A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance

HARRY HEFFES, SENIOR MEMBER, IEEE, AND DAVID M. LUCANTONI

Abstract—**We study the performance of a statistical multiplexer whose inputs consist of a superposition of packetized voice sources and data. The performance analysis predicts voice packet delay distributions, which usually have a stringent requirement, as well as data packet delay distributions. The superposition is approximated by a correlated Markov modulated Poisson process (MMPP), which is chosen such that several of its statistical characteristics identically match those of the superposition. Matrix analytic methods are then used to evaluate system performance measures. In particular, we obtain moments of voice and data delay distributions and queue length distributions. We also obtain Laplace–Stieltjes transforms of the voice and data packet delay distributions, which are numerically inverted to evaluate tails of delay distributions. It is shown how the matrix analytic methodology can incorporate practical system considerations such as finite buffers and a class of overload control mechanisms discussed in the literature. Comparisons with simulation show the methods to be accurate. The numerical results for the tails of the voice packet delay distribution show the *dramatic* effect of traffic variability and correlations on performance.**

I. INTRODUCTION

IN this paper we study the performance of a statistical multiplexer whose inputs consist of a superposition of packetized voice sources together with data traffic. The performance analysis predicts voice packet delay distributions, which usually have a stringent requirement, as well as data packet delay distributions.

While the problem was motivated by the desire to analytically gain insight into the performance of an integrated voice/data netork, the techniques developed are more broadly applicable. The methodology presented here provides new insights and can also be viewed as a step in the direction of reducing the dependence on simulations, which can be expensive and are typically used in studying performance issues for this class of problems. In addition, the analysis enables one to obtain low probability tails and high percentiles of distributions, which cannot be obtained with simulations.

The input process to the statistical multiplexer is a fairly complex process and can possess correlations, in the number of arrivals in adjacent time intervals, which can significantly affect queueing performance. These correlations result from the fact that the aggregate voice packet arrival rate is a *modulated* process obtained by ing the individual voice source packet rate by the numoer of voice sources in their talk spurt, which is itself a correlated process. Even if a component voice process is approximated as a renewal process, with deterministically spaced packets during a talk spurt followed by an exponentially distributed silence period, the superposition process is a complex nonrenewal process. Exact analysis of systems to which this superposition process is offered is intractable, especially when such systems contain finite buffers and overload control mechanisms.

The approach we take in this paper is to approximate the aggregate arrival process by a simpler, correlated, nonrenewal stream, which is modulated in a Markovian manner. The approximating stream is chosen such that several of its statistical characteristics identically match those of the original superposition. In choosing the approximating stream we are driven by the need for analytic simplicity as well as the desire for versatility. A natural choice is the Markov modulated Poisson process (MMPP), which is a doubly stochastic Poisson process where the rate process is determined by the state of a continuous-time Markov chain. This process has previously been used to accurately approximate a superposition of packet arrival processes for a related problem [1]. One advantage of our characterization of the superposition of voice sources and data as an MMPP is that once we obtain the parameters of the process we can feed it into any system we like.

In this paper the packet voice/data multiplexer is modeled by feeding the MMPP into a single-server queue, served first-in–first-out (FIFO), with general service time distribution where the service distribution is the appropriate mixture of the voice and data packet service time distributions. A detailed analysis of this queueing system is presented where matrix analytic-algorithmic procedures [2]–[5] are used to compute, for example,

1) transforms of the delay distributions, which are numerically inverted to yield the *tail of the delay distributions*;
2) the queue length distributions; and
3) moments of the delay distributions.

These measures are obtained for both voice and data traffic. The first quantity is of importance, for example,

Manuscript received May 11, 1985; revised April 2, 1986.
The authors are with AT&T Bell Laboratories, Holmdel, NJ 07733.
IEEE Log Number 8609928.

in systems that have a *performance criterion on the tail of the voice packet delay distribution.*

We also show how the powerful matrix-analytic methodology can be applied to handle finite buffers and a particular type of overload control discussed in the literature. The control mechanism here is to use a variable bit rate on voice packets during congestion (see, e.g., [6]–[8]). This would provide a graceful degradation of system performance during overload. The variable bit rate overload control is incorporated into the model by using state-dependent service times in the matrix-analytic methodology.

There has been a considerable amount of related work on this problem, and we refer the reader to [9]–[17] for details and further references. In a companion paper in this issue [16], and in [17], an approach based on approximating the superposition process by a renewal process with an inflated coefficient of variation for the interarrival time distribution is presented. The inflation factor for the arrival process depends on the system to which the process is offered, and simple closed-form formulas for the first two moments of delay, which capture the qualitative behavior, are given. The multiplexer is modeled as a FIFO queue with infinite buffers and state-independent service times. In [17] it is observed that although successive interarrival times can be nearly independent and exponentially distributed, these low correlations can have a cumulative effect over long time periods and can result in behavior significantly different than that of a Poisson process. In [13] a similar approach is used to analyze a multiplexer serving only packetized voice. For the purposes of comparison, the numerical examples presented here correspond to those in [17].

In [15], approximate queue length distributions are obtained for the case where all sources are identical and have the same deterministic service time. This precludes mixing voice and data sources with different packet lengths.

We finally note that the methodology presented is fairly general, and the application to the voice/data problem may be viewed as an illustration.

II. Statistical Properties of Packetized Voice Processes

In this section we use renewal theory results to evaluate the mean, variance–mean ratio, and third central moment of the number of arrivals in a time interval for a superposition of packetized voice sources.

The packet stream from a single voice source is modeled by arrivals at fixed intervals of T ms during talk spurts and no arrivals during silences. In particular, we consider the packet arrival process from a *single* voice source to be a renewal process with interarrival time distribution given by

$$F(t) = [(1 - \alpha T) + \alpha T(1 - e^{-\beta(t-T)})]\, U(t - T) \quad (1)$$

(where $U(t)$ is the unit step function) and the Laplace–Stieltjes transform (LST)

$$\tilde{f}(s) = \int_0^\infty e^{-st}\, dF(t) = [1 - \alpha T + \alpha T\beta/(s + \beta)]\, e^{-sT}, \quad (2)$$

with the mean packet arrival rate from a single source clearly given by

$$\lambda = -1/\tilde{f}'(0) = 1/(T + \alpha T/\beta). \quad (3)$$

This corresponds to a geometrically distributed number of voice packets (with mean $1/\alpha T$) during an approximately exponentially distributed talk spurt with mean α^{-1} followed by an approximately exponentially distributed silent period with mean β^{-1} [18]–[21]. We note that the actual talk spurt durations here are discrete variables. The parameters used in this paper are given by $\alpha^{-1} \approx 352$ ms, $\beta^{-1} \approx 650$ ms [22], and $T = 16$ ms, which corresponds to the same *single-source* model used in [17]. It should be pointed out that the methodology we present is not restricted to the above voice source characterization. For example, we note that *it is not necessary that the distribution of the silent period be exponential.* This would allow, for example, the use of a silent period distribution which is a mixture of the distributions of silence due to pauses and silence due to listening.

Since each packetized voice process is a renewal process, we can use renewal theory results to study the moments of the number of arrivals in an interval. As we will see in the next section, these quantities will be useful in developing an approximation to the superposition process. Let $N(0, t)$ denote the number of arrivals of a stationary renewal process in the interval $(0, t)$, let

$$M_r(t) = E[N^r(0, t)],$$

and let

$$\tilde{M}_r(s) = \boldsymbol{L}[M_r(t)]$$

where $\boldsymbol{L}(\cdot)$ denotes the Laplace transform. Then it is known (e.g., [23, p. 229]) that

$$\tilde{M}_1(s) = \lambda/s^2 \quad (4a)$$

$$\tilde{M}_2(s) = \frac{\lambda}{s^2}\left(\frac{1 + \tilde{f}(s)}{1 - \tilde{f}(s)}\right) \quad (4b)$$

and

$$\tilde{M}_3(s) = \frac{\lambda}{s^2}\left(\frac{1 + 4\tilde{f}(s) + \tilde{f}^2(s)}{(1 - \tilde{f}(s))^2}\right) \quad (4c)$$

where $\tilde{f}(s)$ is the LST of the interarrival time distribution and λ^{-1} is the mean time between arrivals. Clearly, $M_1(t) = \lambda t$. It is also known [24] that the index of dispersion for counts, $I(t)$, satisfies

$$\lim_{t\to\infty} I(t) = \lim_{t\to\infty} \frac{\text{var}\,(N(0, t))}{M_1(t)} = \frac{\text{var}\,(X)}{E^2(X)},$$

the squared coefficient of variation of the interarrival time X. Applying these results to $F(t)$ and $\tilde{f}(s)$, given by (1) and (2), we have

$$M_1(t) = t/[T + \alpha T/\beta] \tag{5a}$$

and

$$\lim_{t\to\infty} \frac{\text{var } (N(0, t))}{M_1(t)} = \frac{1 - (1 - \alpha T)^2}{(\alpha T + \beta T)^2} \tag{5b}$$

The second and third moments of the number of arrivals in a *finite time* interval, which are needed in Section III, are obtained by numerical transform inversion [25], [30] of $\tilde{M}_2(s)$ and $\tilde{M}_3(s)$ at the desired times, with $\tilde{f}(s)$ defined by (2).

We now consider the superposition of n identical independent voice packet processes and denote by $N_i(0, t)$ the number of packet arrivals in $(0, t)$ from the ith stream. For the results in this paper we assume a fixed number n of active voice sources. The number of arrivals for the superposition is given by

$$N^S(0, t) = \sum_{i=1}^{n} N_i(0, t).$$

Clearly,

$$M_1^S(t) = E[N^S(0, t)] = nM_1(t), \tag{6a}$$

and the index of dispersion for counts satisfies

$$\frac{\text{var } [N^S(0, t)]}{E[N^S(0, t)]} = \frac{\text{var } [N(0, t)]}{E[N(0, t)]} \tag{6b}$$

where we have used the superscript S to denote the *superposition* variables. The third central moment for the superposition process,

$$\mu_3^{*S}(0, t) = E\{[N^S(0, t) - E(N^S(0, t))]^3\},$$

clearly reduces to

$$\mu_3^{*S}(0, t) = n[M_3(t) - 3M_2(t)\, M_1(t) + 2M_1^3(t)] \tag{6c}$$

where $M_2(t)$ and $M_3(t)$ are obtained from Laplace transforms inversion of (4b) and (4c) at the desired times, and $M_1(t)$ is obtained from (5a). From (6b) the variance–mean ratio for the superposition is identical to that of the individual processes. These results clearly enable us to compute the mean, variance–mean ratio, and third cental moment for the superposition over a finite time interval and the variance–mean ratio over an infinite time interval.

We note that the variance–time curve [24] $V^S(t) = \text{var } [N^S(0, t)]$ completely defines the correlation structure of the superposition process, i.e., if

$$C(t) = \text{cov } [N^S(0, t), N^S(t, 2t)],$$

then

$$C(t) = \frac{V^S(2t)}{2} - V^S(t). \tag{7}$$

Thus, to accurately approximate the correlation properties of the superposition it is important that the approximate process provide a good match to the variance–time curve. This is considered in the next section.

III. Approximating the Superposition of Packetized Voice and Data Streams

In this section we present a technique for approximating the superposition of a collection of voice sources and data traffic. Since the superposition process is a correlated nonrenewal stream, we choose the approximating process as a correlated nonrenewal process such that several of its statistical characteristics identically match those of the superposition.

In choosing the approximating process we are driven by the need for analytic simplicity as well as the desire for versatility. A natural choice is the MMPP. An MMPP is a doubly stochastic Poisson process where the rate process is determined by the state of a continuous-time Markov chain. We use a two-state Markov chain where the mean sojourn times in states 1 and 2 are r_1^{-1} and r_2^{-1}, respectively. When the chain is in state j ($j = 1, 2$) the arrival process is Poisson with rate λ_j. This process has been used to accurately approximate a superposition of packet arrival processes and subsequent queueing delays for a related problem [1].

We choose the four parameters of the MMPP so that the following characteristics of the superposition are matched:

1) the mean arrival rate;

2) the variance-to-mean ratio of the number of arrivals in $(0, t_1)$;

3) the long term variance-to-mean ratio of the number of arrivals; and

4) the third moment of the number of arrivals in $(0, t_2)$.

In the previous section we discussed the evaluation of these quantities for the superposition of packet voice processes. We thus have available to us $M_1^S(t) = \lambda t$, var $[N^S(0, t_1)]/E[N^S(0, t_1)]$, $\lim_{t\to\infty}$ (var $[N^S(0, t)]/M_1^S(t)$), and $\mu_3^{*S}(0, t_2)$. Exact expressions for these quantities for the two-state MMPP are obtained in Appendix A using the probability generating function of the number of arrivals in an interval [3],

$$g(z, t) = \pi \exp\{[R + (z - 1)\Lambda]t\}e,$$

where, for the two-state MMPP, the equilibrium probability vector π, is given by

$$\pi = \frac{1}{r_1 + r_2}(r_2, r_1), \qquad e = (1, 1)^T,$$

$$R = \begin{bmatrix} -r_1 & r_1 \\ r_2 & -r_2 \end{bmatrix}, \quad \text{and} \quad \Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}.$$

If we dnote N_t as the number of arrivals of the stationary two-state MMPP over the interval $(0, t)$, then it is shown in Appendix A that

$$\bar{N}_t = E[N_t] = \frac{\lambda_1 r_2 + \lambda_2 r_1}{r_1 + r_2}\, t \tag{8}$$

and

$$\frac{\text{var}\,(N_t)}{\bar{N}_t} = 1 + \frac{2(\lambda_1 - \lambda_2)^2 r_1 r_2}{(r_1 + r_2)^2 (\lambda_1 r_2 + \lambda_2 r_1)} - \frac{2(\lambda_1 - \lambda_2)^2 r_1 r_2}{(r_1 + r_2)^3 (\lambda_1 r_2 + \lambda_2 r_1)t} \cdot (1 - e^{-(r_1 + r_2)t}). \quad (9)$$

Also,

$$E[(N_t - \bar{N}_t)^3] = g^{(3)}(1, t) - 3\bar{N}_t(\bar{N}_t - 1)\frac{\text{var}\,(N_t)}{\bar{N}_t} - \bar{N}_t(\bar{N}_t - 1)(\bar{N}_t - 2) \quad (10)$$

where

$$g^{(3)}(1, t) = \frac{6}{r_1 + r_2}\left[\frac{A_{11}}{6}t^3 + \frac{A_{21}}{2}t^2 + A_{31}t + A_{12}te^{-(r_1 + r_2)t} + A_{41}(1 - e^{-(r_1 + r_2)t})\right]. \quad (11)$$

Expressions for A_{ij} are given in Appendix A (A4) in terms of the four parameters λ_1, λ_2, r_1, r_2. Note that from (9) we clearly have

$$\lim_{t \to \infty} \frac{\text{var}\,(N_t)}{\bar{N}_t} = 1 + \frac{2(\lambda_1 - \lambda_2)^2 r_1 r_2}{(r_1 + r_2)^2 (\lambda_1 r_2 + \lambda_2 r_1)}. \quad (12)$$

If we denote

$$a = \frac{E[N^S(0, t)]}{t}, \quad (13a)$$

$$b_t = \frac{\text{var}\,[N^S(0, t)]}{E[N^S(0, t)]}, \quad (13b)$$

and

$$b_\infty = \lim_{t \to \infty} \frac{\text{var}\,[N^S(0, t)]}{E[N^S(0, t)]}, \quad (13c)$$

then we must solve the following equations:

$$\frac{\lambda_1 r_2 + \lambda_2 r_1}{r_1 + r_2} = a, \quad (14a)$$

$$\frac{2(\lambda_1 - \lambda_2)^2 r_1 r_2}{(r_1 + r_2)^2 (\lambda_1 r_2 + \lambda_2 r_1)} = b_\infty - 1, \quad (14b)$$

$$\frac{1 - e^{(r_1 + r_2)t_1}}{(r_1 + r_2)t_1} = \frac{b_\infty - b_{t_1}}{b_\infty - 1}, \quad (14c)$$

and

$$g^{(3)}(1, t_2) = \mu^{*S}(0, t_2) + 3at_2(at_2 - 1)\, b_{t_2} + at_2(at_2 - 1)(at_2 - 2) \quad (14d)$$

for the parameters λ_1, λ_2, r_1, and r_2. Note that the right-hand sides of (14) are all computable from the results for the superposition of packetized voice sources in Section II.

We can solve for $(r_1 + r_2)$ directly from (14c) if $b_{t_1} > 1$, i.e., the variance-to-mean ratio of the superposition process at t_1 is greater than that of a Poisson process. Denote this solution by $d = (r_1 + r_2)$, which can be obtained, for example, by successive substitution in

$$d = \frac{1}{t_1}\left(\frac{b_\infty - 1}{b_\infty - b_{t_1}}\right)(1 - e^{-dt_1}). \quad (15a)$$

Furthermore, $g^{(3)}(1, t)$ can be written in terms of the parameters of the two-state MMPP, using the results of Appendix A to yield

$$g^{(3)}(1, t_2) = a^3t_2^3 + 3a^2(b_\infty - 1)t_2^2 + \frac{3a(b_\infty - 1)}{d} \cdot \left[\frac{(\lambda_1 - \lambda_2)(r_1 - r_2)}{d} - a\right]t_2 + \frac{3a}{d^2}(b_\infty - 1)[(\lambda_1 - \lambda_2)(r_1 - r_2) + ad] \cdot t_2e^{-dt_2} - \frac{6a}{d^3}(b_\infty - 1) \cdot (\lambda_1 - \lambda_2)(r_1 - r_2)(1 - e^{-dt_2}).$$

We thus see that (14d) can be written in the form

$$(\lambda_1 - \lambda_2)(r_1 - r_2) = K \quad (15b)$$

where K is known. The remaining two equations can be written as

$$(\lambda_1 r_2 + \lambda_2 r_1) = ad \quad (15c)$$

and

$$(\lambda_1 - \lambda_2)^2 r_1 r_2 = \frac{(b_\infty - 1)ad^3}{2} \quad (15d)$$

where we again require that $b_\infty > 1$, i.e., larger long-term variability of the superposition than a Poisson process.

If $K = 0$, then it is necessary that $r_1 = r_2 = (d/2)$ since in (15d) $b_\infty > 1$ and hence $\lambda_1 \neq \lambda_2$. We then solve (15c) and (15d) for λ_1 and λ_2.

If $K \neq 0$, we define

$$e = \frac{(b_\infty - 1)ad^3}{2K^2}.$$

Then the solution to (15) is finally given by

$$r_1 = \frac{d}{2}\left(1 + \frac{1}{\sqrt{4e + 1}}\right)$$

$$r_2 = d - r_1$$

$$\lambda_2 = \left(\frac{ad}{r_2} - \frac{K}{r_1 - r_2}\right)\left(\frac{r_2}{r_1 + r_2}\right)$$

$$\lambda_1 = \frac{K}{r_1 - r_2} + \lambda_2.$$

We note that we still have freedom in choosing the time points t_1 and t_2. Since the variance-time curve $V(t)$ completely specifies the correlation structure of the process

(see Section II), we will choose t_1 and t_2 to get a good fit to $V(t)$, or equivalently $V(t)/E(N(t))$, over the entire range of t.

The data streams are incorporated into the model by noting that the superposition of a Poisson process of rate λ_d and a two-state MMPP with parameters λ_1, λ_2, r_1, r_2 is again a two-state MMPP with parameters $\lambda_1 + \lambda_d$, $\lambda_2 + \lambda_d$, r_1, r_2. Thus, if the superposition of the data streams can be approximated by a Poisson process, then a trivial modification of the MMPP representing the packetized voice traffic will model the aggregate voice and data streams. If the data traffic is not Poisson, then the original methodology of Section II is applied directly to the aggregate stream.

IV. Queueing Performance—The MMPP/G/1 Queue

In this section we discuss the performance of a statistical multiplexer with inputs consisting of the superposition of voice streams together with data streams. The streams are multiplexed onto a high-speed transmission line. This is modeled as a single-server queue where the service time of a packet is its transmission time on the line. In view of the characterization of the superposition process as an MMPP as given in the previous section, we model the multiplexer as a MMPP/G/1 queue. In particular, if there is only voice on the system, then since all voice packets are of equal length we have an MMPP/D/1 queue.

As discussed in Section III, the arrival of data packets can be assumed to be a Poisson process and is trivially incorporated into the representation of the MMPP. We approximate the performance of the voice/data system by using an MMPP/G/1 queue where the service time distribution is an appropriate mixture of the service times of voice and data packets. This is not an exact model of the voice/data system since we are not keeping track of the order of voice and data packets in the queue, but as we will see in the next section, for the range of parameters of interest, the results agree very well with a simulation of the actual system.

Although the MMPP used in this paper has only two states, the results presented below also apply to the general case. In particular, consider an m-state continuous-time Markov chain with infinitesimal generator R. When the Markov chain is in state j there are Poisson arrivals with rate λ_j. These arrivals join a FIFO single-server queue, and the service times of all customers are independent and identically distributed with distribution function $\tilde{H}(\cdot)$. Let the ith moment about the origin of the service times be denoted by $\mu^{(i)}$ and let Λ be a diagonal matrix with the elements λ_j along the diagonal.

Our main performance measures will be the distribution function and moments of the delay seen by voice and data packets. Let the vector of distribution functions $\tilde{W}(x)$ have components $\tilde{W}_j(x)$ where $\tilde{W}_j(x)$ is the joint probability that at an arbitrary time the MMPP is in phase j and that a *virtual* customer who arrived at that time would wait less than or equal to x before entering service. It is easy to show that the virtual waiting time distribution is given by $\tilde{W}(x)e$. The following algorithm describes the computation of $\tilde{W}(x)e$, from which the distributions of delays seen by an arbitrary packet arrival, voice packet arrival, and data packet arrival can be obtained. This is discussed after the presentation of the algorithm. Let π be the stationary distribution of the Markov chain with generator R and let λ be the vector with jth component λ_j. The derivation of the virtual delay distribution is outlined in Appendix B, and the algorithm for computing the transform of the virtual delay distribution and the first two moments of the virtual waiting time is summarized as follows:

$$E(W) = \frac{1}{2(1-\rho)}[2\rho + \mu^{(2)}\pi\lambda - 2\mu^{(1)} \cdot (y_0 + \mu^{(1)}\pi\Lambda)(R + e\pi)^{-1}\lambda], \tag{16}$$

and

$$E(W^2) = \frac{1}{3(1-\rho)}[3\mu^{(1)}[2\mu^{(1)}W'(0)\Lambda - 2W'(0) - \mu^{(2)}\pi\Lambda](R + e\pi)^{-1}\lambda - 3\mu^{(2)}W'(0)\lambda + \mu^{(3)}\pi\lambda] \tag{17}$$

where

$$W'(0) = \mu^{(1)}\pi\Lambda(R + e\pi)^{-1} + y_0(R + e\pi)^{-1} - E(W)\,\pi - \pi,$$

the traffic intensity ρ is given by $\rho = \pi\lambda\mu^{(1)}$, and the vector y_0 is computed by the following algorithm:

1) Compute the stochastic matrix G by the iterative procedure given in (B8). The (i, j) component of G is the probability that a busy period starting with the MMPP in phase i ends in phase j.
2) Compute the stationary distribution of the Markov chain with transition matrix G from
$$g = (g_1, g_2) = \frac{1}{G_{12} + G_{21}}(G_{21}, G_{12}).$$
3) Compute $A = \int_0^\infty e^{Rt}\,d\tilde{H}(t)$. A_{ij} is the probability that a service time ends with the MMPP in phase j given that the service began in phase i. Using the fact that in the two-state case
$$e^{Rt} = e\pi - \frac{e^{-(r_1+r_2)t}}{r_1 + r_2}R,$$
we have that A is given by
$$A = e\pi - \frac{H(r_1 + r_2)}{r_1 + r_2}R$$
where $H(s)$ is the LST of $\tilde{H}(x)$.
4) Compute $U = (\Lambda - R)^{-1}\Lambda$ where U keeps track of the phase during an idle period. That is, U_{ij} is the probability that the first arrival to a busy peroid ar-

rives with the MMPP in phase j given that the last departure from the previous busy period departed with the MMPP in phase i.

5) Compute $\boldsymbol{\beta} = \mu^{(1)}(\pi\lambda)\boldsymbol{e} + (R + \boldsymbol{e}\pi)^{-1}(A - I)\lambda$ where β_j is the expected number of arrivals during a service that began in phase j.
6) Compute $\boldsymbol{\mu} = (I - G + \boldsymbol{eg})[I - A + \boldsymbol{eg} - \boldsymbol{\beta g}]^{-1}\boldsymbol{e}$ where μ_j is the expected number of departures during a busy period that began in phase j.
7) Compute $\boldsymbol{d}$ such that $\boldsymbol{dUG} = \boldsymbol{d}$, $\boldsymbol{de} = 1$. It is clear that d_j is the stationary probability of ending a busy period in phase j.
8) Compute $\boldsymbol{x_0} = (\boldsymbol{dU\mu})^{-1}\boldsymbol{d}$. $(\boldsymbol{x_0})_j$ is the stationary probability that a departure leaves the system empty with the MMPP in phase j. This is just the stationary probability of being in phase j at successive epochs which leave the system empty divided by the expected number of departures between such epochs.
9) Compute $\boldsymbol{y_0} = (\pi\lambda)\,\boldsymbol{x_0}(\Lambda - R)^{-1}$ where $(\boldsymbol{y_0})_j$ is the stationary probability of the system being empty and the phase of the MMPP being in phase j at an arbitrary point in time.
10) Finally, the LST of the virtual delay distribution is given by $\boldsymbol{W}(s)\boldsymbol{e}$ where

$$\boldsymbol{W}(s) = \begin{cases} s\boldsymbol{y_0}[sI + R - \Lambda + \Lambda H(s)]^{-1}, & s > 0, \\ \boldsymbol{\pi}, & s = 0, \end{cases}$$

and $H(s)$ is the LST of $\tilde{H}(x)$.

Remarks:

1) The above algorithm also holds for the m-state case if in steps 2 and 3 the explicit two-state formulas are replaced by computation of the appropriate stationary distribution $\boldsymbol{g}$ and matrix A.

2) Note that if the MMPP has only one state or if $\lambda_j = \lambda$ for all j, then the above expressions reduce to

$$E(W) = \frac{\lambda\mu^{(2)}}{2(1-\rho)},$$

$$\text{var}\,(W) = 2E(W)^2 + \frac{\lambda\mu^{(3)}}{3(1-\rho)},$$

$$W(s) = \frac{s(1-\rho)}{s - \lambda + \lambda H(s)},$$

which are the results for the $M/G/1$ queue [26].

To compute the waiting time distribution seen by an arbitrary arrival, $\tilde{W}_a(x)$, it is easy to show that

$$\tilde{W}_a(x) = (\pi\lambda)^{-1}\,\tilde{\boldsymbol{W}}(x)\lambda$$

where π and λ have been defined earlier. Since the data packets arrive according to a Poisson process, the delay they experience is equivalent to the virtual waiting time (i.e., Poisson arrivals see time averages [28]). Therefore, the delay distribution seen by data packets is given by $\tilde{\boldsymbol{W}}_d = \tilde{\boldsymbol{W}}\boldsymbol{e}$. By conditioning on the type of arrival we can express the distribution of the delay of an arbitrary arrival in terms of the distribution of voice $W_v(x)$ and data delays as follows:

$$\tilde{W}_a(x) = \frac{\lambda_v}{\lambda_v + \lambda_d}\,\tilde{W}_v(x) + \frac{\lambda_d}{\lambda_v + \lambda_d}\,\tilde{W}_d(x),$$

which can be solved for $\tilde{W}_v(x)$. Note that $\lambda_v = a$ and λ_d have been defined in the previous section. Therefore, we can compute the delay seen by voice packets in terms of the parameters of the model and the quantities $\tilde{W}_j(x)$, $j = 1, \cdots, m$.

One advantage of our characterization of the superposition of voice and data as an MMPP is that once we obtain the parameters of the process we can feed it into any system we like. In particular, it enables us to model two very important practical situations: finite buffers and overload control.

One type of overload control which has been discussed in the literature is to use a variable bit rate on voice packets during congestion [6]–[8]. This would provide a graceful degradation of system performance during overload. The variable bit rate can be incorporated into the model by making the service times state dependent. That is, when the buffer lengths exceed some thresholds, the service time of voice packets decreases (i.e., bits are dropped). The effect of finite buffers on the model would be the truncation of the state space of the embedded Markov renewal process (see Appendix B). We illustrate these modifications by displaying the transition probability matrix for a small example (because of space limitations). For example, suppose the system has capacity for four packets. When the number of packets in the buffer is less than or equal to 2 the service time distribution is $\tilde{H}(\cdot)$, and when the number of packets is greater than 2 the service time distribution is $\tilde{H}_1(\cdot)$. Then the transition matrix of the embedded Markov renewal process (see (B1) in Appendix B) is given by

$$\tilde{Q}(x) = \begin{bmatrix} \tilde{B}_0(x) & \tilde{B}_1(x) & \tilde{B}_2(x) & \tilde{B}_3(x) & \sum_{n=4}^{\infty} \tilde{B}_n(x) \\ \tilde{A}_0(x) & \tilde{A}_1(x) & \tilde{A}_2(x) & \tilde{A}_3(x) & \sum_{n=4}^{\infty} \tilde{A}_n(x) \\ 0 & \tilde{A}_0(x) & \tilde{A}_1(x) & \tilde{A}_2(x) & \sum_{n=3}^{\infty} \tilde{A}_n(x) \\ 0 & 0 & \tilde{C}_0(x) & \tilde{C}_1(x) & \sum_{n=2}^{\infty} \tilde{C}_n(x) \\ 0 & 0 & 0 & \tilde{C}_0(x) & \sum_{n=1}^{\infty} \tilde{C}_n(x) \end{bmatrix}$$

where $A_n(\cdot)$ and $B_n(\cdot)$ are as defined in Appendix B, $C_n(x) = \int_0^x P(n, t)\,d\tilde{H}_1(t)$, and $P(n, t)$ is defined in Appendix B. Using Markov renewal theory similar to that used in Appendix B, the performance measures of this system may be derived. Since the MMPP has only two phases, the above matrix is of order $2(N + 1)$ where N is the system size, so that analyzing a system with a capacity of 100 packets is completely feasible.

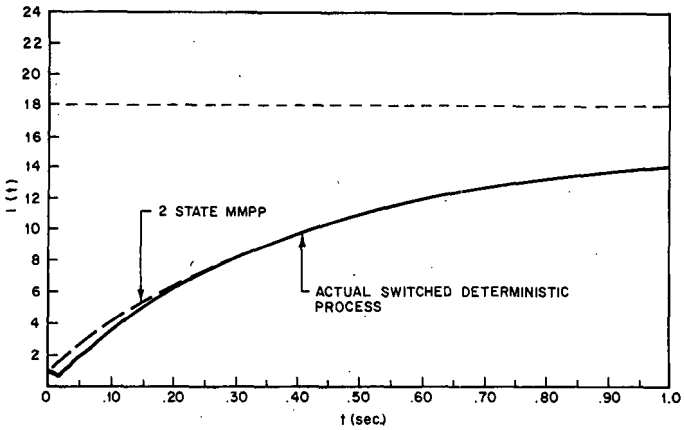


Fig. 1. Index of dispersion for number of arrivals in $(0, t)$.

V. Numerical Results and Discussion

As indicated in Section II, the parameters used for conversational speech will be $\alpha^{-1} = 352$ ms for the mean talk-spurt duration and $\beta^{-1} = 650$ ms for the mean silent-period duration. The time interval between packet arrivals from a voice source in a talk spurt is assumed to be $T = 16$ ms. Fig. 1 shows the variance-mean ratio curve $I(t)$ for the superposition of packetized voice processes. Also shown on the figure is the corresponding curve for the MMPP approximation, where we have chosen $t_1 = 500$ ms to get a good fit over the entire range of t in order to accurately approximate the correlation properties of the input process over the entire range of t. We note that $I(t)$ for the MMPP is insensitive to t_2. Initially, we choose $t_2 = t_1 = 500$ ms.

For the purpose of comparison to the results of [17] we will use the same line speed (1.536 Mbits/s) and voice packet length (64 bytes), which result in a fixed voice packet service time of 1/3 ms. We also use the same data packet length distribution, which is geometrically distributed with mean 50 bytes. As in [17], data packets are assumed to arrive as a Poisson process (although our methodology is not restricted to this).

Fig. 2 shows the mean delay as a function of the number of active voice lines for the case of voice traffic only. Results are shown for a simulation of the actual system and for our MMPP model, which is seen to be very accurate over the entire range. Also shown on the figure are results from [17], which have comparable accuracy above 120 lines, corresponding to a utilization of 0.88 and somewhat poorer accuracy for $\rho < 0.88$. The correlation and traffic variability effects of the input process are seen by comparing the results to the Poisson curve shown. It is seen that these effects become significant above 100 voice lines, which corresponds to a line utilization of $\rho = 0.73$. Fig. 3 shows the standard deviation of delay for the same problem, and Fig. 4 shows the mean delay of an arbitrary packet arrival as a function of utilization for the case of both voice and data traffic. Here we fix the number of voice lines at 80 and vary the intensity of the data traffic. The relative behavior of the results of Figs. 3

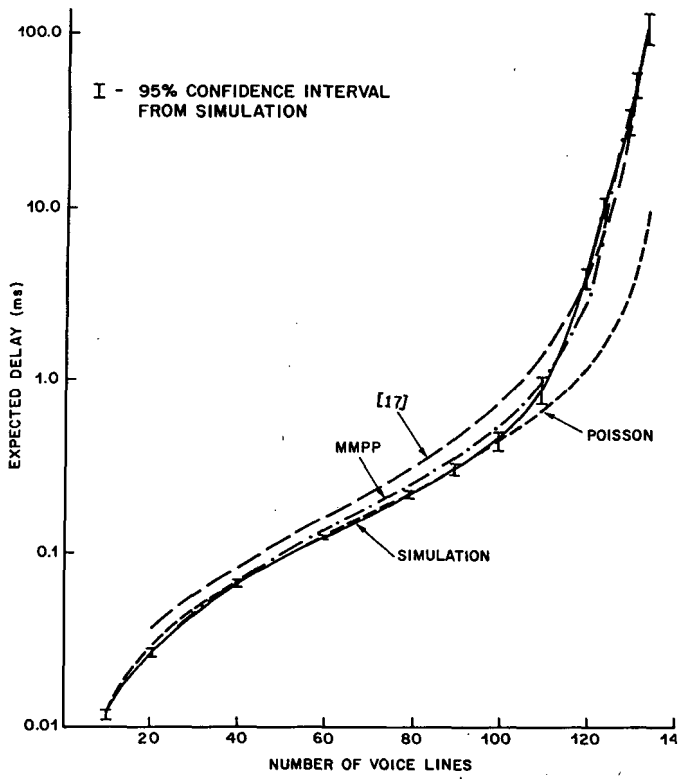


Fig. 2. Expected delay for a packetized voice multiplexer.

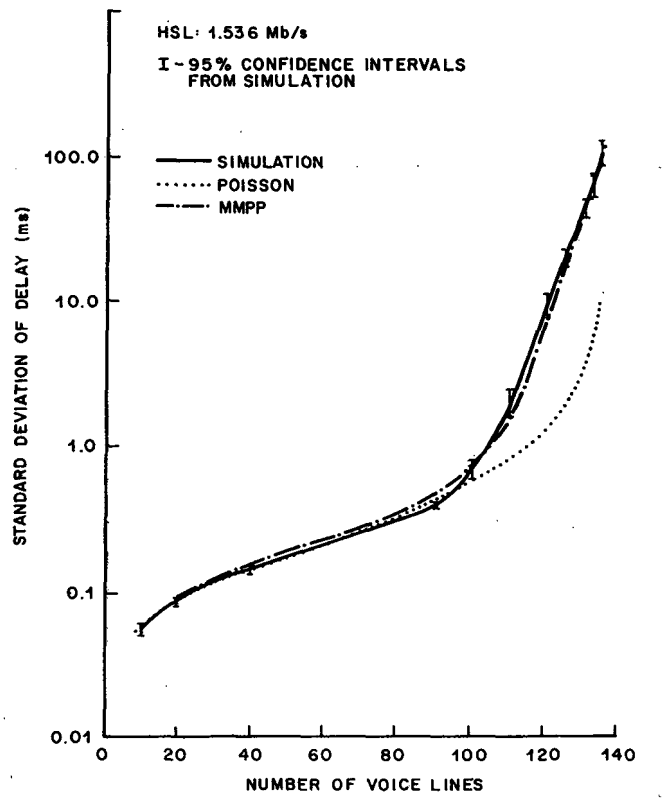


Fig. 3. Standard deviation of delay for packetized voice multiplexer.

and 4 with respect to simulation results in [17], which are for delays seen by an arbitrary arrival, and Poisson input, is similar to that of Fig. 2. We particularly note that the validations with simulation show our model results to be

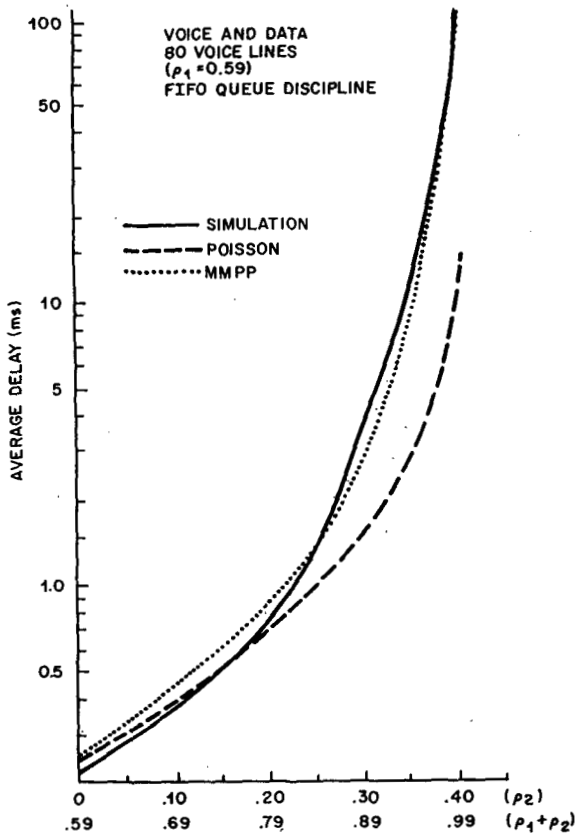


Fig. 4. Average delay versus data utilization.

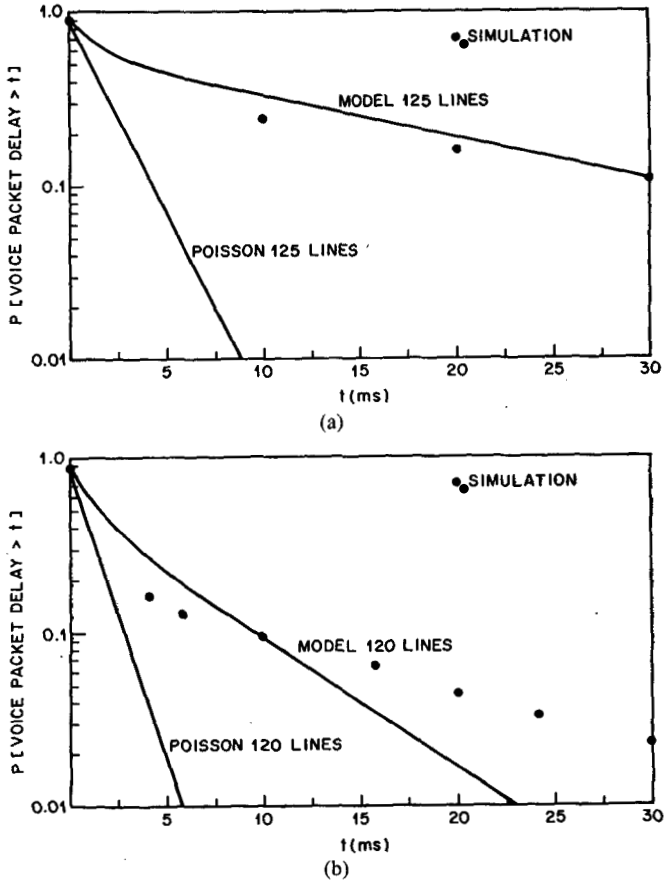


Fig. 5. (a) Complementary voice packet delay distribution (125 lines). (b) Complementary voice packet delay distribution (120 lines).

very accurate. As discussed in [17], for the above parameters the actual delays seen by voice and data packets are close to those seen by arbitrary arrivals.

The deviations of the results from the Poisson input results are further highlighted by comparing the complementary voice packet delay distributions. In Fig. 5(a) and (b) we plot the complementary voice packet delay distributions for the example of Figs. 2 and 3, for the cases of 120 and 125 voice lines. For example, with 125 lines we observe that P[voice packet delay >9 ms], which corresponds to 27 service times, is *more than an order of magnitude* larger than the corresponding Poisson input case, thus showing a *dramatic effect of traffic variability and correlations* on performance. This is also the case for the simulation results shown. For 120 lines the P[voice packet delay >6 ms], which corresponds to 18 service times, for both model and simulation results is more than an order of magnitude larger than the corresponding Poisson input case. We note that accurately capturing the mean delay requires the exact and model results for the complementary delay distributions to cross as shown. We expect that to more accurately capture the long tail behavior would require an MMPP with more levels. The multilevel MMPP/G/1 model described would be useful in this case.

We now turn our attention to the sensitivity of the above results to the parameter t_2. Recall that the actual skewness parameter $\mu_3^{*S}(0, t)$ is matched at $t = t_2$. Results shown in Fig. 6(a) and (b) correspond to $t_2 = 1$ s, which approximately maximizes the high arrival rate of the two-state MMPP. This yields more accurate results for 120 lines and slightly less accurate results for 125 lines. On the other hand, although the distribution tail is somewhat affected by where the skewness parameter is matched, the mean and standard deviation results shown in Figs. 2 and 3 are relatively insensitive to t_2. Furthermore, as noted earlier, the index of dispersion results shown in Fig. 1 are insensitive to t_2. Investigation of the determination of good choices for t_2 is left to future work. Considering the fact that we are using a two-level process to approximate a process of more than 120 levels, the results are amazingly accurate. The use of more levels in the MMPP is also left for future work.

VI. Conclusions

A methodology has been developed for characterizing a superposition of packetized voice sources and data traffic as an MMPP and for evaluating the performance of a statistical multiplexer with the above input. The methodology is rich in that it uses powerful, modern, matrix-analytic techniques and enables us to obtain important performance measures, such as tails of voice packet delay distributions. It is also rich in that it enables, as outlined, the study of bit-dropping overload control mechanisms. We also note that we have avoided the need for using expensive simulations. While the methodology has been applied to an integrated voice/data problem, it can also be

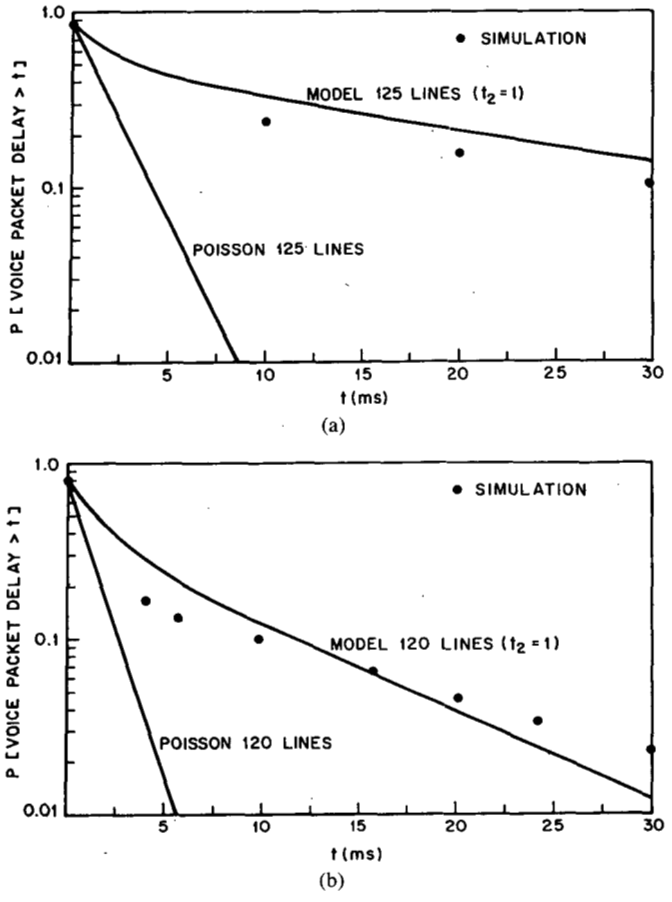


Fig. 6. (a) Complementary voice packet distribution (125 lines) (effect of t_2). (b) Complementary voice packet delay distribution (120 lines) (effect of t_2).

used to characterize more general superpositions of renewal, or even Markov-modulated processes.

We finally observe that the methodology is based on *traffic count statistics* which have been obtained analytically. In a postcutover environment they may also be directly obtained from system measurements. The statistical accuracy of the traffic parameter estimates based on counts, for the class of inputs considered in this paper, is left for future work. We note that the problem of estimating parameters of an MMPP, based on measurements of actual arrival times of customers, has been considered in [31].

Appendix A

We obtain explicit closed-form expressions for the third moment of the number of arrivals in the interval $(0, t)$ for a two-state MMPP. Recalling that an MMPP is a doubly stochastic Poisson process where the rate process is determined by the state of a continuous-time Markov chain, we denote r_1^{-1} and r_2^{-1} as the mean sojourn times in states 1 and 2, respectively, of the underlying chain and λ_j ($j = 1, 2$) as the arrival rate of the Poisson process when the chain is in state j. Also, denote π as the equilibrium probability vector for the Markov chain. Since the two-state MMPP is a special case of the Versatile Markovian Point Process treated in [3], we specialize those results to obtain the probability generating function of the number of arrivals in an interval of length t,

$$g(z, t) = \pi \exp\{[R + (z - 1)\Lambda]t\}e, \qquad \text{(A1)}$$

where, for the two-state MMPP,

$$\pi = \frac{1}{r_1 + r_2}(r_2, r_1), \qquad e = (1, 1)^T$$

$$R = \begin{bmatrix} -r_1 & r_1 \\ r_2 & -r_2 \end{bmatrix}, \quad \text{and} \quad \Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}.$$

Defining N_t as the number of arrivals in $(0, t)$ and $\bar{N}_t = E(N_t)$, we clearly have

$$E[(N_t - \bar{N}_t)^3] = \frac{\partial^3 g(z, t)}{\partial z^3}\bigg|_{z=1} - 3\bar{N}_t(\bar{N}_t - 1)\frac{\text{var}(N_t)}{\bar{N}_t} - \bar{N}_t(\bar{N}_t - 1)(\bar{N}_t - 2). \qquad \text{(A2)}$$

To evaluate the required third derivative we observe that

$$g^{(3)}(1, t) = \frac{\partial^3 g(z, t)}{\partial z^3}\bigg|_{z=1} = \boldsymbol{L}^{-1}\left[\frac{\partial^3 \tilde{g}(z, s)}{\partial z^3}\bigg|_{z=1}\right]$$

where

$$\tilde{g}(z, s) = \pi[sI - R - (z - 1)\Lambda]^{-1}e$$

is the Laplace transform of $g(z, t)$ and $\boldsymbol{L}^{-1}$ denotes the inverse Laplace transform. This gives

$$g^{(3)}(1, t) = \boldsymbol{L}^{-1}\{6\pi[(sI - R)^{-1}\Lambda]^3(sI - R)^{-1}e\} = \boldsymbol{L}^{-1}\left\{\frac{6}{s^2}\pi\Lambda[(sI - R)^{-1}\Lambda]^2 e\right\}, \qquad \text{(A3)}$$

which upon inversion yields

$$g^{(3)}(1, t) = \frac{6}{r_1 + r_2}\left[\frac{A_{11}}{6}t^3 + \frac{A_{21}}{2}t^2 + A_{31}t + A_{12}te^{-(r_1 + r_2)t} + A_{41}(1 - e^{-(r_1 + r_2)t})\right] \qquad \text{(A4)}$$

where

$$A_{11} = \frac{(\lambda_1 r_2 + \lambda_2 r_1)^3}{(r_1 + r_2)^2},$$

$$A_{21} = \frac{2r_1 r_2(\lambda_1 - \lambda_2)^2(\lambda_1 r_2 + \lambda_2 r_1)}{(r_1 + r_2)^3},$$

$$A_{31} = \frac{r_1 r_2(\lambda_1 - \lambda_2)^2[\lambda_1 r_1 + \lambda_2 r_2 - 2(\lambda_1 r_2 + \lambda_2 r_1)]}{(r_1 + r_2)^4},$$

$$A_{41} = \frac{-2r_1 r_2(\lambda_1 - \lambda_2)^3(r_1 - r_2)}{(r_1 + r_2)^5},$$

and finally

$$A_{12} = \frac{r_1 r_2(\lambda_1 - \lambda_2)^2(\lambda_1 r_1 + \lambda_2 r_2)}{(r_1 + r_2)^4}.$$

To evaluate the third moment in (A2) we also require $\overline{N}_t$, which is clearly given by

$$\overline{N}_t = \frac{\lambda_1 r_2 + \lambda_2 r_1}{r_1 + r_2}\, t, \tag{A5}$$

and the variance-to-mean ratio of N_t, which can easily be shown (e.g., [1], [3], [24] or from intermediate results of above analysis) to be given by

$$\frac{\text{var}\,(N_t)}{\overline{N}_t} = 1 + \frac{2(\lambda_1 - \lambda_2)^2 r_1 r_2}{(r_1 + r_2)^2(\lambda_1 r_2 + \lambda_2 r_1)} - \frac{2(\lambda_1 - \lambda_2)^2 r_1 r_2}{(r_1 + r_2)^3(\lambda_1 r_2 + \lambda_2 r_1)t}\,(1 - e^{-(r_1 + r_2)t}). \tag{A6}$$

Appendix B

In this appendix we outline a general procedure for solving the MMPP/G/1 queue. These methods were pioneered by Neuts and have been used successfully to solve a large number of stochastic models [2]–[5]. The results presented here are not new. In fact, the MMPP/G/1 queue is a special case of the much more general $N/G/1$ queue studied by Ramaswami [4]. Our purpose of including this discussion here is to make these methods accessible to a wider audience. Because of space limitations in this appendix, only an outline of the analysis is presented. The reader should supply the proofs and intermediate derivations as they often involve analytic manipulations which are useful in many other applications of the methodology. Similar proofs and derivations are contained in the above references.

The basic philosophy of the methodology is to use probabilistic arguments to derive relationships which will lead to simple stable algorithmic procedures for obtaining quantities of interest. The essential tool is Markov renewal theory [27].

The definition of the MMPP and its parameters has already been given. Here we assume that this process generates arrivals to a single-server queue with a general service time distribution $\tilde{H}(\cdot)$ with moments $\mu^{(i)}$, $i \geq 1$, when they exist. Let $\{\tau_n\colon n \geq 0\}$ denote the successive epochs of departure (with $\tau_0 = 0$), and define X_n and J_n to be, respectively, the number of customers in the system and the phase of the MMPP at τ_n^+. The sequence $\{(X_n, J_n, \tau_{n+1} - \tau_n)\colon n \geq 0\}$ forms a semi-Markov sequence on the state space $\{0, 1, \cdots\} \times \{1, \cdots, m\}$ with transition probability matrix $\tilde{Q}(\cdot)$ given by

$$\tilde{Q}(x) = \begin{bmatrix} \tilde{B}_0(x) & \tilde{B}_1(x) & \tilde{B}_2(x) & \cdots \\ \tilde{A}_0(x) & \tilde{A}_1(x) & \tilde{A}_2(x) & \cdots \\ 0 & \tilde{A}_0(x) & \tilde{A}_1(x) & \cdots \\ 0 & 0 & \tilde{A}_0(x) & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}, \quad x \geq 0, \tag{B1}$$

where the $m \times m$ matrices of mass functions

$$\tilde{A}_n(x) = \int_0^x P(n, t)\, d\tilde{H}(t), \qquad n \geq 0,\ x \geq 0,$$

$$\tilde{B}_n(x) = \tilde{U} * \tilde{A}_n(x), \qquad n \geq 0,\ x \geq 0,$$

and

$$\tilde{U}(x) = \int_0^x e^{(R - \Lambda)t}\, \Lambda\, dt$$

where $P_{ij}(n, t)$ is the conditional probability, given that the underlying Markov process starts in state i at time 0, that at time t it is in state j and there were n arrivals in $(0, t)$. Note that $\tilde{U}_{ij}(x)$ is just the probability that the first arrival to an empty system occurs at or before time x with the MMPP in phase j given that the idle period started at time 0 in phase i.

The goal of this analysis is to derive the virtual waiting time distribution. It is clear that this will involve the stationary distribution of the number in system and phase of the MMPP at an arbitrary time epoch. In order to derive this, we first need the stationary distribution embedded at departure epochs. In particular, we need the stationary probability that a departure leaves the system empty with the MMPP in state j, denoted by $x(0, j)$. This is equal to the reciprocal of the expected number of transitions in the Markov chain with transition matrix $Q = \tilde{Q}(\infty)$, between successive returns to $(0, j)$. We refer to the set of states $\{(i, 1), \cdots, (i, m)\}$, corresponding to i customers in the system, as level $\boldsymbol{i}$. It is clear from the structure of Q that the process may move up many levels in one transition, but may move down only one level at a time (since we are observing at departures). Therefore, the mean number of steps between transitions from $\boldsymbol{i} + \mathbf{1}$ to $\boldsymbol{i}$ is a crucial ingredient in the mean recurrence times to level $\mathbf{0}$. Also, from the structure of Q, the probabilistic structure of transitions from $\boldsymbol{i} + \mathbf{1}$ to $\boldsymbol{i}$ is independent of $\boldsymbol{i}$, for $i \geq 0$. We now study this structure.

Let the matrix $\tilde{G}(n)$ have components $\tilde{G}_{jk}(n)$ defined as the probability that the embedded Markov chain reaches level i for the first time in state (i, k) in exactly n transitions starting in state $(i + 1, j)$. A first passage argument shows that the transform matrix $G(z) = \sum_{n=0}^{\infty} z^n \tilde{G}(n)$ satisfies

$$G(z) = z \sum_{n=0}^{\infty} A_n G(z)^n, \qquad |z| \leq 1, \tag{B2}$$

where $A_n = \tilde{A}_n(\infty)$. The system is stable if and only if the unique matrix $G = G(1)$ is stochastic (i.e., starting in $\boldsymbol{i} + \mathbf{1}$, $\boldsymbol{i}$ will be reached eventually with probability 1). This is the case if and only if $\rho = \pi\lambda\mu^{(1)} < 1$ where $\lambda = \Lambda e$. Let $\boldsymbol{g}$ be the stationary vector of G, i.e., $\boldsymbol{gG} = \boldsymbol{g}$, $\boldsymbol{ge} = 1$. Note that $\boldsymbol{eg}$ is an $m \times m$ matrix and $(I - G + \boldsymbol{eg})$ is nonsingular. Then the vector $\boldsymbol{\mu}$ whose component μ_j is the expected number of transitions from $(i + 1, j)$ to any state in level $\boldsymbol{i}$ is given by

$$\mu = \frac{d}{dz} G(z)\Big|_{z=1} e$$
$$= (I - G + eg)[I - A + eg - \beta g]^{-1} e \quad \text{(B3)}$$

where

$$A = \sum_{n=0}^{\infty} A_n = \int_0^{\infty} e^{Rt}\, d\tilde{H}(t)$$

and

$$\beta = \sum_{n=0}^{\infty} nA_n e = \mu^{(1)}(\pi\lambda)e + (R + e\pi)^{-1}(A - I)\lambda.$$

We now concentrate on the returns to level **0**. Let the matrix $\tilde{L}(n)$ have entries $\tilde{L}_{ij}(n)$ defined as the probability that the first return to level **0** occurs in state $(0, j)$ in exactly n transitions of the embedded Markov chain given that it started in $(0, i)$. The matrix generating function $L(z) = \Sigma_{n=0}^{\infty} z^n \tilde{L}(n)$ is given by

$$L(z) = z \sum_{n=0}^{\infty} B_n G(z)^n = UG(z)$$

where $B_n = \tilde{B}_n(\infty)$ and $U = (\Lambda - R)^{-1}\Lambda$. The matrix $L = L(1) = UG$ is stochastic with stationary vector $\boldsymbol{l}$ defined by $\boldsymbol{l}UG = \boldsymbol{l}$, $\boldsymbol{l}e = 1$. The vector $\boldsymbol{l}^*$ whose component l_j^* is the mean number of transitions in the embedded Markov chain before the first return to any state in level **0**, starting in $(0, j)$, is given by

$$\boldsymbol{l}^* = \frac{d}{dz} L(z)\Big|_{z=1} e = U\mu. \quad \text{(B4)}$$

Now, if we observe the embedded Markov chain with transition matrix Q only at visits to level **0** and keep track of the number of transitions between such visits, then viewing each transition as a discrete time step, the successive states visited, $(0, j)$, and the times between returns to level **0** form a discrete Markov renewal process with transition functions given by $\{\tilde{L}(n)\}$. A classic theorem in Markov renewal theory [27] states that the mean recurrence time of state $(0, j)$ in such a process is given by $\boldsymbol{l}\boldsymbol{l}^* l_j^{-1}$. But this mean recurrence time is none other than the expected number of transitions in the infinite Markov chain between successive returns to state $(0, j)$, and therefore the vector $\boldsymbol{x}_0$ with components $x(0, j)$ is given by $\boldsymbol{x}_0 = \boldsymbol{l}(\boldsymbol{l}U\mu)^{-1}$.

We next derive the stationary probability $y(0, j)$ of being in state $(0, j)$ at an arbitrary point in time. Let $R_{kl}(x)$ be the expected number of visits to state (k, l) in $[0, x]$ by the Markov renewal process $\tilde{Q}(x)$ given that it started in state (i, j) at time 0. Defining $X(t)$ and $J(t)$, respectively, as the number in system and phase of the MMPP at time t, we have, by conditioning on the last transition (i.e., departure) before t, that

$$y_{ij}(0, l, t) \equiv P\{X(t) = 0, J(t) = l \mid X(0) = i, J(0) = j\}$$
$$= \sum_{k=1}^{m} \int_0^t dR_{0k}^{ij}(u)\,[e^{(R-\Lambda)(t-u)}]_{kl}.$$

Applying the key renewal theorem [27] leads to

$$y(0, l) = \lim_{t\to\infty} y_{ij}(0, l, t)$$
$$= \sum_{k=1}^{m} \frac{1}{m(0, k)} \left[\int_0^{\infty} e^{(R-\Lambda)t}\, dt\right]_{kl}$$

where $m(0, k)$ is the mean recurrence time of $(0, k)$ in $\tilde{Q}(\cdot)$. By tracking the continuous time between visits to level **0**, we obtain expressions for $m(0, k)$ which lead to the following expression for the vector $\boldsymbol{y}_0$ [with components $y(0, j)$]:

$$\boldsymbol{y}_0 = (\pi\lambda)\,\boldsymbol{x}_0(\Lambda - R)^{-1}. \quad \text{(B5)}$$

It can be shown that $\boldsymbol{y}_0 e = 1 - \rho$, as one would expect.

The virtual waiting time $V(t)$ is the time a customer who arrives at time t would wait before entering service. By again conditioning on the last departure before time t, we get

$$P\{0 < V(t) \le x, J(t) = j \mid X(0) = i, J(0) = l\}$$
$$= \sum_{k=1}^{m} \sum_{v_1=1}^{\infty} \int_{\tau=0}^{t} dR_{v_1 k}^{il}(\tau) \sum_{v_2=0}^{\infty} P_{kj}(v_2, t - \tau)$$
$$\cdot \int_{w=0}^{x} d\tilde{H}(t + w - \tau)\, \tilde{H}^{(v_1+v_2-1)}(x - w)$$
$$+ \sum_{k=1}^{m} \int_{\tau=0}^{t} dR_{0k}^{il}(\tau) \int_{u=0}^{t-\tau} \sum_{p=1}^{m} [e^{(R-\Lambda)u}\,\Lambda\, du]_{kp}$$
$$\cdot \sum_{v_2=0}^{\infty} P_{pj}(v_2, t - \tau - u)$$
$$\cdot \int_{w=0}^{x} d\tilde{H}(t + w - \tau - u)\, \tilde{H}^{(v_2)}(x - w) \quad \text{(B6)}$$

where $\tilde{H}^{(n)}$ is the n-fold convolution of $\tilde{H}(\cdot)$ with itself. The first term is obtained as follows. The last departure before time t occurred at time τ and left $v_1 \ge 1$ customers in the system. Between τ and t there were $v_2 \ge 0$ arrivals, so that at time t there are a total of $v_1 + v_2$ customers in the system. The service time of the customer who entered service at time τ ends at some time $t + w$. Now, for the customer who arrived at time t to enter service before time $t + x$, there must be $v_1 + v_2 - 1$ departures in $x - w$. The second term is obtained in a similar way except that the last departure at time τ leaves the system empty, and we must keep track of the phase of the MMPP during the idle period. Let the vector $\tilde{\boldsymbol{W}}(x)$ have components $W_j(x) = \lim_{t\to\infty} P\{V(t) \le x, J(t) = J \mid X(0) = i, J(0) = l\}$ and denote the LST of $\tilde{\boldsymbol{W}}(x)$ by $\boldsymbol{W}(s)$. By applying the key renewal theorem to (B6), taking transforms, and performing some tedious manipulations, we finally obtain the joint transform of the virtual waiting time and the phase of the MMPP as

$$\boldsymbol{W}(s) = \begin{cases} s\boldsymbol{y}_0[sI + R - \Lambda + \Lambda H(s)]^{-1}, & s > 0, \\ \boldsymbol{\pi}, & s = 0, \end{cases} \quad \text{(B7)}$$

where $H(s)$ is the LST of $\tilde{H}(x)$. Note that if $\lambda_j = \lambda$ for all j, or if the MMPP has only one phase, then the above expression reduces to $W(s) = s(1 - \rho)/(s - \lambda + \lambda H(s))$, which is the familiar Pollaczek–Khinchin formula for the $M/G/1$ queue.

Remarks on the Numerical Procedures: 1) It is shown in Lucantoni and Ramaswami [29] that the matrix G needed in the above analysis may be computed efficiently by successive substitutions in the following system. Start with $G_0 = 0$, and for $k = 0, 1, 2, \cdots$ compute

$$H_{n+1,k} = [I + \theta^{-1}(R - \Lambda)]H_{n,k} + \theta^{-1}\Lambda H_{n,k}G_k, \quad n = 0, 1, 2, \cdots,$$
$$G_{k+1} = \sum_{n=0}^{\infty} \gamma_n H_{n,k} \tag{B8}$$

where $H_{0,k} = I$, $\theta = \max(\lambda_j - R_{jj})$ and $\gamma_n = \int_0^\infty e^{-\theta x}((\theta x)^n/n!)\, d\tilde{H}(x)$. It is shown in [29] that the sequence G_k converges monotonically to G. This procedure avoids computing and storing the matrices A_n.

2) We note that when $\tilde{H}(\cdot)$ is deterministic with mass at d, then γ_n is computed recursively by $\gamma_0 = e^{-\theta d}$, $\gamma_n = (\theta d/n)\,\gamma_{n-1}$, $n \geq 1$. If $H(\cdot)$ is geometric, i.e.,

$$\tilde{H}(x) = \sum_{n=1}^{[x]} (1 - p)p^{n-1}, \quad nl \leq x < (n + 1)l,\ n = 1, 2, \cdots,$$

where $[x]$ is the integer part of x, then γ_n may be computed recursively by

$$\gamma_0 = \frac{(1 - p)e^{-\theta l}}{1 - pe^{-\theta l}},$$
$$\gamma_n = \frac{(1 - p)\xi_n}{1 - pe^{-\theta l}} + \frac{p}{1 - pe^{-\theta l}} \sum_{k=1}^{n} \xi_k \gamma_{n-k}, \quad n = 1, 2, \cdots,$$

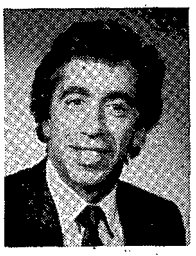
where the sequence $\{\xi_n\}$ is given by $\xi_0 = e^{-\theta l}$, $\xi_n = (\theta l/n)\,\xi_{n-1}$, $n = 1, 2, \cdots$.

Acknowledgment

The authors acknowledge M. Eisenberg for valuable assistance in applying his method for numerical inversion of Laplace transforms and B. Melamed for assistance in the use of the performance analysis workstation [32], which was useful in the validation process. They also acknowledge K. Sriram and W. Whitt for the numerical results in [17] and simulation results produced by A. Anastasio.

References

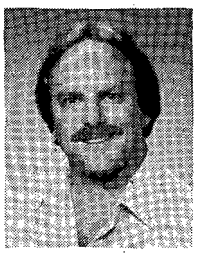
[1] H. Heffes, "A class of data traffic processes—Covariance function characterization and related queueing results," *Bell Syst. Tech. J.*, vol. 59, pp. 897–929, July/Aug. 1980.
[2] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach.* Baltimore, MD: Johns Hopkins Univ. Press, 1981.
[3] M. F. Neuts, "A versatile Markovian point process," *J. Appl. Prob.*, vol. 16, pp. 764–779, Dec. 1979.
[4] V. Ramaswami, "The *N/G/1* queue and its detailed analysis," *Adv. Appl. Prob.*, vol. 12, pp. 222–261, Mar. 1980.
[5] D. M. Lucantoni, *An Algorithmic Analysis of a Communication Model with Retransmission of Flawed Messages.* London: Pitman, 1983.
[6] T. Bially, B. Gold, and S. Seneff, "A technique for adaptive voice flow control in integrated packet networks," *IEEE Trans. Commun.*, vol. COM-28, pp. 325–333, Mar. 1980.
[7] M. Listanti and F. Villani, "Voice communication handling in *X.25* packet switching networks," presented at Globecom '83, San Diego, CA, Nov. 28–Dec. 1, 1983, Paper 2.4.
[8] J. M. Holtzman, "The interaction between queueing and voice quality in variable bit rate packet voice systems," in *Proc. ITC 11*, Kyoto, Japan, Sept. 4–11, 1985, Paper 2.2A-4.
[9] D. Anick, D. Mitra, and M. M. Sondhi, "Stochastic theory of a data-handling system with multiple sources," *Bell Syst. Tech. J.*, vol. 61, pp. 1871–1894, Oct. 1982.
[10] Y. K. Tham and J. N. P. Hume, "Analysis of voice and low-priority data traffic by means of brisk periods and slack periods," *Comput. Commun.*, vol. 6, no. 1, pp. 14–22, Feb. 1983.
[11] A. Weiss and L. Wyatt, private communication.
[12] T. E. Stern, "A queueing analysis of packet voice," in *Proc. IEEE Global Telecommun., Conf.*, San Diego, CA, Dec. 1983, pp. 2.5.1–2.5.6.
[13] Y. C. Jenq, "Approximations for packetized voice traffic in statistical multiplexer," in *Proc. IEEE Infocom*, Apr. 1984, pp. 256–259.
[14] M. L. Luhanga, "Analytical model of a packet voice concentrator," Columbia Univ., New York, Tech. Rep. 1984-02.
[15] J. N. Daigle and J. D. Langford, "Queueing analysis of a packet voice communication system," in *Proc. IEEE Infocom*, Washington, DC, Mar. 1985.
[16] K. Sriram and W. Whitt, "Characterizing superposition arrival processes in packet multiplexers for voice and data," *IEEE J. Select. Areas Commun.*, this issue, pp. 833–846.
[17] K. Sriram and W. Whitt, "Characterizing superposition arrival processes and the performance of multiplexers for voice and data," presented at Globecom '85, New Orleans, LA, Dec. 1985.
[18] P. T. Brady, "A statistical analysis of on–off patterns in 16 conversions," *Bell Syst. Tech. J.*, vol. 47, no. 1, pp. 73–91, Jan. 1968.
[19] J. G. Gruber, "A comparison of measured and calculated speech temporal parameters relevant to speech activity detection," *IEEE Trans. Commun.*, vol. COM-30, pp. 728–738, Apr. 1982.
[20] Y. Yatsuzuka, "Highly sensitive speech detector and high speed voiceband discriminator in DSI-ADPCM system," *IEEE Trans. Commun.*, vol. COM-30, pp. 739–750, Apr. 1982.
[21] K. Sriram, P. K. Varshney, and J. G. Shanthikumar, "Discrete-time analysis of integrated voice–data multiplexers with and without speech activity detectors," *IEEE J. Select. Areas Commun.*, vol. SAC-1, Special Issue on Packet Switched Voice and Data Communications, Dec. 1983.
[22] C. J. May and T. J. Zebo, private communication.
[23] L. Takàcs, *Introduction to the Theory of Queues.* New York: Oxford Univ. Press, 1962.
[24] D. R. Cox and P. A. W. Lewis, *The Statistical Analysis of Series of Events.* London: Methuen, 1966.
[25] D. L. Jagerman, "An inversion technique for the Laplace transform," *Bell Syst. Tech. J.*, vol. 61, no. 8, pp. 1995–2002, 1982.
[26] L. Kleinrock, *Queueing Systems, Volume 1: Theory.* New York: Wiley, 1975.
[27] E. Çinlar, *Introduction to Stochastic Processes.* Englewood Cliffs, NJ: Prentice-Hall, 1975.
[28] R. W. Wolff, "Poisson arrivals see time averages," *Oper. Res.*, vol. 30, no. 2, pp. 223–231, 1982.
[29] D. M. Lucantoni and V. Ramaswami, "Efficient algorithms for solving the non-linear matrix equations arising in phase type queues," *Stochast. Models*, vol. 1, no. 1, pp. 29–52, 1985.
[30] M. Eisenberg, unpublished work.
[31] K. S. Meier, "A statistical procedure for fitting Markov-modulated Poisson processes," Ph.D. dissertation, Univ. Delaware, Newark, DE, Dec., 1984.
[32] B. Melamed, "The performance analysis workstation: An interactive animated simulation package for queueing networks," in *Proc. Fall Joint Comput. Conf.*, Dallas, TX, 1986, to appear.



Harry Heffes (M'66-SM'82) received the B.E.E. degree from the City College of New York, New York, in 1962 and received the M.E.E. degree in 1964 and the Ph.D. degree in electrical engineering in 1968, both from New York University, Bronx, NY.

He joined AT&T Bell Laboratories in 1962 in the Military Systems Area where he applied modern control and estimation theory results to problems relating to guidance, navigation, tracking, and trajectory optimization. Since 1973 his primary concern has been with the modeling and analysis of teletraffic processes and systems. Most recently he has been concerned with the performance analysis of computer-based systems and services, including digital switching systems. He is also an Adjunct Professor of Computer Science at Stevens Institute of Technology, Hoboken, NJ.

Dr. Heffes is the author of over 20 papers in such areas as Kalman filtering, control system theory, approximation theory, communication theory, air traffic control, teletraffic theory, queueing theory, simulation, switching systems, data traffic, overload control, communication network survivability, and integrated voice/data systems. He received the AT&T Bell Laboratories Distinguished Technical Staff Award in 1983. He is a member of Tau Beta Pi, Eta Kappa Nu, American Men and Women of Science, ORSA, and the Association for Computing Machinery.



David M. Lucantoni was born in Baltimore, MD, on August 31, 1954. He received the B.S. degree in mathematics from Towson State University, Baltimore, MD, in 1976 and received the M.S. degree in statistics in 1978, and the Ph.D. degree in operations research in 1981, both from the University of Delaware, Newark, DE.

Since 1981 he has been at AT&T Bell Laboratories where he has worked on the overload control design and performance of switching systems. He is currently working on the overload control and performance analysis of voice/data packet networks. His current research interests are in the area of the algorithmic analysis of stochastic models and queueing theory.

Dr. Lucantoni is a member of ORSA and AMS.