisson process (IPP), which is equivalent to a two-state MMP, into a single exponential server, as do Yechiali and Naor.[9] Some multiserver results are given in Ref. 5, and Eckberg[6] has developed a computational procedure for the multiserver case where the servers have exponentially distributed service times. In the remainder of this section, we present the results of Ref. 7 as applied to our problem.

Before presenting the results, we note that superimposing a Poisson process (with intensity λ_p) to a two-state MMP results in a two-state MMP with parameters

$$\lambda_1 = m + \lambda_p + \sqrt{v/\eta},$$
$$\lambda_2 = m + \lambda_p - \sqrt{v\eta}.$$

The r_1 and r_2 parameters remain unchanged. Denoting

$$\rho = m + \lambda_p, \tag{23a}$$

we have

$$\lambda_1 = \rho + \sqrt{v/\eta} \tag{23b}$$

and

$$\lambda_2 = \rho - \sqrt{v\eta}. \tag{23c}$$

The results of this section are written in terms of the quantities given in eqs. (23) and r_1 and r_2 given by (10a) and (10b).

One problem considered in Ref. 7 is the queuing analysis for the Poisson plus IPP input to a single server with exponentially distributed service times.† The solution for queue size and delay distributions as seen by the Poisson arrivals, the IPP arrivals, or an arbitrary arrival are in terms of the positive, real roots of the cubic

$$T(u) = \lambda_1\lambda_2 u^3 - [\lambda_1\lambda_2 + \lambda_1 + \lambda_2 + \rho(r_1 + r_2)]u^2 + [\lambda_1 + \lambda_2 + r_1 + r_2 + 1]u - 1 = 0, \tag{24a}‡$$

which satisfy (for $\lambda_2 \neq 0$, $\lambda_1 \neq \lambda_2$)§

$$0 < u_1 < 1 < u_2 < \lambda_2^{-1} < u_3 < \infty. \tag{24b}$$

---

* References 8 and 20 present a computational procedure for a special multilevel process into a single server with exponentially distributed service times.

† The unit of time is mean service times. This requires scaling the rates λ_1, λ_2, r_1, r_2, m, λ_p and ρ by μ_s, the mean service rate.

‡ $\rho = m + \lambda_p = (\lambda_1 r_2 + \lambda_2 r_1)/(r_1 + r_2) < 1$ is the server utilization.

§ For $\lambda_2 = 0$, the input process is an IPP, the cubic degenerates to a quadratic. For $\lambda_1 = \lambda_2$, the process is Poisson.

The reciprocals of these roots

$$\omega_i = u_i^{-1} \tag{24c}$$

satisfy

$$0 < \omega_3 < \lambda_2 < \omega_2 < 1 < \omega_1 < \infty. \tag{24d}$$

The distributions are also in terms of the quantities

$$A_2 = \frac{\omega_2(1-\omega_3)(\omega_2 - \lambda_2 u_1)}{(1-\lambda_2 u_1)(\omega_2 - \omega_3)} \tag{25a}$$

and

$$A_3 = \frac{\omega_3(1-\omega_2)(\omega_3 - \lambda_2 u_1)}{(1-\lambda_2 u_1)(\omega_2 - \omega_3)}. \tag{25b}$$

Denoting

Q^I = number of customers in system (waiting and in service) just prior to an IPP arrival,

Q^P = number of customers in system just prior to a λ_2 Poisson arrival,

Q^A = number of customers in the system just prior to an arbitrary arrival,

and lower case q^I, q^P, and q^A, the corresponding quantities not including the customer in service, we have

$$P_j^I = \Pr\{Q^I = j\} = (u_2 - 1)A_2\omega_2^j - (u_3 - 1)A_3\omega_3^j, \tag{26a}$$

$$P_j^P = \Pr\{Q^P = j\} = (1-\rho)\lambda_2^j + \frac{(\rho - \lambda_2)A_2(u_2 - 1)}{\lambda_2 - \omega_2}(\lambda_2^j - \omega_2^j) + \frac{(\rho - \lambda_2)A_3(u_3 - 1)}{\lambda_2 - \omega_3}(-\lambda_2^j + \omega_3^j), \tag{26b}$$

and

$$P_j^A = \Pr\{Q^A = j\} = \left(1 - \frac{\lambda_2}{\rho}\right)P_j^I + \frac{\lambda_2}{\rho}P_j^P. \tag{26c}$$

The mean values of these quantities are

$$\bar{Q}^I = E\{Q^I\} = \frac{1}{u_3 - 1} + \frac{1}{u_2 - 1} - \frac{\lambda_2 u_1}{1 - \lambda_2 u_1}, \tag{27a}$$

$$\bar{Q}^P = E\{Q^P\} = \frac{\lambda_2}{1-\lambda_2} + \frac{(\rho - \lambda_2)}{1-\lambda_2}(1 + \bar{Q}^I), \tag{27b}$$

and

$$\bar{Q}^A = E\{Q^A\} = \left(1 - \frac{\lambda_2}{\rho}\right)\bar{Q}^I + \frac{\lambda_2}{\rho}\bar{Q}^P. \tag{27c}$$

The tail probabilities are given by

$$P^I_{>j} = \Pr\{Q^I > j\} = A_2\omega_2^j - A_3\omega_3^j, \tag{28a}$$

$$P^P_{>j} = \Pr\{Q^P > j\} = \frac{(1-\rho)\lambda_2^{j+1}}{1-\lambda_2} + \frac{(\rho-\lambda_2)A_2(u_2-1)}{\lambda_2-\omega_2}\left[\frac{\lambda_2^{j+1}}{1-\lambda_2} - \frac{\omega_2^{j+1}}{1-\omega_2}\right] + \frac{(\rho-\lambda_2)A_3(u_3-1)}{\lambda_2-\omega_3}\left(\frac{-\lambda_2^{j+1}}{1-\lambda_2} + \frac{\omega_3^{j+1}}{1-\omega_3}\right) \tag{28b}$$

and

$$P^A_{>j} = \Pr\{Q^A > j\} = \left(1-\frac{\lambda_2}{\rho}\right)P^I_{>j} + \frac{\lambda_2}{\rho}P^P_{>j}. \tag{28c}$$

It is a simple matter to obtain the corresponding results for the number of customers in the queue (i.e., not including customer in service). The mean values, for example, are given by

$$\bar{q}^I = E\{q^I\} = \bar{Q}^I - (A_2 - A_3) \tag{29a}$$

$$\bar{q}^P = E\{q^P\} = \bar{Q}^P - \rho \tag{29b}$$

and

$$\bar{q}^A = E\{q^A\} = \left(1-\frac{\lambda_2}{\rho}\right)\bar{q}^I + \frac{\lambda_2}{\rho}\bar{q}^P. \tag{29c}$$

To write the results for delay distribution, we distinguish between
d_I = the delay experienced by an IPP arrival,
d_P = the delay experienced by a λ_2 Poisson arrival, and
d_A = the delay experienced by an arbitrary arrival.
The complementary delay distributions are given by

$$\bar{D}_I(t) = \Pr\{d_I > t\} = A_2 e^{-(1-\omega_2)t} - A_3 e^{-(1-\omega_3)t}, \tag{30a}*$$

$$\bar{D}_P(t) = \rho e^{-(1-\lambda_2)t} + \frac{(\rho-\lambda_2)A_2}{\lambda_2-\omega_2}(e^{-(1-\lambda_2)t} - e^{-(1-\omega_2)t}) + \frac{(\rho-\lambda_2)A_3}{\lambda_2-\omega_3}(e^{-(1-\omega_3)t} - e^{-(1-\lambda_2)t}) \tag{30b}$$

and

$$\bar{D}_A(t) = \Pr\{d_A > t\} = \left(1-\frac{\lambda_2}{\rho}\right)\bar{D}_I(t) + \frac{\lambda_2}{\rho}\bar{D}_P(t). \tag{30c}$$

* Recall that the unit of time is the mean service time.

The mean delays are

$$\bar{d}_I = E\{d_I\} = \frac{A_2}{1-\omega_2} - \frac{A_3}{1-\omega_3}, \tag{31a}$$

$$\bar{d}_P = E\{d_P\} = \frac{1}{1-\lambda_2}(\rho + (\rho - \lambda_2)\bar{d}_I), \tag{31b}$$

and

$$\bar{d}_A = E\{d_A\} = \left(1 - \frac{\lambda_2}{\rho}\right)\bar{d}_I + \frac{\lambda_2}{\rho}\bar{d}_P. \tag{31c}$$

Clearly, since the unit of time is the mean service time, the results of (31) are identical to the corresponding results of (27). We note that the statistical quantities as seen by an arbitrary arriving customer are indexed with the letter A (e.g., $\bar{d}_A$, $\bar{Q}^A$, etc.), whereas the statistical quantities as seen by the Poisson arrivals are indexed with the letter P (e.g., $\bar{D}_P(t)$, $\bar{Q}^P$, etc.).

We further note that the results are very easy to compute, the most difficult operation being to obtain the roots of the cubic (24a).

This completes the specification of the two-state MMP/M/1 solution. Numerical results using this analysis are presented in Section VII.

VI. ACCURACY RESULTS

We consider the accuracy question by comparing our results with known computational results for special cases. In addition, we discuss limiting cases where the results become exact.

As mentioned in the previous section, computational results are available when a multilevel process is offered to a queuing system and the number of levels is not large. For example, in Ref. 8, special cases of the process shown in Fig. 3, superimposed with a Poisson process, is offered to a single server with exponentially distributed service times. These special cases correspond to $N_i = 1$ for all $i = 1, 2, \cdots J$ superimposed with a Poisson process of rate λ_p. Blocked calls are cleared. The numerical results are for $J \leq 8$.*

In Fig. 7, we show the mean queue size (including the customer being served) as seen by the Poisson arrivals (i.e., $\bar{Q}^P$) as a function of the number of levels of arrival rates (i.e., $J + 1$). Results are presented for occupancies of $\rho = 0.6$ and $\rho = 0.8$. Other parameters for the processes are $\lambda'/\mu = 6$ the mean number of arrivals per call, background Poisson traffic of $\lambda_p = \rho/8$, and utilization due to one call in progress of $\lambda' = 0.2$. The exact results, which were obtained from Fig. 1 of Ref. 8, are seen to be indistinguishable from our approximate results, even

* I.e., up to nine levels of arrival rates.

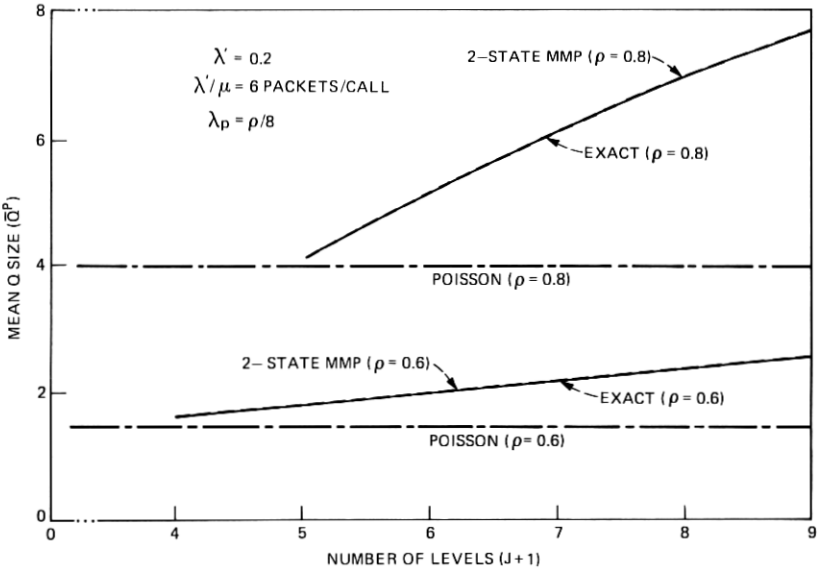


Fig. 7—Accuracy of queuing results.

where both differ significantly from the Poisson result of $\rho/(1-\rho)$.

We note that, since the mean number of packets/call in this example was small, it is of interest to determine the accuracy for larger values of λ'/μ. In Fig. 8 we show the mean queue length results ($\bar{Q}^P$) as a function of the mean number of packets/call. Other parameters are $J+1=6$ levels of arrival rates, $\lambda'=0.2$ occupancy due to one call in progress, and $\lambda_p=0$ background Poisson traffic. Results are shown for an occupancy of $\rho=0.8$. We note that the exact results, which are taken from Fig. 4 of Ref. 8, are closely approximated* by our results over a wide range of mean packets/call.

As mentioned earlier, for a single call type with a call limit of one call in progress and blocked calls cleared superimposed with a Poisson process, our results become exact. Also, if one considers the case where the calling rate λ_c and the mean number of packet arrivals per call λ'/μ are held fixed and the call duration approaches zero, then the limiting process corresponds to a batched Poisson process with geometric batch size distribution as does our approximating process. Results for both the blocked calls delayed and cleared models then become exact.

* Comparisons for lower occupancies are also very close. For the class of processes considered in Ref. 20, an approximation is presented which can significantly underestimate queuing.

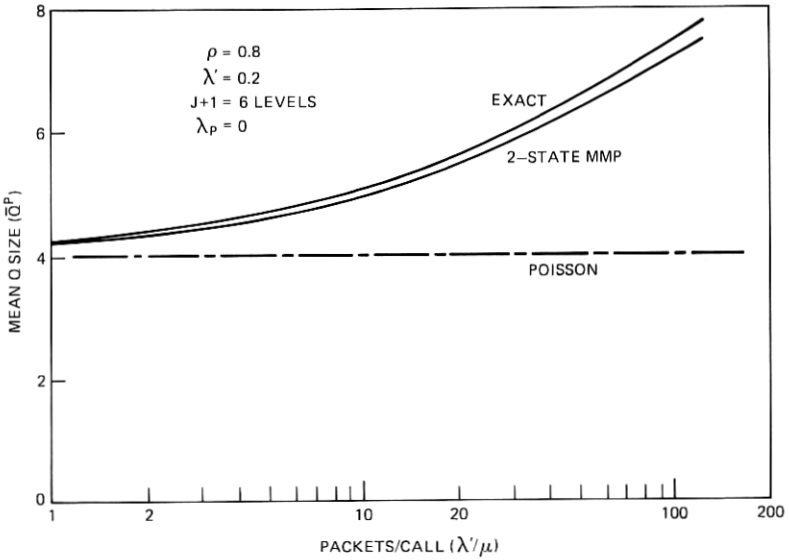


Fig. 8—Accuracy results.

The accuracy comparisons presented are limited in that the number of arrival rate levels was not large and the blocked calls delayed results were not directly validated. An indirect accuracy check we have for the BCD case was the accuracy of the covariance function results.

VII. NUMERICAL RESULTS

In this section, we present some numerical results which show the effect of traffic mix on queuing system performance and capacity. We also show the effect of controlling the packet rate by controlling the maximum allowable number of calls in progress. We see that queuing delays can be quite sensitive to the traffic mix and the speed of packet arrivals per call.

In Fig. 9, we consider the superposition of a Poisson process, with intensity λ_p, and the packet process resulting from a call type with no call limit ($N = \infty$), a mean call duration of $\mu^{-1} = 1000$,* and with mean number of packets per call of $\lambda'/\mu = 240$. Results are shown for a range of mixes λ_p/ρ.

We note the extreme sensitivity of the mean queue size, $\bar{Q}^A$, to the traffic mix and note the sensitivity of the occupancy at which a given mean queue size is achieved [denote this by ρ_{cap} (mix)] to the mix

* Recall that this corresponds to 1000 mean packet service times.

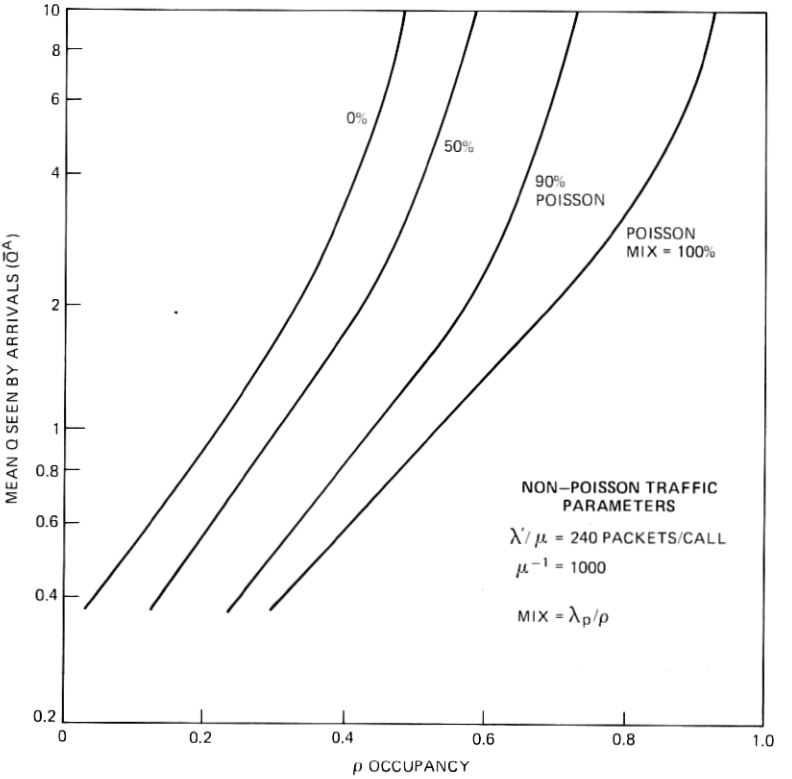


Fig. 9—Effects of traffic mix.

parameter. This is more clearly depicted in Fig. 10 where we see that, for example, with 10-percent non-Poisson traffic the capacity (occupancy at which $\bar{Q}^A = 2$)* is 83 percent of the Poisson capacity. Also shown on the figure are results for a call type with the same mean number of packets per call; however, with twice the mean call duration. Spreading the packets over a longer call duration has the effect of reducing the mean queue size and reducing the sensitivity to mix. We note that, for 10 percent non-Poisson traffic, ρ_{cap} is 90 percent of the Poisson capacity.

In Fig. 11 we consider a call type with an average of 10 packets per call, spread over a mean call duration of 4000 packet service times, for different limits on the maximum allowable number of calls in progress. We note that a limit of 160 calls in progress corresponds to a maximum occupancy of 0.4 and a limit of 240 calls in progress corresponds to a

* The results of Fig. 10 are not very sensitive to the queue size level used to define capacity.

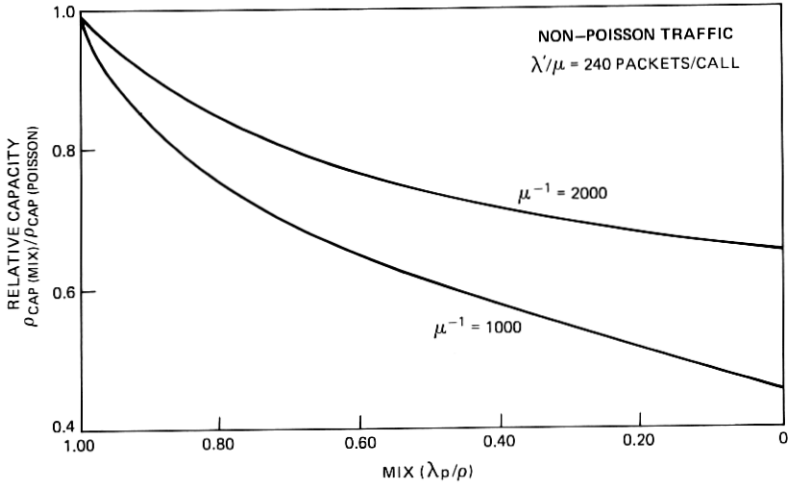


Fig. 10—Relative capacity as function of mix.

maximum occupancy of 0.6. The mean queue size at an arbitrary instant of time $\bar{Q}^P$ is plotted against $\rho_0 = \rho/(1 - B)$, which represents the mean number of packet arrivals per service time that would occur if no calls were rejected, i.e., in general

$$\rho_0 = \sum_{i=1}^{J} a_i \lambda'(i).$$

The limiting values of $\bar{Q}^P$ are closely predictable from M/M/1 results with loads of 0.4 and 0.6. These result in mean queue sizes (including customer in service) of $\rho_{max}/(1 - \rho_{max})$ equal to 0.67 and 1.5, respectively. We also note that the unconstrained results are very close to the M/M/1 results. Figure 12 shows the corresponding results for the tail of the virtual waiting time distribution.

In Fig. 13, we show mean Q size for a call type with a large number of packets/call. The limits of one, two, or three simultaneous calls in progress correspond to occupancy limits of 0.24, 0.48 and 0.72, respectively. The corresponding mean Q limits are 0.32, 0.92, and 2.57. We note that, as expected, the limiting effect is not as sharp as in Fig. 11, for example.

These examples illustrate the nature of the results that can be obtained and the effect of traffic parameters on delays experienced by packets. While we illustrated the results with one or two call types, in any particular application a mix of many call types can be considered. Furthermore, while our examples only considered either no constraint on the number of calls in progress or blocked calls cleared, similar numerical results can be obtained for delayed call setups.

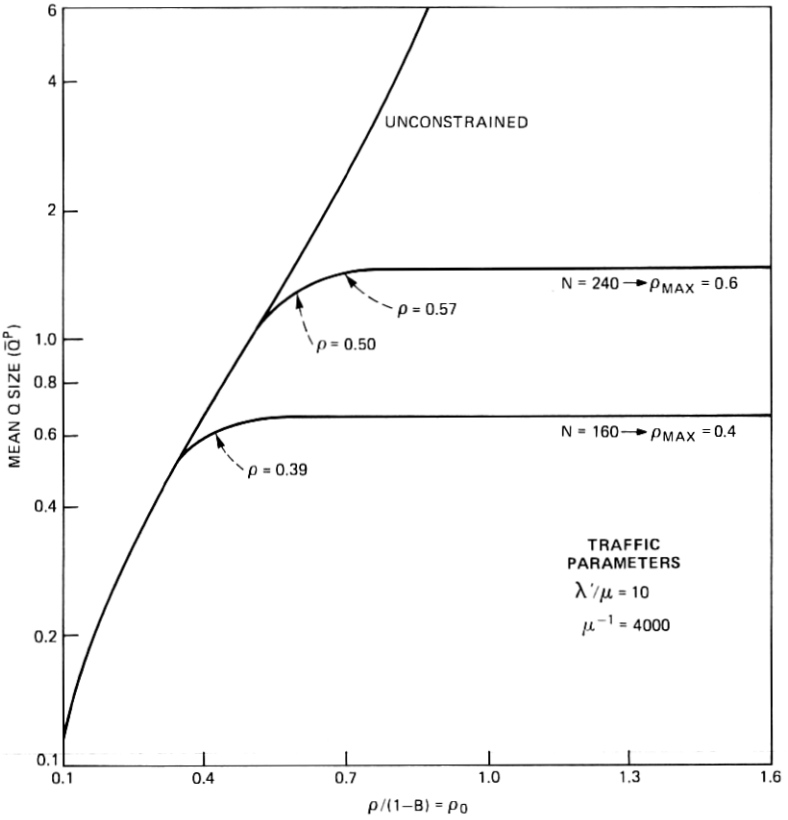


Fig. 11—Effect of call limit (BCC).

VIII. DISCUSSION AND FURTHER WORK

We have obtained a simple characterization for the packet arrival process resulting from a collection of calls in progress. The characterization, which is based in part upon time constants associated with the arrival rate covariance function, depends on the statistics of the requests for call setups and call duration, the rate of packet arrivals per call in progress, and the discipline for limiting the number of calls in progress. Queuing models which deal with the above class of processes have been discussed and numerical examples illustrated the nature and accuracy of the results.

Some direct validation of queuing results for the BCC case and some indirect validation for both BCC and BCD cases were performed. These accuracy results proved to be very favorable.

The results of this paper allow call setup requests either to queue up or to be rejected if more than a given number of calls of its type are in

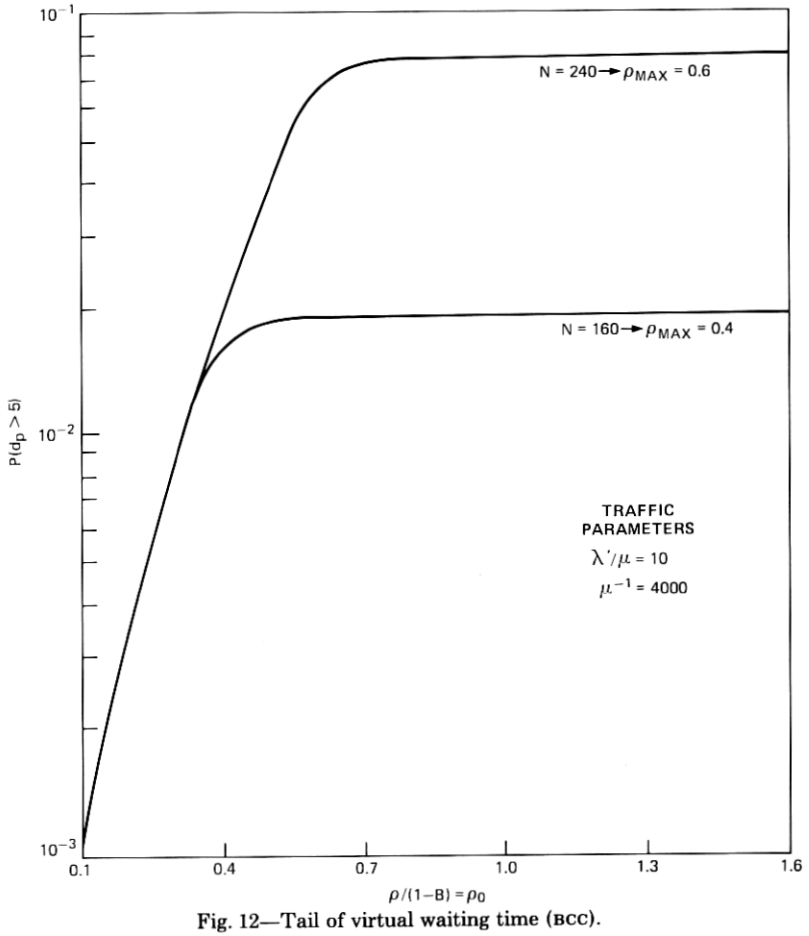


Fig. 12—Tail of virtual waiting time (BCC).

progress. Extension to the loss-delay case, where only a finite number of call setup requests can be queued up, is fairly straightforward, combining the approaches of Appendices A through C. The blocking or delay condition for call setup requests used in this paper allowed up to a maximum number of calls of a particular type to be in progress. Other blocking conditions, such as constraining a linear combination of the number of calls of each type in progress, would allow one to limit the overall packet rate. The moment and covariance function properties of the resulting packet rate are of interest.

As mentioned earlier, while our numerical queuing results are for systems with exponentially distributed service times, results are available for the general service time case.[2] Furthermore, because of the simplicity of the input characterization, one can consider multiserver queuing systems.[5,6]

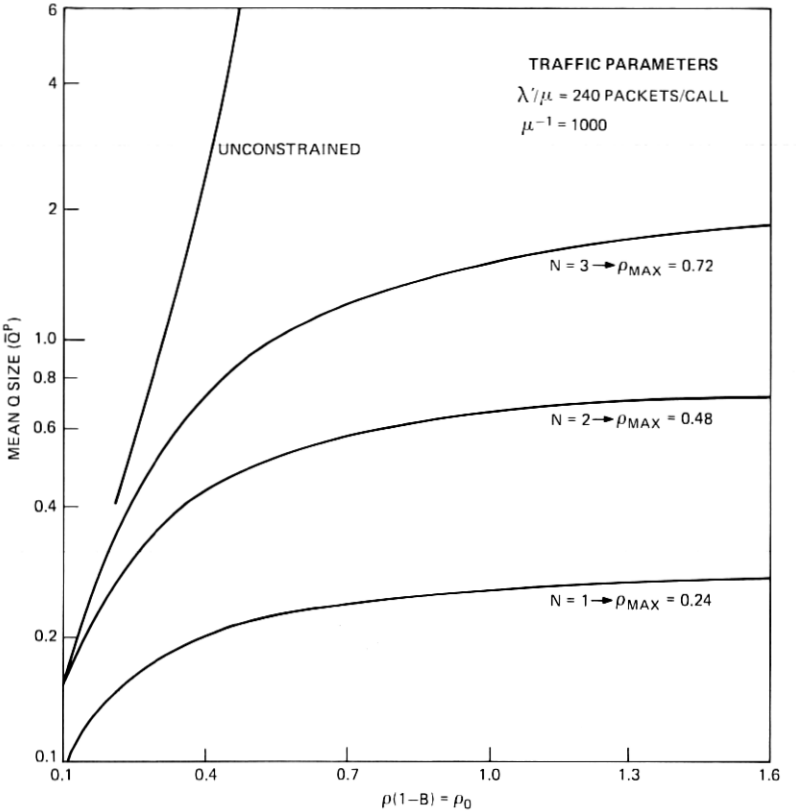


Fig. 13—Effect of call limit (BCC).

IX. ACKNOWLEDGMENTS

Valuable discussions with A. E. Eckberg, D. L. Jagerman, and H. Zucker are gratefully acknowledged, along with the contributions of D. C. D'Angelo, R. H. Harris, and D. F. Shearer to the numerical results.

APPENDIX A

Covariance Function Transform: Delay Case

In this appendix we are interested in obtaining an expression for

$$\tilde{R}(s) = \int_0^\infty R(\tau) e^{-s\tau}\, d\tau, \qquad (32)^*$$

* Numerical inversion allows the accuracy of the exponential covariance function approximation to be determined and evaluating at $s = 0$ gives the covariance integral. The unit of time in this appendix is mean call durations unless otherwise stated.

where $R(\tau)$ is the covariance function of the *number of busy servers* in an M/M/N queuing system. More precisely,

$$R(\tau) = E\{(x(t) - \bar{x})(x(t+\tau) - \bar{x})\}. \tag{33}*$$

Here $\{x(t)\}$ is the stationary process of the number of busy servers and $\bar{x}$ is the mean number of busy servers.

Denoting $n(t)$ as the number of customers in the queuing system, we can write (33) as

$$R(\tau) = \sum_{m=0}^{\infty} \alpha_m P_m \sum_{n=0}^{\infty} \alpha_n p_{mn}(\tau) - \bar{x}^2 \tag{34}$$

where

$$\alpha_m = \begin{cases} m & m \le N \\ N & m > N, \end{cases}$$

$$P_m = \Pr\{n(t) = m\}$$

are the well-known stationary probabilities and

$$p_{mn}(\tau) = \Pr\{n(t+\tau) = n \mid n(t) = m\}$$

are the transition probabilities. The Laplace transform of (34) can be written

$$\tilde{R}(s) = L[R(\tau)] = \sum_{m=0}^{\infty} \alpha_m P_m \tilde{L}_m(s), \tag{35}$$

where

$$\tilde{L}_m(s) = \sum_{n=0}^{\infty} \alpha_n \tilde{p}_{mn}(s) - \frac{a}{s}, \tag{36}$$

$a = \lambda_c/\mu$ is the offered load in erlangs, and

$$\tilde{p}_{mn}(s) = L\{p_{mn}(t)\}. \tag{37}\dagger$$

Writing

$$\tilde{L}_m(s) = \sum_{n=0}^{N-1} n\tilde{p}_{mn}(s) + N \sum_{n=N}^{\infty} \tilde{p}_{mn}(s) - \frac{a}{s}, \tag{38}$$

and the summations in terms of $\tilde{p}_{mN}(s)$ and $\tilde{p}_{m,N-1}(s)$ using the difference equations satisfied by $\tilde{p}_{m,n}(s)$[21] gives

$$\begin{aligned} s(s+1)\tilde{L}_m(s) &= \tilde{p}_{mN}(s)N(a - N - s) + a(N - a)\tilde{p}_{m,N-1}(s) \\ &\quad + (a + ms)U(N - 1 - m) + (s+1)(NU(m - N) - a) \end{aligned} \tag{39}$$

* In Ref. 27, Descloux investigates the covariance function for a large class of delay and loss systems.

† The quantities $\tilde{p}_{mn}(s)$ have been studied, for example, in Refs. 21 and 22.

where

$$U(k) = \begin{cases} 1 & k \geq 0 \\ 0 & k < 0 \end{cases}.$$

Based upon some algebraic manipulation of the results in Refs. 21 and 22, we obtain

$$\tilde{p}_{mN}(s) = \frac{a^{N-m} D_m(s)}{D_e(s)} \quad m < N$$

$$\tilde{p}_{mN}(s) = \frac{x_2^{N-m}(s) D_N(s)}{D_e(s)} \quad m \geq N$$

$$\tilde{p}_{m,N-1}(s) = \frac{a^{N-1-m} x_2(s) N D_m(s)}{D_e(s)} \quad m < N$$

and

$$\tilde{p}_{m,N-1}(s) = \frac{N D_{N-1}(s) x_2(s)^{N-m}}{D_e(s)} \quad m \geq N, \tag{40}$$

where

$$D_{i+1}(s) = (a + i + s) D_i(s) - a i D_{i-1}(s); \quad i > 0$$

with initial conditions

$$D_0(s) = 1, \quad D_1(s) = a + s, \tag{41}$$

$$D_e(s) = (N + s) D_N(s) - N a D_{N-1}(s) + (a - N x_1(s)) D_N(s), \tag{42}$$

$$x_1(s) = \frac{s + N + a - \sqrt{(s + a + N)^2 - 4Na}}{2N}$$

and

$$x_2(s) = \frac{s + N + a + \sqrt{(s + a + N)^2 - 4Na}}{2N}. \tag{43}$$

For $m < N$, we have from (40)

$$\tilde{p}_{m,N-1}(s) = \frac{N x_2(s)}{a} \tilde{p}_{mN}(s); \quad m < N. \tag{44}$$

Inserting this into (39) gives

$$s(s + 1)\tilde{L}_m(s) = N\tilde{p}_{mN}(s)[(N - a)(x_2(s) - 1) - s] + s(m - a); \quad m < N. \tag{45}$$

Similarly, for $m \geq N$ we can show

$$s(s+1)D_N(s)\tilde{L}_m(s) = N\tilde{p}_{mN}(s)[(a-N-s)D_N(s) + a(N - a)D_{N-1}(s) + \frac{N-a}{N}D_e(s)x_2^{m-N}(s)] + s(N-a)D_N(s); \quad m \geq N. \tag{46}$$

The transform, defined by (35), is evaluated by computing the sum over $m = 0, 1, \cdots N-1$ and evaluating the infinite sum, over $m = N, N+1, \cdots$, in closed form.

We note that taking the limit as $s \to 0$ gives

$$\tilde{R}(0) = V + aC,$$

where V is the variance of the number of busy servers and $C = C(N, a)$ is the Erlang C function. In Appendix B, eq. (54), it is shown that

$$V = a(1-C)$$

thus giving the simple result

$$\tilde{R}(0) = \int_0^\infty R(\tau)\, d\tau = a. \tag{47}*$$

Figures 5a, 5b, and 5c compare $R_e(t)$, given by (18), with $R(t)$ for three different load levels corresponding to "moderate," "low," and "high" probabilities of call originations being delayed. $R(t)$ is obtained by a Laplace transform inversion program written by A. E. Eckberg.† The higher accuracy at low delay probability is not surprising since, for low-delay probabilities, the delay system looks like an infinite server group whose covariance function is a single exponential.[14] We note from the figures that, even at moderate and high delay, probabilities matching the covariance integral do a good job of matching the covariance function.

APPENDIX B

Moments of Calls in Progress for M/M/N Delay System

The distribution function of the number of busy servers is well known. Rather than compute the moments directly from this distribution, we want closed-form expressions for the variance and third

* Recall that the unit of time in (47) is the mean call duration. In Ref. 26, Ott studies the covariance integral of the virtual waiting time in the M/G/1 queue.

† The program employs the Laplace transform inversion method of D. L. Jagerman (Ref. 19).

moments of interest. We obtain these by specializing GI/M/N results to the Poisson case.

From Ref. 23, eq. (III-2), the variance of the number of busy servers is given by

$$V = a[U_0 + U_1 + N(1 - U_0) - a], \tag{48}$$

where

$$a = \frac{\lambda_c}{\mu} = M \tag{49}$$

is the mean number of calls in progress,

$$U_j = \frac{a}{j} U_{j-1} - \binom{N}{j}\left(1 - \frac{a}{N}\right)C \tag{50}*$$

with initial condition

$$U_0 = 1 - C \tag{51}$$

$$C = C(N, a) \tag{52}$$

is the Erlang C function. From (50) we get

$$U_1 = a - NC \tag{53}$$

and from (48) the variance of interest is

$$V = a(1 - C) \tag{54}$$

The third moment about the origin is given by Ref. 23, eq. (V-13):

$$\mu_3 = a[U_0 + 3U_1 + 2U_2] + aN^2(1 - U_0). \tag{55}$$

From (50),

$$U_2 = \frac{a}{2} U_1 - \frac{(N-1)}{2}(N - a)C. \tag{56}$$

Putting (56), (53), and (51) in (55) and using (54) gives

$$\mu_3 = V + a[3a - 2NC + a^2 - aC]. \tag{57}$$

The third central moment is

$$\mu_3^* = V - 2a(N - a)C. \tag{58}$$

* These quantities (i.e., U_j) are defined in Chapter 2 of Ref. 24.

APPENDIX C

Covariance Function Integral: Blocking Case

In this appendix, we are interested in obtaining

$$\int_0^\infty R(\tau)\, d\tau \tag{59}*$$

where $R(\tau)$ is the covariance function of the number of busy servers (calls in progress) in an M/M/N blocking system. More precisely,

$$R(\tau) = E\{(x(t) - \bar{x})(x(t+\tau) - \bar{x})\}, \tag{60}\dagger$$

where $\{x(t)\}$ is the stationary process of the number of busy servers and $\bar{x}$ is the mean number of busy servers. Clearly, (59) is the Laplace transform of $R(\tau)$ evaluated at $s = 0$; i.e.,

$$\tilde{R}(0) = \mathsf{L}[R(\tau)]_{s=0} = \int_0^\infty R(\tau)\, d\tau. \tag{61}$$

Beneš (Ref. 14, p. 214) considers this exact problem and gives expressions for evaluating $\tilde{R}(0)$ involving sums of "sigma functions" which can be recursively computed. Jagerman[19] also gives an expression for $\tilde{R}(s)$ in terms of quantities which are recursively evaluated.

In particular, from Ref. 19 we have

$$\tilde{R}(s) = \frac{V + M^2}{1+s} + \frac{aM}{s(1+s)} - \frac{aNB}{(1+s)^2} - \frac{M^2}{s} + \frac{aNB}{s(1+s)^2}\,\alpha_N(-s-1, a), \tag{62}$$

where M and V are the mean and variance, respectively, of the number of busy servers, $a = \lambda_c/\mu$ is the offered load and $B = B(N, a)$ is the Erlang B function. The function $\alpha_j(x, a)$ satisfies the recursion

$$\alpha_{j+1}(x, a)^{-1} = \frac{x - j - a}{a} - \frac{j}{a}\,\alpha_j(x, a) \tag{63}$$

with initial condition

$$\alpha_1(x, a) = \left(\frac{x}{a} - 1\right)^{-1}.$$

We note that (62) is indeterminate at $s = 0$. Rewriting (62) and taking

---

* The unit of time in this appendix is mean call duration, unless otherwise stated.

† $R(\tau)$ has been studied extensively in telephone trunking theory. It is given by a linear combination, with nonnegative coefficients, of N exponentials (Ref. 14).

the limit gives

$$\tilde{R}(0) = V + M^2 - aNB - aM + \lim_{s\to 0}\left[\frac{aM - M^2}{s} + aNB\left(\frac{1}{s} - \frac{2}{1+s} + \frac{s}{(1+s)^2}\right)\alpha_N(-s-1, a)\right]. \quad (64)$$

Taking the limit of (64)* gives

$$\tilde{R}(0) = V + M^2 - aM - aNB[1 + 2\alpha_N(a) + \alpha'_N(a)] \quad (65)$$

where

$$\alpha_N(a) = \alpha_N(-1, a)$$

satisfies

$$1/\alpha_{j+1}(a) = -(j + 1 + a)/a - (j/a)\alpha_j(a), \quad \text{and} \quad (66a)$$

$\alpha'_j(a)$ satisfies

$$\alpha'_{j+1}(a) = [(j\alpha'_j(a) - 1)/a][\alpha_{j+1}(a)]^2 \quad (66b)\dagger$$

with initial conditions

$$\alpha_1(a) = -a/(1 + a) \quad (66c)$$

and

$$\alpha'_1(a) = -a/(1 + a)^2. \quad (66d)$$

Another form of (65) is

$$\tilde{R}(0) = V - aNB[1 + M/N + 2\alpha_N(a) + \alpha'_N(a)]. \quad (66e)$$

The mean and variance of the number of busy servers are well known to be[14]

$$M = a(1 - B) \quad (67a)$$

and

$$V = M\left[1 - aB\left(\frac{N}{M} - 1\right)\right]. \quad (67b)$$

Equations (66) and (67) and the Erlang B function completely

---

* Using results in Refs. 19 and 25, it can be shown that $aNB\alpha_N(-1, a) = -(aM - M^2)$.

† $\alpha'_j(a) = (d/dx)(\alpha_j(x, a))_{x=-1}$.

specify $\tilde{R}(0)$. For later reference, the third moment about the origin is

$$\mu_3 = (a^2 + 3a + 1)M - N(N + a + 2)(a - M) \qquad (68a)^*$$

and the third central moment

$$\mu_3^* = \mu_3 - 3MV - M^3 \qquad (68b)$$

In addition to (65) being an exact result for the covariance integral, it can also be used in an exponential approximation, $(R_e(t))$, to $R(t)$,

$$R_e(t) = Ve^{-(V/\tilde{R}(0))t}, \qquad (69)\dagger$$

which has the correct variance and matches the covariance integral. It is straightforward‡ to show that $R_e(t)$ crosses $R(t)$ once and bounds $R(t)$ from above prior to the crossover. Figures 6a and 6b compare $R_e(t)$ and $R(t)$§ for high- and low-blocking cases, respectively. The high accuracy at low loads is not surprising since $R(t)$ approaches a single exponential; however, we also see that even at high loads matching the covariance integral does a good job of matching the covariance function.

REFERENCES

1. D. R. Cox and P. A. W. Lewis, *The Statistical Analysis of Series of Events*, London: Methuen, 1966.
2. M. F. Neuts, "A Queue Subject to Extraneous Phase Changes," Adv. Appl. Prob., *3* (1971), pp. 78–119.
3. M. F. Neuts, "A Versatile Markovian Point Process," J. Appl. Prob., to appear.
4. M. F. Neuts, "The M/M/1 Queue with Randomly Varying Arrival and Service Rates," Opsearch, *15*, No. 4 (1978), pp. 139–157.
5. M. F. Neuts, "Further Results on the M/M/1 Queue with Randomly Varying Rates," Opsearch, *15*, No. 4 (1978), pp. 158–168.
6. A. E. Eckberg, "MMP/M/N Queuing Analysis," unpublished work.
7. A. Kuczura, "Queues with Mixed Renewal and Poisson Inputs," B.S.T.J., *51*, No. 6 (July–August 1972), pp. 1305–1326.
8. J. N. Daigle, "Queuing Analysis of a Packet Switching Node with Markov-Renewal Arrival Process," IEEE 1977 Comm. Conf. Proceedings, June 1977, pp. 12.5-279–283.
9. U. Yechiali and P. Naor, "Queuing Problems with Heterogeneous Arrivals and Service," Operations Research, *19* (1971), pp. 722–734.
10. W. W. Chu and L. C. Liang, "Buffer Behavior for Mixed Input Traffic and Single Constant Output Rate," IEEE Trans. Commun., *COM-20* (April 1972), pp. 230–235.
11. F. Closs, "Packet Arrival and Buffer Statistics in a Packet Switching Node," Third Data Commun. Symp., 1973, pp. 12–17.
12. B. Gopinath and J. A. Morrison, "Discrete-Time Single Server Queues with Correlated Inputs," B.S.T.J., *56*, No. 9 (November 1977), pp. 1743–1768.
13. M. L. Chaudhry, "Queuing Problems with Correlated Arrivals and Service Through Parallel Channels," Can. Oper. Res. Jour., *5*, No. 1 (February 1967), pp. 35–46.

---

* Can be obtained simply from the generating function for the distribution of calls in progress.

† Analogous to a Beneš type of approximation (Ref. 14) with somewhat different time constant.

‡ Based on the log convexity of $R(t)$ (Ref. 19).

§ $R(t)$ was obtained from a program written by D. F. Shearer.

14. V. E. Beneš, *Mathematical Theory of Connecting Networks and Telephone Traffic*, New York: Academic Press, 1965.
15. A. Kuczura, "The Interrupted Poisson Process as an Overflow Process," B.S.T.J., *52*, No. 3 (March 1973), pp. 437–448.
16. A. E. Eckberg, "A Generalization of Peakedness to Arbitrary Arrival Processes and Service Time Distributions," unpublished work.
17. M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Vol. I, Third Edition, Hafner, New York, 1969.
18. A. E. Eckberg, "Sharp Bounds on Laplace-Stieltjes Transforms with Applications to Various Queuing Problems," Math. O.R., *2*, No. 2 (May 1977), pp. 135–142.
19. D. L. Jagerman, "An Inversion Technique for the Laplace Transform with Application to Approximation," B.S.T.J., *57*, No. 3 (March 1978), pp. 669–710.
20. J. N. Daigle, "Queuing Analysis of a Packet-Switching Node in a Data Communications System," Doctoral Dissertation, Columbia University, School of Engineering and Applied Science, 1977.
21. H. Zucker, "The Auto-Correlation Function of the Multiserver Queue," unpublished work.
22. T. L. Saaty, "Time-Dependent Solution of the Many-Server Poisson Queue," Oper. Res., *8*, No. 6 (November–December 1960), pp. 755–772.
23. H. Heffes and J. M. Holtzman, "Peakedness in Switching Machines: Its Effect and Estimation," *Proc. of Eighth Int. Teletraffic Cong.*, Melbourne, Australia, 1976, pp. 343/1–343/7.
24. L. Takacs, *Introduction to the Theory of Queues*, New York: Oxford University Press, 1962.
25. D. L. Jagerman, "Some Properties of the Erlang Loss Function," B.S.T.J., *53*, No. 3 (March 1974), pp. 525–551.
26. T. J. Ott, "The Covariance Function of the Virtual Waiting-Time in an M/G/1 Queue," Adv. Appl. Prob., *9* (1977) pp. 158–168.
27. A. Descloux, "Variance of Load Measurements in Markovian Service Systems," B.S.T.J., *54*, No. 7 (September 1975), pp 1277–1300.