Models for Analysis of Packet Voice Communications Systems

JOHN N. DAIGLE, SENIOR MEMBER, IEEE, AND JOSEPH D. LANGFORD

Abstract—**In a packet voice communication system, packets are fed to a common queue by a number of independent voice sources and are removed from this queue on a first-come-first-serve basis for transmission over a communication link of finite capacity. Each voice source alternates between active periods, during which packets are generated at regular intervals, and inactive periods, during which no packets are generated. In this paper, we discuss three models, a semi-Markov process model, a continuous-time Markov chain model, and a uniform arrival and service model, to assess the queueing behavior of such systems. Numerical results obtained from each of the three models are compared to each other, to results obtained from a discrete event simulation program, and to results obtained from an *M/D/1* analysis. Parameters of the model are the average duration of active and inactive periods, the packet generation rate, the communication link capacity, and the total number of voice sources. Conclusions are drawn regarding which models appear to be most appropriate in the parameter ranges investigated.**

I. INTRODUCTION

THE widespread deployment of packet-switched computer communication networks and the resource sharing that they make possible [1] has stimulated interest in transmission of packetized voice signals over these networks [2]. But, since active voice sources generate periodic packet streams, the statistical properties of the voice packet arrival process differ from those of the bursty non-voice traffic which is typically carried on computer communication networks [3]. Hence, in order to design networks which meet the strict delay constraints required for acceptable voice reconstruction [4], models which accurately reflect the statistical properties of packet voice systems are needed.

Typical behavior of a voice source, which generates packets from a voice signal, is illustrated in Fig. 1. A voice source is "active" when the talker is actually speaking. During active periods, the voice source generates fixed length packets at regular intervals which are depicted in Fig. 1 by vertical arrows. The voice source is "inactive" and generates no packets during periods in which the speaker is silent. In normal conversation, the duration of active periods fits the exponential distribution reasonably well while the duration of inactive periods is approximated less well by the exponential distribution [5]. Nonetheless, to facilitate analysis, the lengths of both active and inactive periods are assumed to be exponentially distributed in this paper. The quality of the resulting model has been found to be quite good if the number of voice sources is more than about 25 and less suitable if the number of voice sources is less than about 10 [6].

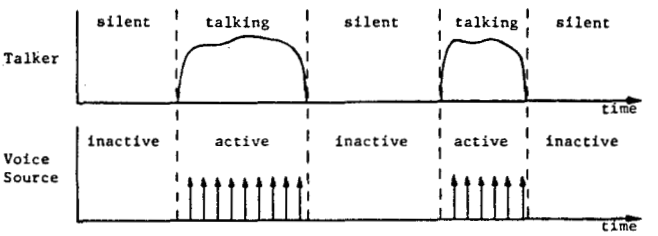


Fig. 1. Typical voice source behavior.

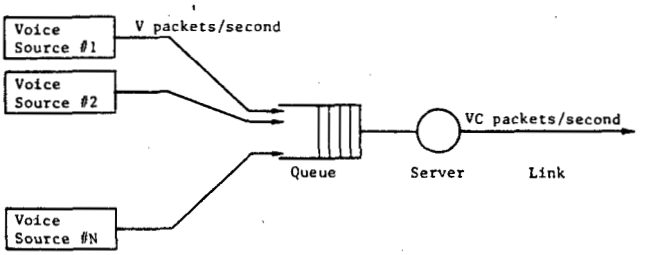


Fig. 2. Packet voice statistical multiplexing system.

Fig. 2 depicts the system being modeled. There are N independent voice sources, each of which generates a packet every $1/V$ s while active. The packets so generated are fed instantaneously to a common queue from which a server removes them for transmission over a communications link, the required transmission time per packet being $1/VC$ s. Hence, C active voice sources will just saturate the link, and C is called the link capacity. During periods in which there are more than C active voice sources, packets accumulate in the queue, and this backlog of packets is eliminated during periods in which the number of active sources falls below C.

Few analytical treatments of problems of the type depicted in Fig. 2 have appeared in the literature; for a review of those treatments, see [7]. In Section II of this paper, we discuss three such analytic models for approximating the equilibrium queue length distribution for the packet voice statistical multiplexing system depicted in Fig. 2. The queue length distribution is important because it is closely related to the delay that packets experience in the system. Thus, the models provide analytic tools for assessing the performance of the multiplexing system as a function of the system parameters. Numerical results

Manuscript received November 4, 1985; revised March 6, 1986.

The authors are with the Graduate School of Management, University of Rochester, Rochester, NY 14626.

IEEE Log Number 8610300.

obtained from each of the three models are presented and compared in Section III. We then use discrete event simulation to assess the quality of the results obtained from each of the models and from an $M/D/1$ analysis.

II. Model Descriptions

In this section, we describe three analytical models for computing approximate queue length distributions for the packet voice communication system depicted in Fig. 2. In each case, the alternating renewal process nature and the independent operation of the voice sources are maintained, but the packet arrival process during active periods and the nature of the service process are altered in each of the models to obtain analytic tractability.

In the first model, the rate of change in queue length is proportional to the difference between the line capacity and the number of active sources. During a period in which j voice sources are active the queue length grows by a geometric number of packets where the parameter of the geometric distribution is based upon the difference between C and j. This change in queue length may be either positive or negative depending upon whether j is larger or smaller than C. The approximations implied by this behavior are identical to those underlying the "underload/overload" analysis of Stern [8]. Under these conditions, if the state of the system is taken to be the number of active voice sources and the number of packets in the queue, the resulting model is a semi-Markov process (SMP). The Markov chain governing the state transitions of the semi-Markov process can be solved using Neuts' matrix geometric techniques [9], and then the ergodic probabilities for the semi-Markov process are obtained by weighting the equilibrium probabilities of the underlying Markov chain.

The second model is a continuous-time Markov chain in which the state is defined to be the number of active sources and the number of packets in the system. In this model, each active source produces packets with exponentially distributed lengths according to a Poisson process. The equilibrium distribution of the number of packets in the system is computed directly from the equilibrium balance equations using the matrix geometric solution technique of Neuts [9].

The third model, analyzed by Pilc [10] and later analyzed more elegantly by Anick, Mitra, and Sondhi [11], [12] in the context of data communications, supposes that each active source transmits information uniformly, and the server removes information from the queue in the same manner. The latter authors use a forward equation approach to derive a differential equation which is solved to yield the equilibrium distribution of the queue content and the number of active sources. They present a computational procedure in a greatly simplified form in which the eigenvalues of the system are found as roots of quadratic equations and the remaining calculations require only algebraic manipulation. This model differs from the SMP model in that it lacks a concept of packetization.

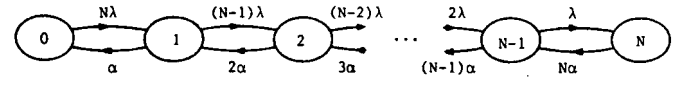


Fig. 3. The phase process.

A. System Properties

Since the durations of active and inactive periods of the voice sources are assumed to be exponentially distributed with parameters α and λ, respectively, the number of active sources as a function of time can be modeled as a continuous-time Markov chain as shown in Fig. 3. Following Stern [8], we shall refer to this process as the "phase process," and we shall denote the phase state (that is, the number of active voice sources) by $\wp$. As was pointed out in the introduction, each active source delivers one packet to the queue every $1/V$ s, and the server removes one packet from the queue (when it is nonempty) every $1/VC$ s. Three conditions are possible:

1) $\wp = j \leq C$. In this case, the average arrival rate of packets to the queue is less than the average service rate. The queue length decreases at an average rate of $V(C - j)$ packets/s so long as the queue is nonempty. If the queue is empty, then it remains empty in an average sense so long as $\wp = j \leq C$.

2) $\wp = j = C$. Under this condition, which occurs only if C is an integer, the packet arrival and service rates are equal so that the rate of change of the queue length is zero.

3) $\wp = j \geq C$. Here, the average arrival rate of packets to the queue exceeds the service rate. The length of the queue increases at an average rate of $V(j - C)$ packets/s.

Each of the models discussed in this paper captures this behavior in an average sense, but the dynamics of the models differ from those of the actual system.

B. Semi-Markov Process Model

For the semi-Markov process (SMP) model, which is described in detail in [13], the following approximations are adopted. So long as the phase process is in state $\wp = j = C$, the queue length does not change. While the phase process is in state $\wp = j < C$, then so long as the queue is not empty, the queue length is decreased by one packet every $1/[V(C - j)]$ s starting from the point in time when the phase process entered state j. Finally, while $\wp = j > C$, the queue length is increased by one packet every $1/V(j - C)]$ s starting from the point in time at which the phase process entered state j.

Some implications of these approximations are now discussed. First, the approximations adopted for this model ignore the "high frequency" fluctuations present in the system being modeled. For example, in the real system any time a voice source generates a packet during the transmission of another packet, the queue length would increase by one packet independent of the current phase state. Yet, the SMP model assumes that while $\wp = C$, the queue length *never* changes and while $\wp < C$ *any* queue length change is a decrease. Similarly, in the real system, any time the server completes the service of a packet, the

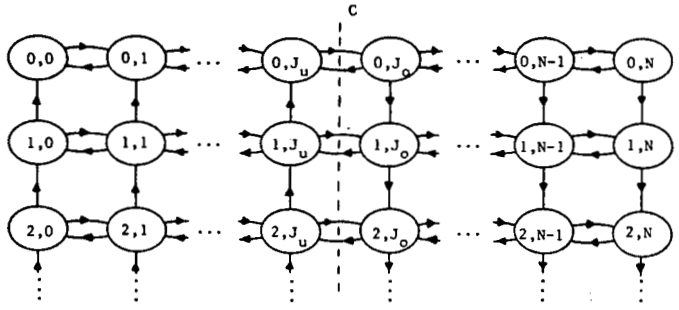


Fig. 4. State diagram for the semi-Markov process model.

server removes one packet from the queue (if the queue is not empty) and the queue length decreases, again no matter what phase the system is in. The SMP model assumes, however, that *any* queue length change that occurs while $\wp > C$ is an increase. Results obtained from this model will clearly overestimate the probability that the queue is empty.

A second implication of the approximations is that they imply that a renewal occurs at the instants of phase transition (call this "resynchronization"). That is, no fractional queue length growth (reduction) is accrued to account for packets that are in the process of arriving (being serviced) between the time of the latest queue length change prior to a given phase transition and the time of the phase transition. It is expected that the effect of both of these implications will decrease as the packet-generation rate, V, and the service rate increase relative to the expected duration of the voice source active time.

We now turn our attention to the mathematical description of the model Denote the state of the process, which is depicted in Fig. 4, by $(l(t), \wp(t))$, where $l(t)$ is the number of packets in the queue at time t and $\wp(t)$ is the number of active sources at time t, or the phase state. It is argued in [13] that the assumption of resynchronization of the process at instants of phase transition leads to the conclusion that the transition probabilities for this process depend solely upon the current state of the process. Hence, there exists a Markov chain embedded at instants of phase state changes, queue increments and queue decrements. Further, the expected sojourn time in any state depends only upon the state. Therefore, the process of Fig. 4 is a semi-Markov process, and once the equilibrium probabilities for the embedded Markov chain have been determined the queue length distribution can be determined in a straightforward manner.

In particular, let $p_{i,j} = \lim_{t\to\infty} P\{l(t) = i, \wp(t) = j\}$. Also, let $q_{i,j}$ denote the equilibrium probabilities for the embedded Markov chain; that is, $q_{i,j}$ is the equilibrium fraction of transitions which enter state (i, j). Finally, let m_{ij} denote the expected sojourn time in state (i, j) of the process. Then, from renewal theory [14],

$$p_{i,j} = \frac{q_{i,j}m_{i,j}}{\sum_{k=0}^{\infty}\sum_{l=0}^{N} q_{k,l}m_{k,l}}. \tag{1}$$

Our strategy is to first compute the $q_{i,j}$'s using matrix geometric techniques, and then to compute the $p_{i,j}$'s by using (1).

We now review our procedure, which is presented in detail in [13], for computing the $q_{i,j}$'s. Referring to the state diagram of Fig. 4, let the collection of states $(m, 0)$, $(m, 1)$, $\cdots$, (m, N) be termed "level m," and let q_i denote the row vector $[q_{i,0}, q_{i,1}, \cdots, q_{i,N}]$. Also, let $P_{m\to i}$ denote the matrix of transition probabilities from states at level m to states at level i. Since it is clear from Fig. 4 that state transitions can occur only within the same level or to adjacent levels, it follows that for $i = 0$ the transition balance equations are $q_0 = q_0P_{0\to 0} + q_1P_{1\to 0}$, or in Neuts' notation,

$$q_0 = q_0B_0 + q_1B_1. \tag{2}$$

Similarly, for $i > 0$, $q_i = q_{i-1}P_{i-1\to i} + q_iP_{i\to i} + q_{i+1}P_{i+1\to i}$. Now, for $i > 0$, the transition probabilities are independent of level so that if we let A_0 denote $P_{i-1\to i}$, A_1 denote $P_{i\to i}$, and A_2 denote $P_{i+1\to i}$, then for $i > 0$ we find

$$q_i = q_{i-1}A_0 + q_iA_1 + q_{i+1}A_2. \tag{3}$$

Equations (2) and (3) and the law of total probability define the equilibrium probabilities for the embedded Markov chain. These equations can be solved readily using the matrix geometric solution technique developed by Neuts [9]. In particular, Neuts shows that under certain conditions, which the above system is shown to satisfy in [13], the solution to the system defined by (2) and (3) is given by

$$q_i = q_0R^i = q_{i-1}R \qquad \text{for } i \geq 1. \tag{4}$$

The $(N + 1) \times (N + 1)$ matrix R is the minimal nonnegative solution to the matrix equation

$$R = A_0 + RA_1 + R^2A_2 \tag{5}$$

and q_0 is given by the solution to

$$q_0 = q_0[B_0 + RB_1] \tag{6}$$

normalized by

$$q_0(I - R)^{-1}\, e = 1 \tag{7}$$

where e is the $(N + 1)$ column vector of 1's. Neuts [9] shows that (5) can be solved for R by using successive approximations while (6) and (7) form a linear system of equations which can be solved for q_0. This then settles computation of the equilibrium probabilities, $q_{i,j}$, for the embedded Markov chain.

It remains to apply (1) to compute the equilibrium probabilities, $p_{i,j}$, for the semi-Markov process and to sum these over j to obtain the queue length distribution. It should be noted that it is not necessary to perform an infinite summation to evaluate the denominator of (1). Instead, one can rewrite the denominator of (1) as

$$\sum_{i=0}^{\infty}\sum_{j=0}^{N} q_{i,j}m_{i,j} = \sum_{j=0}^{N} q_{0,j}m_{0,j} + \sum_{i=1}^{\infty}\sum_{j=0}^{N} q_{i,j}m_{i,j}. \tag{8}$$

But, for $i > 0$, it is readily seen that $m_{i,j} = m_{1,j}$, thus (8) can be rewritten as

$$\sum_{i=0}^{\infty}\sum_{j=0}^{N} q_{i,j}m_{i,j} = \sum_{j=0}^{N} q_{0,j}m_{0,j} + \sum_{j=0}^{N} m_{1,j}\left(\sum_{i=0}^{\infty} q_{i,j} - q_{0,j}\right). \quad (9)$$

It is also straightforward to show that $\Sigma_{i=0}^{\infty} q_{i,j} = [q_0(I - R)^{-1}]_j$, that is, the jth element of the vector $q_0(I - R)^{-1}$. Thus, (9) reduces to

$$\sum_{i=0}^{\infty}\sum_{j=0}^{N} q_{i,j}m_{i,j} = \sum_{j=0}^{N} q_{0,j}m_{0,j} + \sum_{j=0}^{N} m_{1,j}\left([q_0(I - R)^{-1}]_j - q_{0,j}\right). \quad (10)$$

In summary, the SMP model is solved according to the following procedure. Using the formulae given in [13], compute the probabilities $P\{m, n \to i, j\}$ that the next state of the embedded Markov chain is (i, j) given that the present state is (m, n) for $m = 0, 1, 2$, and $n = 0, 1, \cdots, N$. Using these transition probabilities, form the transition matrices $B_0 = P_{0\to 0}$, $B_1 = P_{1\to 0}$, $A_0 = P_{0\to 1}$, $A_1 = P_{1\to 1}$, and $A_2 = P_{2\to 1}$. Next, use (5) to solve for R by successive approximations, and then find q_0 by solving (6) and (7). Use the simple expression given at the end of Section IV of [13] to compute $m_{i,j}$ for $i = 0, 1$, and $j = 0, 1, \cdots, N$. For each level i, use (4) to obtain q_i, the right-hand side of (10) to compute the infinite sum in (1), and then use (1) to determine $p_{i,j}$, for $j = 0,1, \cdots, N$. Finally, sum over j to determine π_i, the probability that the queue length is i; that is

$$\pi_i = P\{l = i\} = \sum_{j=0}^{N} p_{i,j}. \quad (11)$$

C. Continuous-Time Markov Chain Model

As before, each voice source becomes active and inactive according to an alternating renewal process in which the durations of the active periods and the inactive periods are exponentially distributed with parameters α and λ, respectively. However, in the continuous-time Markov chain (CTMC) model, the generation of packets by an active voice source occurs according to a Poisson process with rate β rather than at a constant rate of one packet every $1/V$ s as in the system being modeled. Since the superposition of independent Poisson processes is a Poisson process with rate equal to the sum of the rates of the individual processes, it follows that if the phase process is in state $\wp = j$, then the arrival process of packets to the system is Poisson with parameter $j\beta$. Similarly, we assume in this model that service times are exponentially distributed with parameter ν. Thus, so long as the *system*, that is, the server and the queue, is nonempty, the departure process from the queue is Poisson with parameter ν rather than deterministic with a departure every $1/CV$ s as is the case in the system being modeled. As before, we may think of the channel capacity, C, as the number of active sources that will just saturate the server; namely $C = \nu/\beta$.

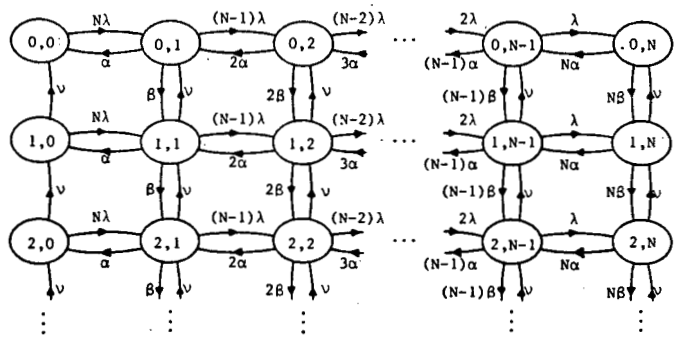


Fig. 5. State diagram for the continuous-time Markov chain model.

To see how the CTMC model dynamics compare to those of the system under investigation, suppose that $\wp = j = C$ during some interval of time. Then, the system occupancy would change at a Poisson rate of $j\beta + \nu$ during this interval of time so long as the system occupancy remained above zero, and at a Poisson rate of $j\beta$ during periods in which the system occupancy fell to zero. In both the real system and the CTMC model, the average rate of growth in system occupancy is zero, but the length of time required to observe this zero average growth rate would be quite different for the two cases. For example, if the queue length at the beginning of an interval during which the phase state is $\wp = j = C$ were, say, 10 packets, then during the entire interval one would expect the queue length to remain near 10 packets with only brief excursions above and below 10 packets. But, in the CTMC case, starting with a system occupancy of 10 packets, the queue length would with probability 1 return to 10 packets once having exceeded 10 packets, but the expected length of time required to do so would be infinite.

The behavior of the model over a given period of time during which the phase process is in state $\wp = j$, $j = 0, 1, \cdots, N$ is identical to the transient behavior of the $M/M/1$ system with arrival rate $j\beta$, service rate ν and initial occupancy equal to that observed at the beginning of the given period of time. The CTMC model, therefore, experiences higher frequency fluctuations in system occupancy than does the real system. It is then expected that the CTMC model would yield conservative estimates for probability of buffer overflow and waiting time. That is, if the system were engineered to achieve a certain probability of overflow based on the CTMC model, then the probability of overflow in the system would probably be less than that predicted by the model.

The state diagram of the CTMC model is shown in Fig. 5. Let the state of this CTMC be denoted by $(s, \wp)$, where s is the number of packets in the system and $\wp$ is the phase state, and let $P_{i,j}$ denote the stochastic equilibrium probability that $s = i$ and $\wp = j$. As explained in elementary texts covering the theory of continuous-time Markov

chains (see Ross [15], for example), it is straightforward to specify equilibrium balance equations from the state diagram. From Fig. 5, the equilibrium balance equations for our system are then readily obtained. These equations are identical to the system (3) of [18].

These balance equations can then be rewritten into the form

$$0 = p_0 B_0 + p_1 B_1, \tag{12}$$

and for $i > 0$,

$$0 = p_{i-1} A_0 + p_i A_1 + p_{i+1} A_2 \tag{13}$$

where p_i denotes the row vector $[p_{i,0}, p_{i,1}, \cdots, p_{i,N}]$ and B_0, B_1, A_0, A_1, and A_2 are constant $(N + 1) \times (N + 1)$ matrices containing appropriate transition rates. The exact transition rate matrices and a procedure for solving these equations using the matrix geometric solution approach of Neuts [9] are presented in [16]. To complete the solution, we note that $\pi_i = P\{l = i\} = P\{s = i + 1\}$ where π_i is the probability that the queue length is i; thus

$$\pi_i = P\{s = i + 1\} = \sum_{j=0}^{N} p_{i+1,j}. \tag{14}$$

Convergence to a solution based on the iterative technique suggested by Neuts [9] was found to be slow, but tolerable. A modification to Neuts' approach for a system of equations very similar to the above in the context of integrated voice/data systems was offered by Williams and Leon-Garcia [21]. They found their solution technique converged much more quickly than the standard approach. An efficient approach to solving the identical system of equations based upon probability generating functions and discrete Fourier transforms is detailed in [22].

D. Uniform Arrival and Service Model

We now turn to the model analyzed in Anick, Mitra, and Sondhi [11], [12]. In that model, which we shall refer to as the "uniform arrival and service" (UAS) model, each active source generates information to the transmission buffer at a uniform rate of 1 unit of information per unit of time, and the server removes information from the buffer at a uniform rate not to exceed C units of information per unit of time. As in the SMP and CTMC models, the server's capacity is C. While the phase process is in state $\varphi = j > C$, the buffer content increases at a constant rate of $j - C$ units of information per unit of time. So long as the buffer is nonempty, while the phase process is in state $\varphi = j < C$, the buffer content is reduced at a constant rate of $C - j$ units of information per unit of time. If the buffer empties while the phase process is in state $\varphi = j < C$, the buffer remains empty until the number of sources again exceeds C. For reasons of analytic tractability, the capacity C cannot be integer valued in this model.

The approximations of this model ignore "high frequency" variations in buffer content as compared to the real system. That is, in the real system, information does not enter the transmission buffer, and therefore cannot be transmitted, until a particular active source completes generation of a packet. In the UAS model, on the other hand, it is possible for information to be transmitted while it is being received. It would appear that the effect of this approximation on the UAS model's accuracy would be slight since the difference between the behavior of this model and that of the real system is greatest when the buffer content is low and the number of active sources is below the server's capacity.

In the Anick, Mitra, and Sondhi papers, the unit of time is taken to be the average duration of an active period or $1/\alpha$ s. The unit of information is taken to be the amount of information that would be generated by a source during an active period of average duration; a unit of information would therefore be equivalent to V/α packets.

We now turn to the mathematical model of the system. Let $B(t)$ and $\varphi(t)$ denote the buffer content and the number of active sources, respectively, at time t. Further, let $P_i(t, b) = P\{\varphi(t) = i, B(t) \leq b\}$ for $0 \leq i \leq N$, $t \geq 0$, $b \geq 0$. Anick, Mitra, and Sondhi first write a set of simple partial differential equations for $P_i(t, b)$ and then obtain a differential equation for $F_i(b) = \lim_{t \to \infty} P_i(t, b)$. The resulting set of equations, in which it is understood that $F_k(b) = 0$ for $k < 0$ and $k > N$, is as follows:

$$(i - C)\frac{dF_i(b)}{db} = (N - i + 1)\frac{\lambda}{\alpha} F_{i-1}(b) - \left\{(N - i)\frac{\lambda}{\alpha} + i\right\} \cdot F_i + (i + 1) F_{i+1}(b). \tag{15}$$

The above system of equations reflects the fact that phase transitions can occur only to states adjacent to the current state.

Anick, Mitra, and Sondhi perform a thorough analysis of the system and they derive simple analytic formulae to compute the complementary occupancy distribution, $G(b)$. This complementary occupancy distribution is, however, specified in "units of information." Since we have pointed out that a unit of information is equivalent to V/α packets, we define the complementary queue length distribution for this system, $P\{l > i\}$, to be the probability that the buffer occupancy exceeds $(\alpha/V)i$ units of information; that is,

$$P\{l > i\} \stackrel{\text{def}}{=} G\left(\frac{\alpha}{V} i\right) \tag{16}$$

for the UAS model.

A thorough discussion of the details of our implementation of the Anick, Mitra, and Sondhi solution technique is given in [23] which may be obtained by writing the authors.

III. Results

For each of the three models discussed in Section II, a computer program which computes the equilibrium dis-

tribution of the number of packets in the queue was written. These programs have been used to generate queue length distributions for a number of parameter sets. The parameters of the system are λ^{-1} (seconds), the mean inactive time for the voice source, α^{-1} (seconds), the mean active time for the voice source, V (packets/s), the voice source packet-generation rate, N, the number of voice sources, and C, the link capacity. Three parameter sets, PS0, PS1, and PS2, are considered. The parameters of PS0 are chosen on the basis of speech activity models that have been reported in the literature, and those of PS1 and PS2 are chosen to assess the usefulness of the models over a broad range of parameters. In order to evaluate the accuracy of the models, a GPSSV [17] simulation program which computes approximate queue length distributions for the packet voice statistical multiplexing system was written.

A. Parameter Choices

In a frequently referenced study, Brady [5] determined that the mean voice source inactive time ($1/\lambda$) and the mean voice source active time ($1/\alpha$) for conversational speech are approximately equal to 1.8 and 1.2 s, respectively. It is easily shown that the long-run fraction of time that each voice source is active, known as the speaker "activity fraction," is given in the models by the quantity $\lambda/(\alpha + \lambda)$. Thus, for normal conversational speech, the activity fraction is approximately 0.4. In all of the results presented here, the activity fraction was taken to be 0.45; hence, in an actual system operating with real voice signals one would expect to be able to multiplex a somewhat larger number of voice sources on a link of given capacity than is done in the parameter sets here.

Also, in all of the results, the packet generation rate, V, was taken to be 62.5 packets/s, which corresponds to a packet intergeneration time of 16 ms. Thus, for PSO we take $\lambda^{-1} = 1.65$ s, $\alpha^{-1} = 1.35$ s, $V = 62.5$ packets/s, and various values of N and C which yield the desired server utilization. In order to examine the sensitivity of the model results to magnitude of the mean active and inactive period durations, the parameters of PS1 were chosen to be $\lambda^{-1} = 0.825$ s, $\alpha^{-1} = 0.625$ s, and V, N, and C the same as for PS0, and the parameters of set 2 were chosen to be $\lambda^{-1} = 3.3$ s, $\alpha^{-1} = 2.7$ s, and V, N, and C the same as for set 0. Thus, if we call the sum of the mean active and inactive times the "average on/off cycle time," then the average on/off cycle time of PS1 is half that of PS0, while the average on/off cycle time of PS2 is twice that PS0.

The packet-generation process for the 1.5 s average on/off cycle time system is less bursty than that for the 3 s system, which is in turn less bursty than that for the 6 s system. An alternate way of varying the level of burstiness is to hold the average on/off cycle time constant while changing the packet-generation rate. That is, a system with an on/off cycle time of 3 s and a packet-generation rate of $2V$ is simply a time-scaled version of a system with an average on/off cycle time of 6 s and a packet-generation rate of V. Thus, PS2 will yield the same results as would be obtained by replacing the packet generation rate of set 0 by $2V$ while holding all other parameters constant.

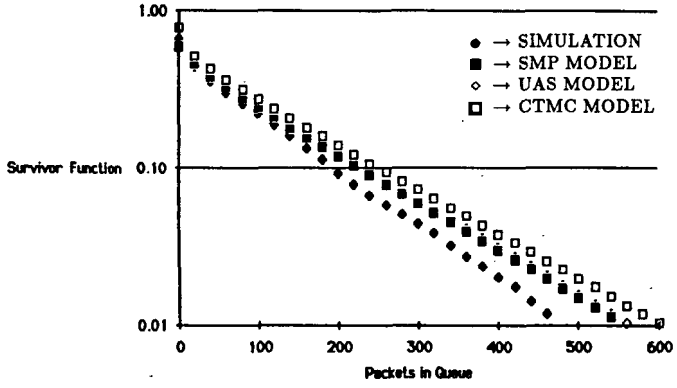


Fig. 6. Queue length survivor function for parameter set 0, $N = 8$, server utilization = 0.85.

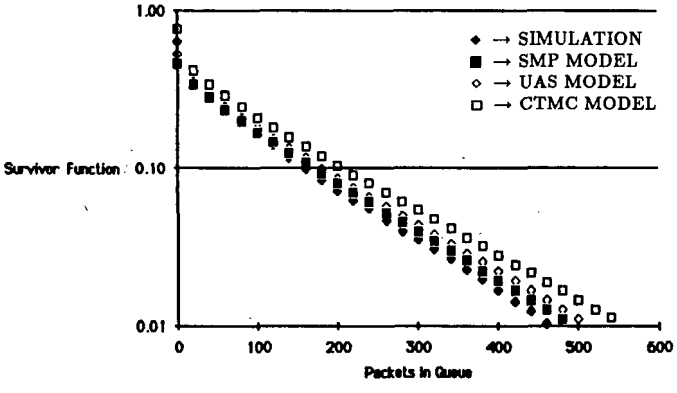


Fig. 7. Queue length survivor function for parameter set 0, $N = 15$, server utilization = 0.85.

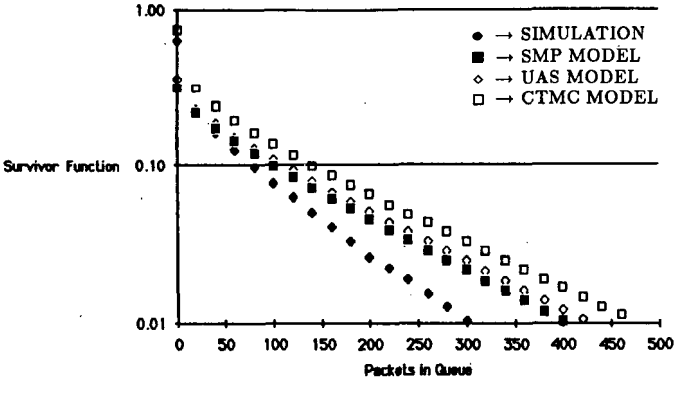


Fig. 8. Queue length survivor function for parameter set 0, $N = 30$, server utilization = 0.85.

Figs. 6–9 show the complementary cumulative distribution functions (or survivor functions) for parameter set 0 with various values of N and C chosen to produce a server utilization of 0.85. Fig. 10 shows the survivor functions for the same parameter set with values of N and C chosen to produce a server utilization of 0.7. These figures will now be discussed. Curves for the other parameter sets are deferred to [23] to conserve space, but results are discussed below as appropriate.

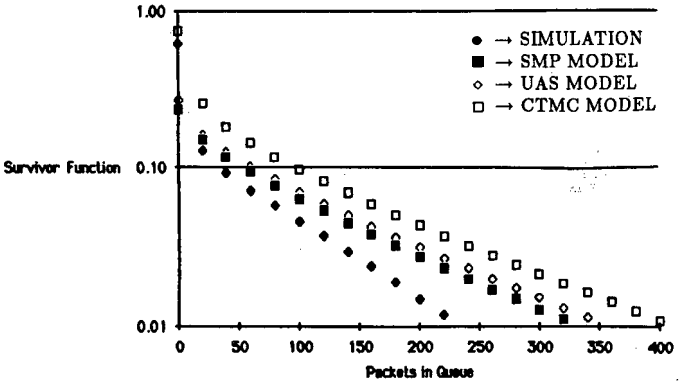


Fig. 9. Queue length survivor function for parameter set 0, $N = 45$, server utilization = 0.85.

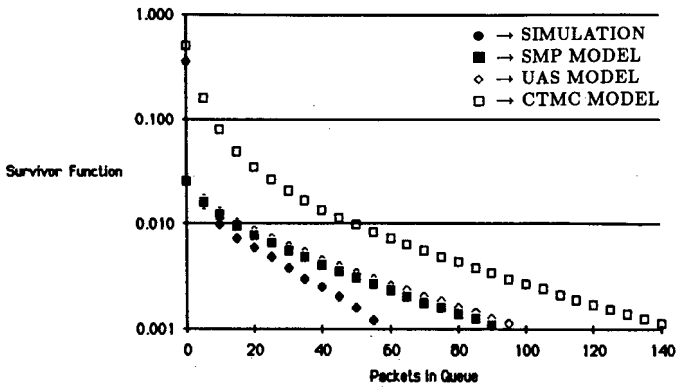


Fig. 10. Queue length survivor function for parameter set 0, $N = 30$, server utilization = 0.70.

B. Numerical Results

A brief examination of Figs. 6–9 reveals that the SMP, CTMC, and UAS models all predict longer queue lengths than the simulation. The SMP and UAS results are in most cases quite close to each other, with the SMP computed survivor function being slightly closer to the survivor function computed by the simulation. In general, the CTMC model predicts longer queue lengths than the other models; this is consistent with the remarks regarding the greater variability of the CTMC model dynamics presented in Section II.

Also, it is seen that both the SMP and UAS models clearly overestimate the probability that the queue is empty. This results from the fact that when the number of active voice sources is less than C neither model allows for an increase in queue length, whereas in the actual system the queue length will fluctuate around zero. As was noted in the discussions of the models in Section II, the number of active voice sources is likely to be less than C when the queue content is near zero.

A closer examination of Figs. 6–9, which correspond to PS0, reveals that the analytical model results tend to be closer to the simulation results for smaller values of N. For $N = 8$ (Fig. 6), the buffer occupancy which yields a survivor function value of 0.01 is approximately 550 packets for the SMP model and 480 for the simulation, a 15 percent difference. For $N = 15$, 30, and 45 (Figs. 7, 8, and 9), the differences between the SMP model and simulation buffer occupancies yielding a survivor function value of 0.01 are approximately 7, 33, and 43 percent, respectively. The corresponding numbers for UAS model are very close to these. The differences between the CTMC model and simulation buffer occupancies which yield the survivor function value of 0.01 for $N = 8$, 15, 30, and 45 are 25, 20, 40, and 75 percent, respectively.

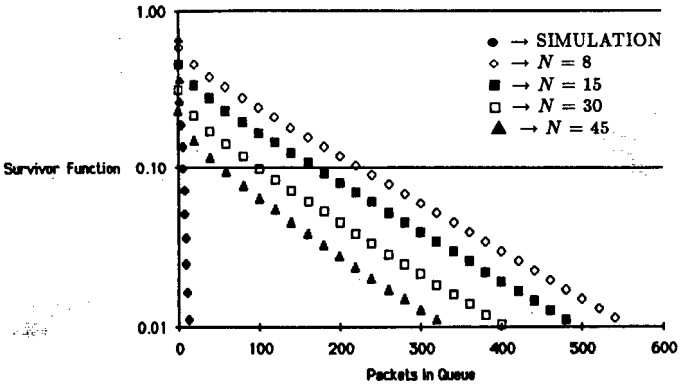


Fig. 11. Queue length survivor function for parameter set 0, server utilization = 0.85, SMP versus $M/D/1$.

A similar examination of the numerical results obtained for PS1 and PS2 shows that the same trend exists for parameter sets 1 and 2. Thus, it appears that the survivor functions predicted by the analytic models become less close to the simulation results as N increases.

Fig. 10 shows results for PS0 with $N = 30$ sources and C chosen to yield a server utilization of 0.7. From Fig. 10 it is seen that the analytic models again predict longer queue lengths than are found by the simulation. In particular, it appears that the CTMC model predicts the queue length survivor function markedly less well for the server utilization of 0.7 than it did for the server utilization of 0.85. Similar observations hold for PS1 and PS2 [23].

Fig. 11 shows the effect on the queue length distribution of changing the total number of voice sources and the link capacity while holding the remaining parameters constant. Here, α, λ, and V correspond to PS0, the parameter set representing typical packet voice operation; N and C are varied in proportion to maintain the server utilization at 0.85. From Fig. 11, it is seen that as N and C are increased, the queue lengths become shorter. This behavior is an instance of the well-known observation that "one fast server is better than two slow servers." That is, if there are ten voice sources that require service, then one fast link serving all ten provides better delay performance than two slower links each serving five sources. This efficiency results from the fact that in the two-server case, one server may be idle while the other server has a long queue, but in the single server case, full service effort is expended whenever there is any work to be done.

Fig. 11 also shows the $M/D/1$ queue length survivor function, the data for which was generated by a GPSSV simulation. It is seen that the approximate survivor func-

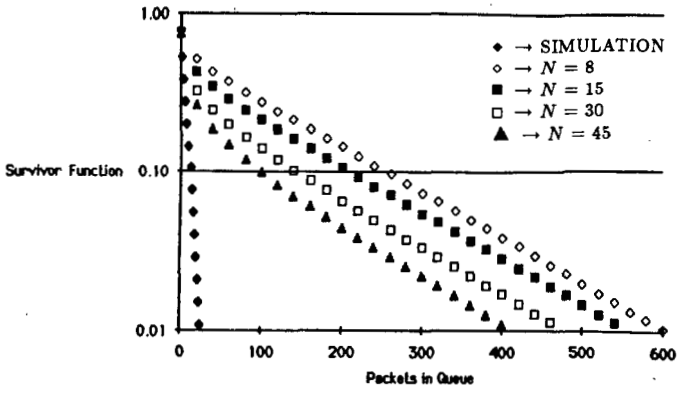


Fig. 12. Queue length survivor function for parameter set 0, server utilization = 0.85, CTMC versus *M/M/1*.

tions computed by the SMP program fall off much less rapidly than the *M/D/1* queue length survivor function. This is due to the correlation in the packet-generation process [12], [18], [19]. From Fig. 11, it appears that the performance of the SMP model approaches that of the *M/D/1* system for very large N and C. This appears to result from the fact that as the packet service time decreases with respect to the packet intergeneration time of a single source, the effect of correlation in the packet arrival process is reduced. Thus, as N and C are increased without bound, the effect of the correlation is completely lost and the performance approaches that of the *M/D/1* system [13]. The effect of correlation in the arrival process would, of course, be lost as C is increased even if N were not increased with C. In that case, the traffic intensity would go to zero, and the analysis would be pointless.

Fig. 12 depicts, in a fashion completely parallel to Fig. 11, the survivor function curves generated by the CTMC program and the survivor function for the *M/M/1* system. It is again observed that the *M/M/1* survivor function falls off much more rapidly than the CTMC model curves, but that as N and C are increased the performance of the CTMC model does tend toward that of the *M/M/1* system with the same server utilization. However, the convergence of the CTMC model performance to that of the *M/M/1* model is very slow.

We now turn to a discussion of the limiting models. Since the lengths of the active periods of each voice source are exponentially distributed, the packets from a given voice source can readily be shown to occur according to a general renewal process; this is also true of the packet arrivals from each source in the CTMC model. Now, it is well known that in the superposition of a large number of independent general renewal processes the distribution of the interevent times converges to the exponential distribution [20]. Thus it follows that the packet interarrival time distributions in the CTMC model and the simulation representing the real system converge to the exponential distribution. Since we have argued intuitively that, from a practical point of view, the effect of correlation in successive interarrival times also vanishes with increasing N and C, one would expect the packet arrival process to be approximated reasonably well by the Poisson process. Thus, one would expect the CTMC model and the SMP model, which tracks the simulation results reasonably well, to converge to the *M/M/1* and *M/D/1* models, respectively. The results of the above numerical investigations demonstrate that this convergence does not occur rapidly. A reasonably detailed discussion of this phenomenon is presented in the paper by Sriram and Whitt [24], which appears in this issue.

IV. Conclusions

Three models for assessing the queueing behavior of a packet voice statistical multiplexing system were described and evaluated. The solution techniques for all three models were readily implemented and no computational difficulties were encountered. By comparison to results from a GPSSV simulation model, it was shown that the queue length distributions obtained from the SMP and UAS models are reasonable for engineering purposes. Queue length distributions obtained from these models might be used, for example, to predict the buffer capacity needed to ensure that the fraction of packets lost due to buffer overflow is less than some maximum value. Of course, the analytic models are also useful as checks on the reasonableness of the simulation results.

It was also shown that the queue length survivor function for the *M/D/1* queueing system decreases at a much higher rate than that for the packet voice statistical multiplexing system when the number of voice sources is not large and the link capacity is low. As N and C are increased at the same rate while holding all other parameters fixed, the packet voice survivor function begins to approach that of the *M/D/1* system, but the convergence is very slow. Reasons for this are suggested at the end of Section IV. Based on the results of our numerical work, it does not appear that the *M/D/1* model would provide useful results unless the number of voice sources is large, say hundreds.

The computational effort required to solve the SMP and CTMC models and to perform the simulations is substantial. The UAS model requires negligible computational effort, but there is some reduction in the quality of results compared to the SMP model. While both the SMP model and the UAS model provide results which appear to be useful, better models are needed to predict performance accurately.

References

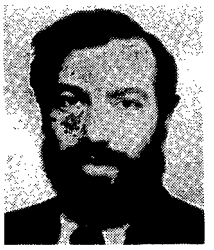
[1] L. Kleinrock, "Principles and lessons in packet communications," *Proc. IEEE* (Special Issue on Packet Communication Networks), vol. 66, pp. 1320–1329, Nov. 1978.

[2] L. G. Roberts, "The evolution of packet switching," *Proc. IEEE* (Special Issue on Packet Communication Networks), vol. 66, pp. 1307–1313.

[3] E. Arthurs and B. W. Stuck, "A theoretical traffic performance analysis of an integrated voice-data virtual circuit packet switch," *IEEE Trans. Commun.* (Special Issue on Digital Switching), vol. COM-27, pp. 1104–1111, July 1979.

[4] W. A. Montgomery, "Techniques for packet voice synchronization," *IEEE J. Select. Areas Commun.* (Special Issue on Packet Switched Voice and Data Communication), vol. SAC-1, pp. 1022–1028, Dec. 1983.

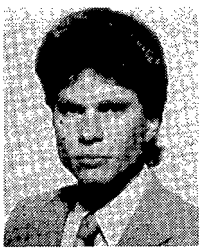
[5] P. T. Brady, "A model for generating on-off speech patterns in two-way conversations," *Bell Syst. Tech. J.*, vol. 48, pp. 2445-2472, Sept. 1969.
[6] C. J. Weinstein, "Fractional speech loss and talker activity model for TASI and for packet-switched speech," *IEEE Trans. Commun.*, vol. COM-26, pp. 1253-1257, Sept. 1978.
[7] J. N. Daigle and J. D. Langford, "Operations research methods in the communications fields," Univ. of Rochester, Rochester, NY, Working Paper QM 8502, Jan. 1982.
[8] T. E. Stern, "A queueing analysis of packet voice," in *Conf. Rec. 1983 IEEE Global Telecommun. Conf.*, vol. 1, San Diego, CA, 1983, pp. 71-76.
[9] M. F. Neuts, *Matrix Geometric Solutions in Stochastic Models*. Baltimore, MD: The John Hopkins University Press, 1981.
[10] R. J. Pilc, "A derivation of buffer occupancy statistics in an asynchronous time-division multiplexor used with bursty sources," Bell Labs Tech. Memo. (internal report), Dec. 1968.
[11] D. Anick, D. Mitra, and M. Sondhi, "Stochastic theory of a data-handling system with multiple sources," in *Conf. Rec. 1980 Int. Conf. Commun.*, vol. 1, Seattle, WA, 1980, pp. 13.1.1-13.1.5.
[12] D. Anick, D. Mitra, and M. Sondhi, "Stochastic theory of a data handling system with multiple sources," *Bell Syst. Tech. J.*, vol. 61, no. 8, pp. 1871-1894.
[13] J. N. Daigle and J. D. Langford, "Queueing analysis of a packet voice communication system," *Conf. Rec. IEEE INFOCOM'85*, Washington, DC, Mar. 1985, pp. 18-26.
[14] J. Kohlas, *Stochastic Methods of Operations Research*. Cambridge, MA: Cambridge University Press, 1982.
[15] S. M. Ross, *Introduction to Probability Models*. New York: Academic, 1980.
[16] J. D. Langford, "Queueing analysis of a packet voice communication system via a semi-Markov process," Masters' thesis, Clemson University, Clemson, SC, 1984.
[17] P. A. Bobillier, B. C. Kahan, and A. R. Probst, *Simulation with GPSS and GPSS V*. Englewood Cliffs, NJ: Prentice-Hall, 1976.
[18] J. N. Daigle, "Queueing analysis of a packet switching node with Markov renewal arrival process," *Conf. Rec. 1977 Int. Conf. Commun.*, vol. 1, Chicago, IL, 1977, pp. 279-283.
[19] L. Kleinrock, *Queueing Systems—Vol. I: Theory*. New York; Wiley, 1975.
[20] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. II. New York: Wiley, 1966.
[21] G. F. Williams and A. Leon-Garcia, "Performance analysis of integrated voice and data hybrid-switched links," *IEEE Trans. Commun.*, vol. COM-32, pp. 695-706, June 1984.
[22] J. N. Daigle, "Queueing analysis of a packet switching node in a data communication system," Ph.D. dissertation, Columbia University, New York, NY, 1977.
[23] J. N. Daigle and J. D. Langford, "Models for analysis of packet voice communication systems," Univ. Rochester, Rochester, NY, Working Paper QM 8542 Oct. 1985.
[24] K. Sriram and W. Whitt, "Characterizing superposition arrival processes in packet multiplexers for voice and data," this issue.



John N. Daigle (M'68-SM'84) received the B.S.E.E. degree from Louisiana Tech University, Ruston, LA, in 1968, the M.S. degree in electrical engineering from Virginia Polytechnic Institute and State University, Blacksburg, VA, in 1969, and the Eng.Sc.D. in operations research from Columbia University, New York, NY, in 1977.

He is currently Associate Professor and Area Coordinator of Computer and Information Systems at the Graduate School of Management, University of Rochester, Rochester, NY, and is a consultant to GTE Laboratories. He has research interests in design and analysis of computer communication networks and information systems. He has extensive practical experience in mathematical and simulation analysis of many aspects of computer communication network design, including operations, and has authored more than 40 articles and reports in these areas. Prior to his current appointment, he held appointments in the electrical and computer engineering departments of Clemson University (from 1982 to 1984) and Washington State University (from 1980 to 1982). He led the systems analysis group at the Communication Systems Division of the NCR Corporation from 1977 to 1980 and was a member of Technical Staff at Bell Laboratories from 1972 to 1977. He has served on the conference board of IEEE INFOCOM and as general chairman of INFOCOM'85. He is a past chairman of the IEEE Communication Society's Technical Committee on Computer Communications, and he has served on the technical program committees of numerous IEEE conferences.

Mr. Daigle is a member of the Eta Kappa Nu, Omega Rho, and Sigma Xi.



Joseph D. Langford (S'82) received the B.S.E.E. degree from Washington State University, Pullman, WA, in 1983, and the M.S.E.E. degree from Clemson University, Clemson, SC, in 1984.

During 1983, he was with Harris Corporation, Melbourne, FL, where he worked on performance analysis of satellite-based naval communication systems. His current interests are in the general area of computer networking with emphasis on design and analysis of information systems. He is currently pursuing the Ph.D. degree in that area at the University of Rochester. He was awarded a three-year National Science Foundation Graduate Fellowship in 1983.

Mr. Langford is a member of Tau Beta Pi.