Approximation Techniques for Computing Packet Loss in Finite-Buffered Voice Multiplexers [1]

Ramesh Nagarajan[2], James F. Kurose and Don Towsley[3]

Abstract

In this paper we examine three different approximation techniques for modeling packet loss in finite-buffer voice multiplexers. The performance models studied differ primarily in the manner in which the superposition of the voice sources (i.e., the arrival process) is modeled. The first approach models the superimposed voice sources as a renewal process and performance calculations are based only on the first two moments of the renewal process. The second approach is based on modeling the superimposed voice sources as a Markov Modulated Poisson Process (MMPP). Our choice of parameters for the MMPP attempts to capture aspects of the arrival process in an alternate, more intuitive, manner than previously proposed approaches for determining the MMPP parameters and is shown to compute loss more accurately. Finally, we also evaluate a fluid flow approximation for computing packet loss. For all three approaches, we consider as a unifying example, the case of multiplexing voice sources over a T1-rate link. The main conclusion of this paper is that both the new MMPP model examined here and the fluid flow approximation can provide accurate loss predictions for parameter ranges of practical interest; these predictions are also shown to be *many orders of magnitude* better than modeling the superimposed voice sources simply as a Poisson process. We also discuss the problem of modeling buffer overflow for general arrival processes and also consider modeling approaches for analyzing finite-buffer multiplexers with general arrival and service processes in a network environment.

[1]This work was supported in part by the Office of Naval Research under contract N00014-87-K-304, the Defense Advanced Research Projects Agency under contract NAG2-578 and an NSF equipment grant, CERDCR 8500332.

[2]Dept. of Elect. and Comp. Engg., UMASS, Amherst.

[3]COINS Dept., UMASS, Amherst.

1 Introduction

Traditionally, computer communication networks have primarily carried data traffic and have been engineered to meet the attendant performance requirements of this class of traffic. More recently, however, there has been interest in supporting *real-time* communication applications such as control, command, and interactive voice and video applications in a packet-switched environment. Such real-time traffic differs from traditional data traffic in several important ways. Most importantly, real-time traffic is delay sensitive (loss insensitive) while data traffic is loss sensitive (delay insensitive). Hence it is natural to engineer communication networks, that support real-time traffic, so that delays are bounded at the expense of some loss. However, the magnitude of this loss determines the quality of service and hence it is critical to predict this loss accurately in order to provide an acceptable grade of service.

In this paper we study packet loss in a finite-buffered statistical multiplexer for superimposed packetized voice sources. Given the fixed length packets typical of voice sources and FCFS service at a multiplexer, imposing a buffer size of K is essentially equivalent to imposing a time constraint of Kd, where d is the fixed transmission time of a packet [25]. In both cases, a packet is lost (and hence not transmitted) if its delay would exceed Kd. The performance of a voice multiplexer under finite buffer assumptions is thus of considerable interest since (from a practical standpoint) there are no truly infinite buffers and (from an applications standpoint) finite buffers can be used to bound queueing delays for time-constrained, loss-tolerant traffic.

We will study three different *approximate* approaches for computing loss in finite-buffered voice multiplexers. Our work differs from [25] in that our modeling approaches are approximate rather than exact. We can thus consider large systems without needing to explicitly solve for the joint distribution of the number of active sources and the number of queued packets. (In some of the examples we consider, for example, a transition matrix with over 10^7 entries would have to be considered in an exact analysis). Further, the solution technique in [25] is tailored to the voice source model and hence has a more limited application as a general model for bursty sources. Our first approach is similar in spirit to [10, 24], in that the superposition of voice sources is modeled as a renewal process and performance calculations are based only on the first two moments of the renewal process. Our work differs from the general approach of [10, 24] (see also [21]) in that we introduce an additional heuristic needed to handle the case of finite buffers; we examine several possible ways in which this approximation can be done. Our second approach is based on modeling the superimposed voice sources as a Markov Modulated Poisson processes (MMPP), a technique used successfully in [7] to compute *average delay* through an infinite buffer voice multiplexer. It is suggested in [7] that the analysis of finite-buffered multiplexers can be handled in a similar fashion. We use the MMPP model to approximate the superposition but present a simpler analytical approach to compute packet loss and show that the MMPP parameters which successfully approximated average delay in [7] do not work as well in the case of analyzing packet loss; we derive an alternate set of MMPP parameters which is shown to compute loss more accurately. Recent work by Liao and Mason [12] has shown that the parameter set proposed in [7] is inadequate for accurately predicting even the

average delay when the MMPP is used to model a superposition of sources that are more bursty than the voice source i.e., the sources have either a higher peak rate in talkspurts or have longer burst lengths. In this context, [12] independently reports success in accurately predicting the average delay but not loss using a parameter set that is identical in spirit to ours. Finally, we evaluate a fluid flow approximation for computing packet loss based on the technique in [1]. For all three approaches, we consider as a unifying example, the case of multiplexing voice sources over a T1-rate link. The main conclusion of this paper is that both the new MMPP model examined here and the fluid flow approximation can provide accurate loss predictions for parameter ranges of practical interest; these predictions are also shown to be many *orders of magnitude* better than modeling the superimposed voice sources simply as a Poisson process. We also discuss the applicability of our techniques to alternate source traffic models and for the case of a multiplexer operating in a *network* environment (in which case the inter-nodal network flows must also be considered). In the latter case the renewal process approximation is used to model the network flows. The network setting considered is a Homogenous network with error control. The error control protocol that we examine is the end-to-end scheme as outlined in [3]. This error control protocol is designed to recover from packet loss due to errors and buffer overflows by retransmission of packets from source to destination (end-to-end basis). In our analysis we consider, for simplicity, packet loss due to buffer overflows only.

The remainder of this paper is structured as follows. In section 2 we describe the model which characterizes the behavior of a single packet voice source. The two-moment approximation, MMPP, and fluid flow models are then discussed in sections 3, 4 and 5, respectively. Section 6 discusses the results and section 7 concludes this paper. Much of the mathematical detail of the models has been relegated to the appendices.

2 The Voice Source Model

The model we assume for a single voice source is a standard one (see, e.g., [5, 7, 21]) whose basic premise is that an active voice source periodically generates fixed length packets when a speaker is speaking (talkspurt) and otherwise remains idle. We briefly describe this model here; the reader is referred to the above references, in particular [21], for additional details and discussion. The voice packetization period is assumed to be 16 msec. and the talkspurt is assumed to contain a geometrically distributed number of packets, with the mean number of packets in a talkspurt being 22. The mean length of a talkspurt is thus 352 ms. The period between talkspurts, known as the silence period and denoted by X, is assumed to be exponentially distributed with a mean length of 650 ms. The speech activity ratio is thus 0.351 and each source generates on the average 22 packets every second.

Given the above model, the interarrival times between packets generated by a *single* source form a renewal process. With probability $\frac{1}{22}$, the interarrival time is 16 ms. and with probability $\frac{21}{22}$, the interarrival time is $16 + X$ ms. [21]. We note that this arrival process is quite ***bursty***; the

squared coefficient of variation of the interarrival time is 18.1, while it is 1.0 for a Poisson process.

The input traffic to the multiplexer is taken to be the superposition of a finite population of voice sources (say M), each of which is characterized as above. In our loss calculations in the following sections, we will assume 64 byte packets [21] and that the voice sources are being multiplexed over a T1-rate (1.536 Mbits/sec.) link. A line utilization of 100 percent would then result from a superposition of approximately 136 voice sources.

3 A Renewal Process Approach

Since the superposition of a number of renewal processes generally does not result in a renewal process, our first approach to modeling the finite-buffer voice multiplexer will be to approximate the superimposed voice arrival process by a single renewal process, i.e., we will approximate the interarrival times of packets to the multiplexer as i.i.d random variables. Fixed length packets and a finite number of buffers, K, will be assumed. Our goal is thus to approximate packet loss in a $GI/D/1/K$ system.

Let us define N_∞ to be the number of packets queued in the *infinite* buffer $GI/D/1/\infty$ queue with the above interarrival and service time distributions and N_K as the number of packets queued in the corresponding finite-buffer $GI/D/1/K$ queue. There are two parts to the renewal approximation.

- **Calculate the approximate occupancy distribution, $P(N_\infty = i)$, for an *infinite* buffered voice multiplexer.** To do this, we will need to know:

  λ : the aggregate arrival rate of the superimposed voice sources. This is simply equal to the sum of the average packet generation rates of the individual voice sources.

  c_a^2 **:** the squared coefficient of variation of the renewal interval (interarrival time) of the approximating arrival process. We discuss c_a^2 below.

  μ^{-1}, c_s^2 **:** the first moment and squared coefficient of variation of the service time distribution. Given fixed packet lengths, $c_s^2 = 0$.

  Given these four parameters, the first and second moments of the occupancy distribution for the infinite buffer queue are then computed using approximate formulae from the literature [24]. A discrete or a continuous distribution can then be chosen which matches the computed moments and used to approximate $P(N_\infty = i)$.

- **Calculate the approximate occupancy distribution, $P(N_K = i)$, for the *finite* buffered voice multiplexer with K buffers.** Given the distribution $P(N_\infty = i)$, the probability of packet loss, $P(N_K = K)$, for a multiplexer with K buffers is approximated by,

$$P(N_K = K) = \frac{P(N_\infty = K)}{P(N_\infty \leq K)} \tag{1}$$

Here $P(N_K = K)$ provides our first approximation of the loss probability for the finite-buffered packetized voice multiplexer. A similar result holds for the occupancy probabilities at departure instants for the $M/G/1/K$ queue and the corresponding $M/G/1/\infty$ queue [2].

We now describe the above approach in more detail. The squared coefficient of variation of the arrival process is approximated using a heuristic based on the so-called asymptotic approximation method and the stationary interval method (refer to appendix A.1).

$$c_a^2 = wc_1^2 + (1 - w) \tag{2}$$

where c_1^2 is the squared coefficient of the interarrival time in a single voice source, $w = [1 + 4(1 - \rho)^2(M - 1)]^{-1}$ and $\rho = \lambda/\mu$.

Given the mean and squared coefficients of variation of the interarrival and service times, we can now use approximation formulae for the $GI/G/1/\infty$ queue to obtain the following expressions for the average number of customers queued, $E[N_\infty]$, and $E[N_\infty^2]$, the second moment of the number of queued customers (via equations (47) and (66) in [24]):

$$E[N_\infty] = \rho + \frac{\rho^2(c_a^2 + c_s^2)g}{2(1 - \rho)} \tag{3}$$

$$E[N_\infty^2] = E^2[N_\infty](c_N^2 + 1) \tag{4}$$

where g, a bias factor, and c_N^2 are given by equations (45) and (65) in [24]. We also note that the probability that the server is busy at an arbitrary point in time, $P(N_\infty > 0)$, is given by:

$$P(N_\infty > 0) = \rho \tag{5}$$

and that this relationship is exact even for stationary non-renewal arrival processes [24].

Although the first and second moments of the occupancy distribution can now be computed via equations (3) and (4), the *distribution itself* is needed in order to use the conditioning heuristic of equation (1) to compute packet loss. We approximate this distribution by choosing a distribution with a known form whose first and second moments match those values computed in equations (3) and (4) above. Equation (5) provides another measure that can be used to match the distribution to analytically computed values. Since equations (3) and (4) together imply that the squared coefficient of variation of the number of queued customers is greater than one for the range of input traffic loads of interest, a mixture of geometric distributions seems natural. The mixture of two geometric distributions shown in Figure 1 has three unknowns: p, p_1, and p_2. Let $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$. If N is the random variable whose distribution is given by the mixture shown in Figure 1, then the distribution, mean, second moment, variance and probability that $N > 0$ are given by:

$$P(N = k) \quad = \quad pp_1q_1^k + (1 - p)p_2q_2^k \tag{6}$$

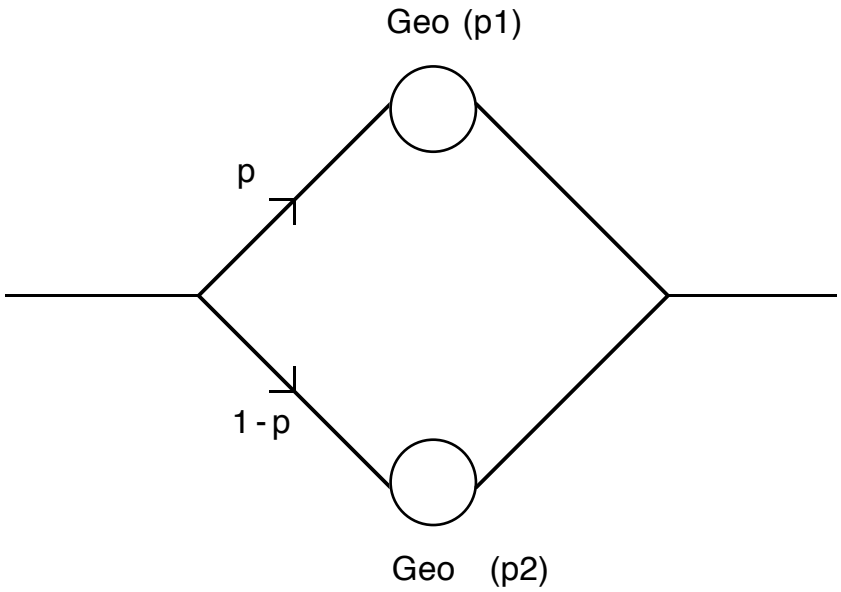

Figure 1: Fitting a Discrete Distribution

$$E[N] = \frac{pq_1}{p_1} + \frac{(1-p)q_2}{p_2} \tag{7}$$

$$E[N^2] = p(\frac{q_1}{p_1^2} + \frac{q_1^2}{p_1^2}) + (1-p)(\frac{q_2}{p_2^2} + \frac{q_2^2}{p_2^2}) \tag{8}$$

$$\sigma_N^2 = E[N^2] - E^2[N] \tag{9}$$

$$P(N > 0) = 1 - P(N = 0)$$

$$= 1 - pp_1 - (1-p)p_2 \tag{10}$$

Numerical values for $E[N], E[N^2], \sigma_N^2$ and $P(N > 0)$ are calculated for the infinite buffer system using equations (3), (4) and (5). Equations (7), (8) and (10) then provide three equations in three unknowns (p, p_1, and p_2). Attempts to solve these nonlinear equations simultaneously using standard software packages proved difficult as we encountered problems of convergence and difficulty in the choice of an appropriate set of initial parameters. Hence, we introduce a new parameter, θ, and reduce the number of parameters to two by using one of the following set of expressions,

$$p_1 = 1/2p\theta, p_2 = 1/2(1-p)\theta \tag{11}$$

or

$$q_1/p_1 = 1/2p\theta, q_2/p_2 = 1/2(1-p)\theta. \tag{12}$$

Note that equation (12) reduces the system to a mixture of geometric distributions with balanced means. Given these transformations, the unknowns are now p and θ, and there are three equations in these two unknowns. The following pairs of the equations (7), (8) and (10) were considered for determining p, p_1 and p_2.

1. Equations (7) and (10) were used to solve for p and θ as shown below, and p_1 and p_2 were then determined using equation (11).

$$\theta = 1/(1 - P(N > 0))$$

$$p = \frac{1 \pm \sqrt{1 + (1 + E[N] - 2\theta)/\theta}}{2}$$

2. The equations used were the same as above but equation (12) was used to reduce the system to two unknowns. The values of the parameters of the system are obtained as

$$\theta = 1/E[N]$$

$$p = \frac{1 \pm \sqrt{1 + \frac{1-P(N>0)(1+2\theta)}{\theta(1-P(N>0)\theta)}}}{2}$$

3. Here we use equations (7), (8) and and (12). Solving for θ and p we obtain,

$$\theta = \frac{1}{E[N]} \tag{13}$$

and p as the solution of a quadratic equation shown below. p_1 and p_2 were then determined using equation (12).

$$p = \frac{1 \pm \sqrt{1 - \frac{2}{\theta(\theta E[N^2]-1)}}}{2} \tag{14}$$

It is important to note that these heuristics are not guaranteed to yield a fit. This is due to the fact that the heuristics require solutions of quadratic equations which yield real values for the roots only if the discriminant is positive in value. Moreover, in some cases, the solutions did not correspond to valid probabilities i.e., values in the interval [0,1]. Heuristic (3) was successful when the number of external voice sources was greater than 120, while heuristic (2) was successful in the lower traffic ranges.

Given p, p_1 and p_2, the distribution of the number in the system, N_∞, can then be computed via equation (6). The probability of packet loss in the finite-buffered statistical multiplexer can then be computed using our conditioning heuristic in equation (1).

We also fit a continuous distribution, $F(.)$, to the moments and the individual probabilities were obtained as

$$P(N_\infty = i) = F(i + 0.5) - F(i - 0.5) \tag{15}$$

$$F \sim \textit{Continuous distribution fit for the number in the system}$$

This was then used in the conditioning heuristic 1 to obtain the loss probabilities. The continuous distribution $F(.)$ was determined as in [24, eqs. 55-61].

Case 1:$c_N^2 > 1.01$

$$Prob(N > x) = p(e^{-\gamma_1 x} - e^{-\gamma_2 x})$$

$$where\ p = (1 + \sqrt{\frac{c_N^2 - 1}{c_N^2 + 1}})/2$$

$$and\ \gamma_1 = 2p/EN, \gamma_2 = 2(1-p)/EN$$

Case 2:$0.99 \leq c_N^2 \leq 1.01$

$$Prob(N > x) = e^{-x/EN}$$

Case 3:$0.501 \leq c_N^2 < 0.99$

$$Prob(N > x) = \frac{(\gamma_1 e^{-\gamma_2 x} - \gamma_2 e^{-\gamma_1 x})}{(\gamma_1 - \gamma_2)}$$

$$where\ \gamma_1^{-1} = EN - \gamma_2^{-1}$$

$$and\ \gamma_2^{-1} = \frac{EN + \sqrt{2VAR(N) - (EN)^2}}{2}$$

Case 4:$c_N^2 < 0.501$

$$Prob(N > x) = e^{-\gamma x}(1 + \gamma x)$$

$$where\ \gamma = 2/EN$$

where $F(x) = 1 - P(N > x)$.

Figure 2 plots the computed loss probabilities as a function of the traffic intensity (number of voice sources). Loss probabilities were estimated by simulation where each estimate was based on simulation over approximately one million packets. The Occupancy Normalization (discrete fit) plots the results obtained from the analysis above. The results of the continuous distribution fitting are also shown in Figure 2 and are seen to be similar to the results obtained by fitting a discrete distribution. Also shown in Figure 2 are the loss probabilities resulting from a Poisson model of the superimposed voice sources. Note that the values based on the Poisson assumptions underestimate the loss by several orders of magnitude, thus highlighting the difficulties (previously noted by others, e.g., [21]) in modeling superimposed voice sources by a Poisson process.

We note that the loss probabilities obtained by our approximations are consistently lower than the simulation values. To understand the source of the discrepancy between the analysis and the simulation results, we carried out a simulation to estimate the variance of the system occupancy distribution. The point estimates in the simulation were obtained by independent replications and the 95% confidence intervals were obtained by assuming a student-t distribution and using the jackknifing technique [11]. Table ?? shows that the estimates were fairly accurate for traffic intensities

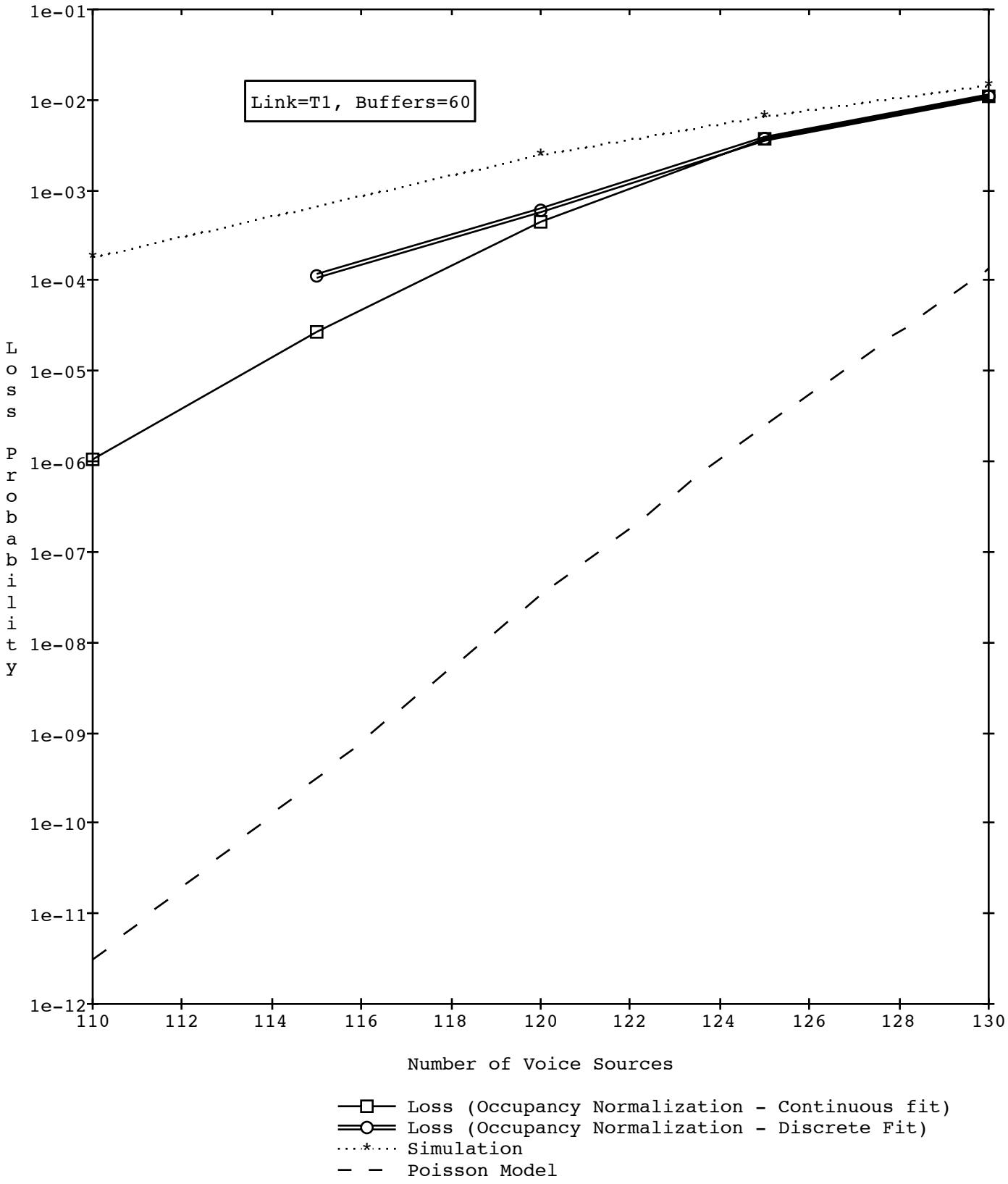

Figure 2: Loss in a Packet Voice Multiplexer (Renewal Model)

Sources	VAR(N)	
	Analysis	Simulation
110	1.67e+01	4.86e+01± 0.91e+01
125	1.27e+02	9.38e+02± 0.74e+02
130	0.67e+04	2.26e+04± 0.36e+04

Table 1: Variance of occupancy distribution for an infinite buffer multiplexer

approaching one (as expected, since a number of the approximations in [24] are asymptotically correct in heavy traffic) but were underestimated in other cases. Similar observations have been made in [21] but their discussion was limited to the first moments. We conjecture that these inaccuracies (which appear in our approximation technique even before the distribution fitting approximation is performed) may have contributed to the deteriorating accuracy of the approximation as the traffic intensity decreases.

3.1 Performance of Multiplexer in Network Environment

In this section we extend the renewal process approximation technique to communication networks; that is to the case of a statistical multiplexer in a *network* setting. We look at the accuracy of this technique in a Homogenous network with end to end error control and shared buffer pools [3]. For the communication network, we assume a Virtual Circuit environment as in [3]. We look at Homogenous networks for the case that the number of hops in the Virtual Circuit are two and three. The analytical model of a single node under this scheme is shown in Figure 3.

The arrival process at this node is the superposition of external traffic and departure processes from other similar nodes. Given the homogeneity assumption, these departure processes are the same as the departure processes from the node under consideration and this forms the basis of the iteration scheme described below. The iteration process is graphically illustrated in Figure 3 for three hop Virtual Circuits. The calculations and the values of the statistical characteristics of the processes shown in the figure are enumerated here.

- Initial values for the iteration scheme :

$$\lambda_1 = \lambda_{av} M$$

$$where\ \lambda_{av} = 1/(T + \alpha T/\beta)$$

$$with\ \alpha^{-1} = 352msec, \beta^{-1} = 650msec\ and\ T = 16msec\ [21]$$

$$and\ M = Number\ of\ external\ Voice\ Sources$$

$$c_1^2 = 18.1\ (Asymptotic\ Approximation)$$

$$\lambda_2 = \lambda_3 = 0.0$$

$$c_2^2 = c_3^2 = 1.0$$

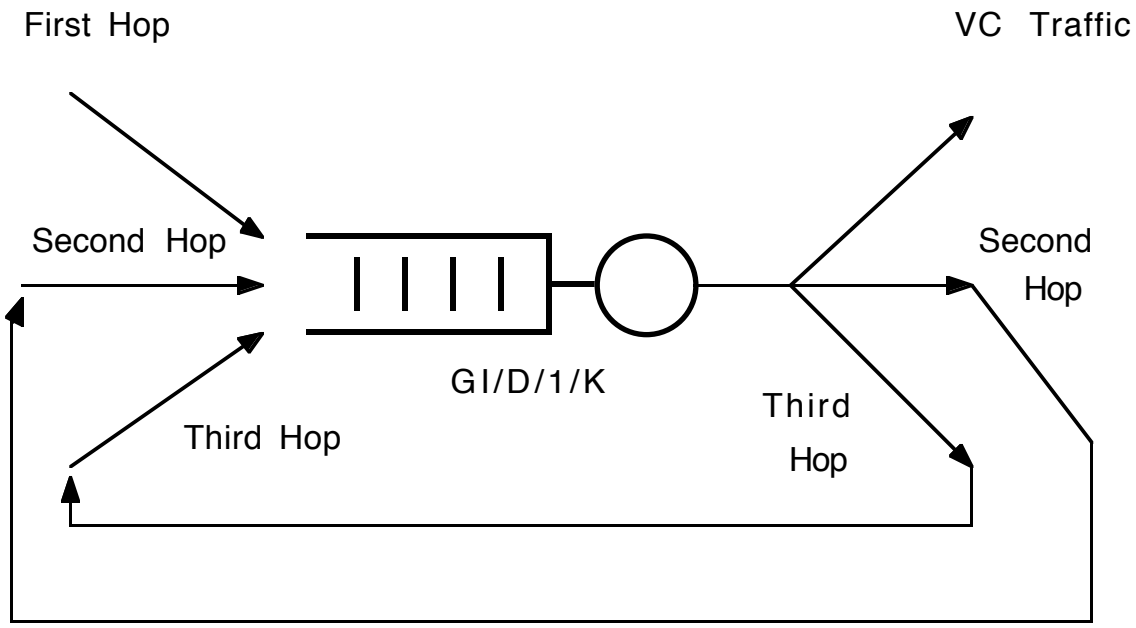

Figure 3: A single node in isolation (Iterative scheme)

- Iteration process :

$$\lambda_{sup} = \lambda_1 + \lambda_2 + \lambda_3$$

1. Compute c^2_{sup} =Superposition operation applied to λ_1, λ_2 and λ_3 as outlined in section A.1.

2. Estimate loss by the Occupancy Normalization Approximation (continuous fit).

3. Split λ_{sup} to account for loss i.e.,

$$\lambda_{sup} = \lambda_{sup}(1 - loss)$$

4. Obtain c^2_d from the input process and queue characteristics as outlined in section A.3.

5. Compute

$$\lambda_2 = \lambda_1(1 - Loss)\ and\ \lambda_3 = \lambda_2(1 - Loss)$$

$$\lambda_{vc} = \lambda_3(1 - Loss)$$

$$\lambda_1 = \lambda_{vc}/(1 - Loss)^3$$

6. The squared coefficients of variation of the split departure processes are obtained from c_d^2 and the splitting probabilities as in section A.2.

The iteration process is terminated when both c_{sup}^2 and λ_{sup} as obtained from successive iterations are within 10^{-4} of each other or the number of iterations exceeds 500. In instances wherein the number of iterations exceeded 500, it was found that both c_{sup}^2 and λ_{sup} were within about 10^{-3} of each other and iteration process could be halted. The performance measures of the node such as blocking probability and average delays can then be obtained and these are identical for all nodes along the VC [3]. The end to end performance can then be obtained easily.

Figure 4 shows the results of the analysis and simulation for the two hop Homogenous network. It can be seen that the analytical results are good indicators of the average loss probabilities for the range of parameters of interest[4] (although this range is not fully explored in the figure, it is clear that the analytical results appear to be within an order of magnitude of the simulation values for this range). The figure also shows the poor performance of the Poisson assumption for the network traffic.

The results for the three hop network are shown in Figure 5. Here again we see order of magnitude correspondence between the simulation and analysis in the range of interest. The parameters used for the simulation and the analysis are as in [7] and are also shown in the figures. The simulation was carried out for about 35 minutes of operating time which amounts to about 50 million packets arriving into the network. The specific values depend on the parameters of the simulation which include the number of hops and the number of external voice sources. In the simulation, the incoming packet to each node is branched with equal probability to one of the output links. The input traffic to the network reaches its destination after it traverses a number of nodes equal to the number of hops in the Virtual Circuit.

We note that the asymptotic approximation has been used for the superposition of the external voice sources and this is evident in the higher error probabilities as compared to the simulation results. It has been noted in [21] that the approximation for the superposition of renewal processes is more accurate if the processes being superposed are of the same rate. The asymptotic approximation for the external traffic enables us to consider it as a single stream whose rate is comparable to that of the other internal network traffic. As specified earlier, we reiterate here that our Occupancy Normalization Approximation in the network environment employs the continuous fit to obtain the occupancy distribution.

4 An Approach Based on Markov Modulated Poisson Processes

Our second approximation technique for computing loss models the superposition of voice sources as a correlated non-renewal process known as a Markov Modulated Poisson process (MMPP). This technique was used in [7] to successfully model average delay of voice packets through an

[4] 1 to 10% packet loss

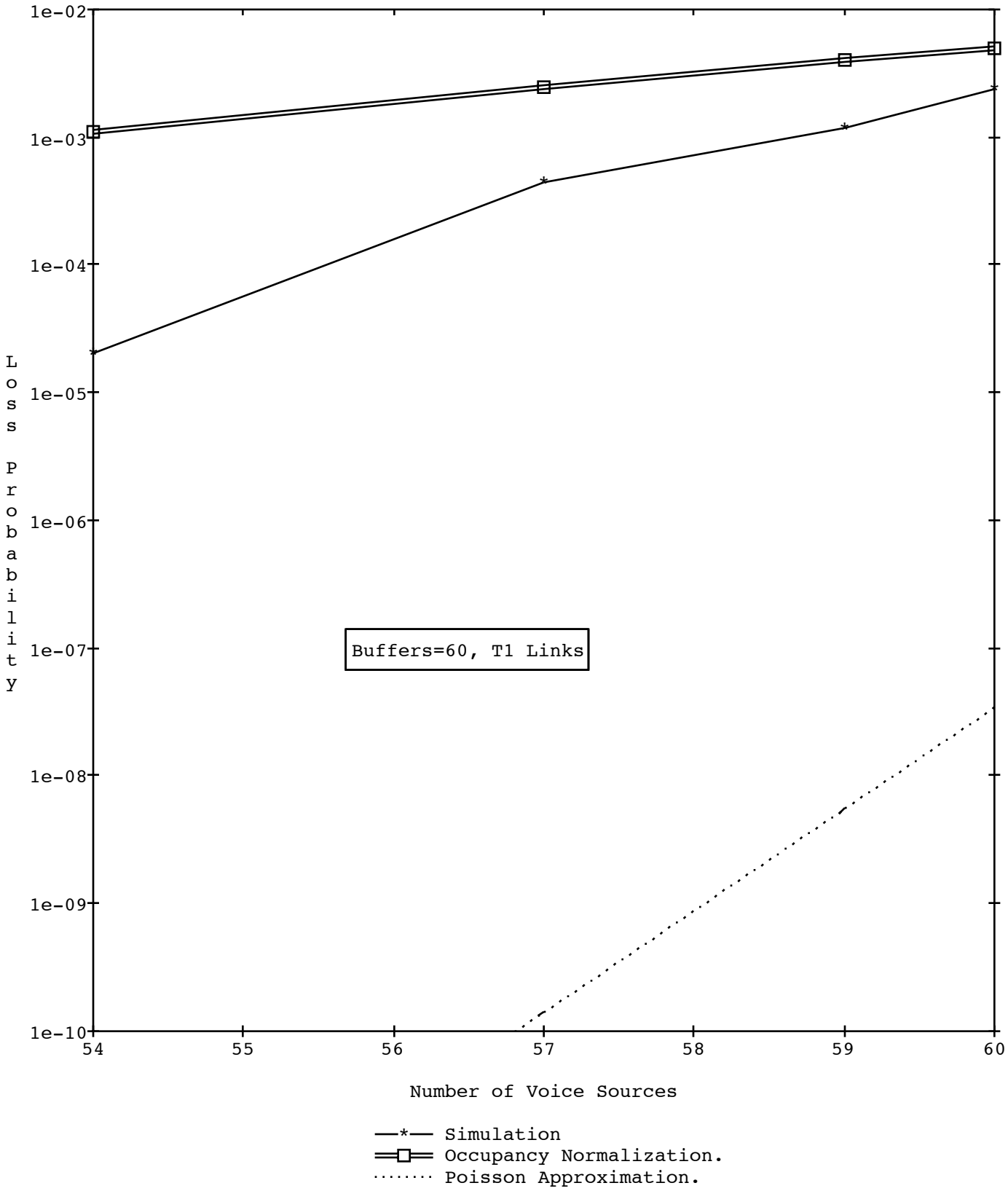

Figure 4: Loss in a two hop network

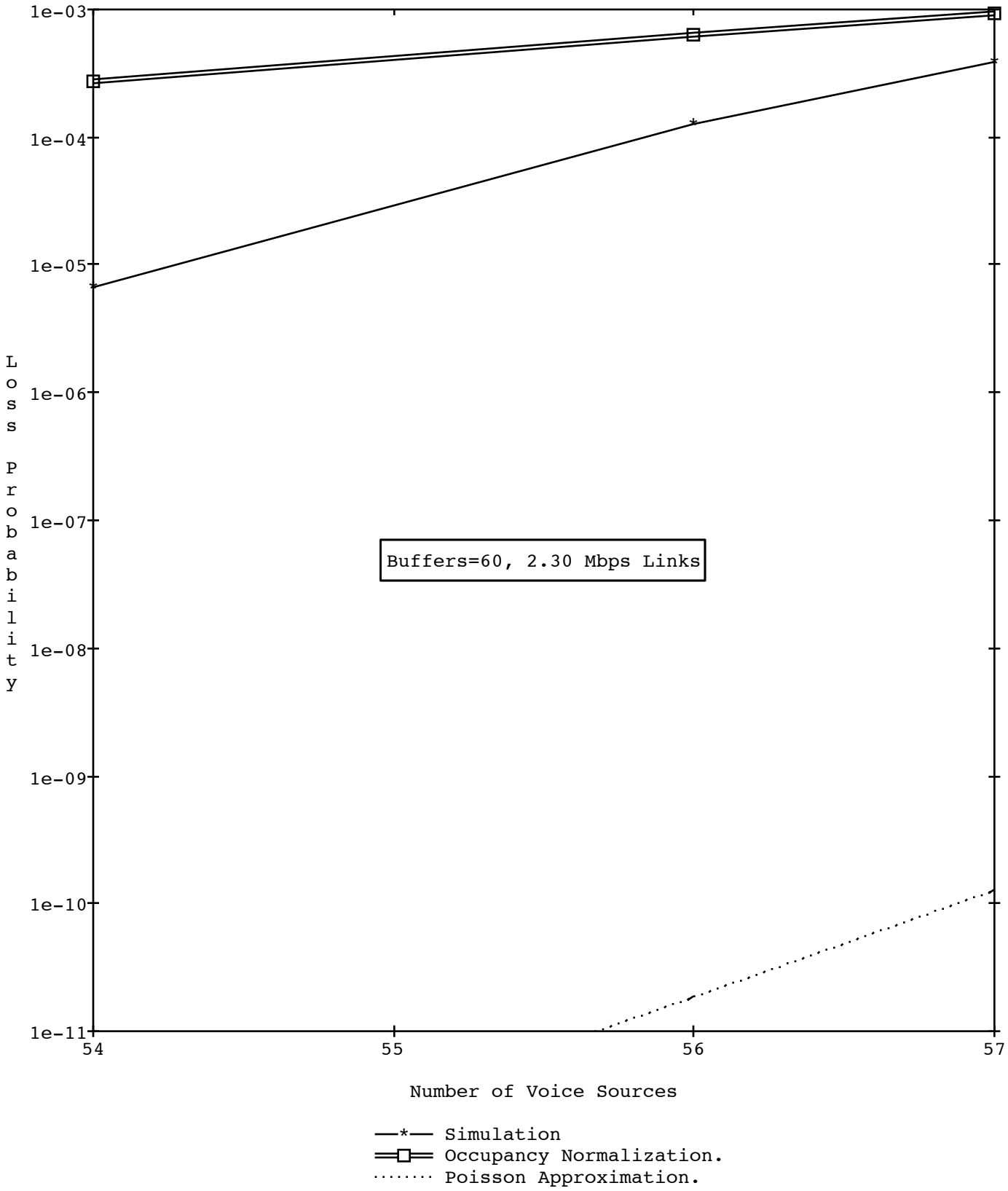

Figure 5: Loss in a three hop network

infinite buffer multiplexer. It is suggested in [7] that the analysis of finite-buffered multiplexers can be handled in a similar fashion. We also use the MMPP to model the superposition but present a simpler analytical approach to compute packet loss and show that the MMPP parameters, which successfully approximated average delay in [7], do not work as well in the case of analyzing packet loss; we derive an alternate set of MMPP parameters which is shown to compute loss more accurately.

In section 4.1 we define the MMPP formally and discuss its use for approximating the arrivals of voice packets at a statistical multiplexer. Following this, in section 4.2 we derive formulae for computing packet loss in the $MMPP/D/1/K$ queue. In section 4.3, we discuss several alternative ways in which the parameters of the MMPP model can be determined. We re-examine the issue of modeling (approximating) an arrival process with a large number of arrival states and present certain intuitive statistical measures that are shown to be important in the modeling of such systems (at least in the case of the performance measure of interest here).

4.1 The MMPP and Superimposed Traffic Flows

The MMPP is a special case of the versatile Markovian process [16]. Arrivals are governed by a continuous time, discrete state Markov chain, $\mathcal{M}$, in the following manner. Let M be the number of states in $\mathcal{M}$ labeled $m = 1, \cdots, M$. When the process is in state m, packets arrive according to a Poisson process with parameter λ_m.

As in [7], we use a two-state ($M = 2$) MMPP to approximate the superposition of voice sources. The parameters of the two-state MMPP, $\lambda_1, \lambda_2, r_1$ and r_2, will be determined by matching certain statistical characteristics (e.g., the mean arrival rate) of the MMPP, expressed in terms of $\lambda_1, \lambda_2, r_1$ and r_2, to the corresponding characteristics of the superposition of the voice sources. These latter values will be computed by directly analyzing the superimposed arrival stream generated by the voice sources. A wide variety of statistical characteristics could potentially be matched and we will see that matching certain statistics provides more accurate packet loss predictions than others. For the moment, we will assume that $\lambda_1, \lambda_2, r_1$ and r_2 are known and analyze packet loss in a queueing system in which the arrival process is modeled by an $M-$ state MMPP. Note that although we use a two state MMPP to model the superposition, the analysis in the next section is for the general case.

4.2 Analysis of the $MMPP/D/1/K$ queue

In contrast to the analysis in [7], we focus on modeling a statistical multiplexer with finite capacity. Our analysis employs the technique of uniformization in a fashion similar to that by Grassmann [6] and later by Lucantoni and Ramaswamy [13] in the analysis of phase type queues. More recently Blondia [4] has carried out, independently, a detailed analysis of the $N/G/1/K$ queue (the $MMPP/D/1/K$ queue is a special case) which is essentially identical to our analysis, except for the computation of the loss probability. Our approach for the computation of the loss probability

leads to a much simpler expression.

In analyzing the $MMPP/D/1/K$ queue, we will find it useful to study the state of this queue at departure instants. Let $(\tau_n : n \geq 0)$ be the successive epochs of departure and X_n and J_n be, respectively, the number in the queue and the state of the MMPP at τ_n^+. The sequence of triples $(X_n, J_n, \tau_{n+1} - \tau_n)$ then forms a Markov renewal sequence. We can write down the transition probability matrix for the embedded Markov chain and then solve for the joint distribution of the number in the queue and the state of the MMPP at departure instants. This, in turn, can be used to solve for the performance measures of interest. We now present the details of the solution technique.

Consider a buffer that can hold up to K packets that is served by a single server. We assume that each packet requires τ units of service. The arrival process is generated by a M-state MMPP as outlined earlier. Let $\boldsymbol{Q}$ be the infinitesimal generator of $\mathcal{M}$. Also, let $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_M)$.

Now we study the behavior of the discrete time Markov chain imbedded at the instants of departures. Let (M_i, N_i) denote the state of the queue where M_i is the state of $\mathcal{M}$ immediately after the i-th departure and N_i is the number of packets left behind by the departing packet. We are interested in the behavior of $(M, N) = \lim_{i \to \infty}(M_i, N_i)$, when it exists. This chain is aperiodic, recurrent and irreducible provided that $\mathcal{M}$ is aperiodic, recurrent, and irreducible and $\lambda_m < \infty$, $1 \leq m \leq M$. Let $\boldsymbol{T} = [t_{m,n;k,l}]$ be the transition probability matrix for this process, i.e., $t_{m,n;k,l} = P[M_{i+1} = k, N_{i+1} = l | M_i = m, N_i = n]$.

In order to determine $\boldsymbol{T}$, we first consider a version of $\mathcal{M}$ uniformized over a Poisson process with parameter $\lambda \geq \max\{\lambda_m - q_{m,m}\}$ [20]. This discrete time Markov chain has transition probabilities $r_{m,n}$ given by

$$r_{i,j} = \begin{cases} q_{i,j}/\lambda, & j \neq i, \\ (\lambda + q_{i,i})/\lambda, & j = i, \end{cases} \tag{16}$$

where an arrival occurs at state transitions with probabilities

$$p'_{i,j} = \begin{cases} 0, & i \neq j, \\ \lambda_i/(\lambda + q_{i,i}), & j = i. \end{cases} \tag{17}$$

We define $r^{(l)}_{i,j,k}$ to be the probability that the state is j, and that k arrivals occur in l transitions given that the initial state was i. If we let $\boldsymbol{P}$ be the matrix with elements $(1 - p'_{i,j})r_{i,j}$ and $\boldsymbol{P'}$ be the matrix with elements $p'_{i,j}r_{i,j}$, then

$$\boldsymbol{R}^{(l)}_k = \begin{cases} \boldsymbol{I}, & k = 0, l = 0, \\ \boldsymbol{R}^{(l-1)}_k \boldsymbol{P} + \boldsymbol{R}^{(l-1)}_{k-1} \boldsymbol{P'}, & 1 \leq k \leq l. \end{cases} \tag{18}$$

where $\boldsymbol{I}$ is the identity matrix.

We now define $M \times M$ matrices $\boldsymbol{A}_i$ and $\boldsymbol{B}_i$, $i \geq 0$ which define the transition probability matrix $\boldsymbol{T}$. Let $\boldsymbol{A}_i = [a_{i;m,k}]$, where $a_{i;m,k}$ is the probability that i packets arrive and the resulting state

of the uniformized version of $\mathcal{M}$ is k at the end of a service interval given that it was m at the beginning of the service interval. Let $\boldsymbol{B}_i = [b_{i;m,k}]$; where $b_{i;m,k}$ denotes the probability of i packet arrivals and the resulting state of the uniformized version of $\mathcal{M}$ is k given that an idle period preceded the service period and that the state of the arrival process was m at the start of the idle period. Also, let $\boldsymbol{U} = [u_{i,j}]$; where $u_{i,j}$ is the probability that an idle period terminates with the MMPP in state j given that the idle period started with the MMPP in state i. These matrices are given as

$$\boldsymbol{A}_i = \sum_{j=i}^{\infty} \frac{(\lambda\tau)^j e^{-\lambda\tau}}{j!} \boldsymbol{R}_i^{(j)}, \quad i \geq 0, \tag{19}$$

$$\boldsymbol{B}_i = \boldsymbol{U}\boldsymbol{A}_i, \tag{20}$$

and $\boldsymbol{U} = (\boldsymbol{\Lambda} - \boldsymbol{Q})^{-1}\boldsymbol{\Lambda}$ [7]. To evaluate the probability matrices $\boldsymbol{A}_i$ we truncate the infinite summation using techniques developed in [19]. In our solution, we truncate the series so as to obtain an accuracy of about ten significant digits.

The transition probability matrix can now be written as:

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{B}_0 & \boldsymbol{B}_1 & \boldsymbol{B}_2 & \cdots & \Sigma_{i=k-1}^{\infty}\boldsymbol{B}_i \\ \boldsymbol{A}_0 & \boldsymbol{A}_1 & \boldsymbol{A}_2 & \cdots & \Sigma_{i=k-1}^{\infty}\boldsymbol{A}_i \\ 0 & \boldsymbol{A}_0 & \boldsymbol{A}_1 & \cdots & \Sigma_{i=k-2}^{\infty}\boldsymbol{A}_i \\ & & & \vdots & \\ 0 & 0 & 0 & \cdots & \Sigma_{i=1}^{\infty}\boldsymbol{A}_i \end{bmatrix} \tag{21}$$

Let $\pi_{m,n} = P[M = m,\ N = n]$. These probabilities can be obtained by solving

$$\boldsymbol{\Pi}\boldsymbol{T} = \boldsymbol{\Pi}, \qquad \boldsymbol{\Pi}\boldsymbol{e} = 1. \tag{22}$$

To determine the probability of loss in the $MMPP/D/1/K$ queue, we use the the principle of flow conservation. Let λ_{av} be the average arrival rate of the MMPP and μ denote the throughput of successfully transmitted packets. Also, let D be the interdeparture time random variable and P_l be the probability of loss.

$$P_l = 1 - \frac{\mu}{\lambda_{av}} = 1 - \frac{1}{\lambda_{av}E[D]}. \tag{23}$$

It is easily seen that the mean interdeparture time is given as (consider departure instants)

$$E[D] = \tau + E[I]P(N = 0) \tag{24}$$

where I is the idle period random variable and N is the number of packets left behind by a departing packet. We compute the mean idle period next.

Since the conditional (on MMPP states) mean idle period of the MMPP is in matrix form $(\boldsymbol{\Lambda} - \boldsymbol{Q})^{-1}$, the unconditional mean idle period is easily obtained as

$$E[I] = \boldsymbol{P}(\boldsymbol{\Lambda} - \boldsymbol{Q})^{-1}\boldsymbol{e} \tag{25}$$

where $\boldsymbol{P}$ is the conditional distribution vector at departure instants (that leave the system empty) of the MMPP states and $\boldsymbol{e}$ is the unit vector. Substituting the above in equation (24), we obtain for the mean of the interdeparture distribution

$$E[D] = \tau + \boldsymbol{\pi_0}(\boldsymbol{\Lambda} - \boldsymbol{Q})^{-1}\boldsymbol{e} \tag{26}$$

where $\boldsymbol{\pi_0}$ is a probability vector with ith element, $\pi_{i,0}$.

For the two state MMPP we obtain

$$E[D] = \tau + \frac{(\pi_{1,0}(\lambda_2 + r_1 + r_2) + \pi_{2,0}(\lambda_1 + r_1 + r_2))}{((\lambda_1 + r_1)(\lambda_2 + r_2) - r_1 r_2)}. \tag{27}$$

The average arrival rate of the two state MMPP required to calculate the loss probability in equation (23) is defined in [7].

4.3 Calculating Loss Using various MMPP Models

Recall from our discussion in section 4.1 that the four parameters of the two-state MMPP, $\lambda_1, \lambda_2, r_1$ and r_2 are determined by matching certain statistical characteristics of the MMPP, expressed in terms of $\lambda_1, \lambda_2, r_1$ and r_2, to the corresponding characteristics of the superposition of the voice sources and that these latter values are computed by directly analyzing the superimposed voice sources. The first set of statistics we match, are those proposed in [7]:

MMPP Model I

1. The average 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 number of arrivals in $(0, t_2)$.

The equations that define these statistics in terms of the parameters of the MMPP and the superposition of voice sources are formulated in [7]. Since the choice of the time-scale (t_1 and especially t_2) in [7] is arbitrary (not analytically defined), the values we use to obtain the parameters of the MMPP are the same as in [7]. The four equations obtained by matching the above four statistics of the MMPP to those of the superimposed voice sources can be solved to obtain the parameters of the approximating MMPP; the solution technique is presented in section 3 of [7]. Figure 6 plots the loss probability in the concentrator as a function of the number of voice sources being superposed at the input for the MMPP model. Note that the loss predicted by the MMPP model under model I is very similar (in accuracy) to that of the earlier renewal model. The MMPP model can be seen to perform better in heavy traffic but the results for 80% utilization (110 sources) are almost the same in both models.

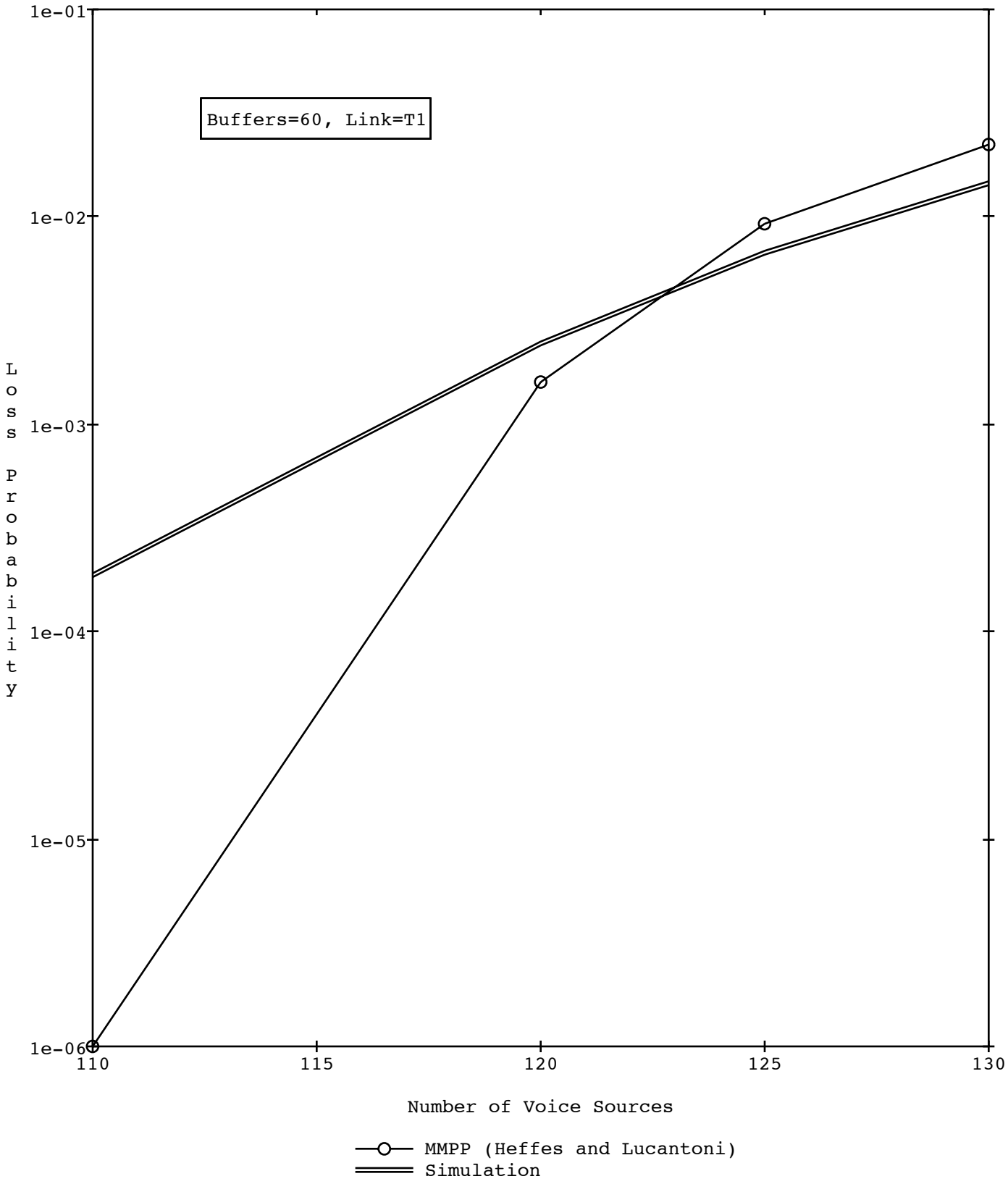

Figure 6: Loss in a Packet Voice Multiplexer (MMPP model (Heffes and Lucantoni))

Previous work [7, 17, 21, 25] has indicated that correlation in the arrival process plays a key role in determining the values of the performance measures of interest. In order to improve the accuracy of the loss predictions of the MMPP model, we conjectured that it might be possible to explicitly consider the effects of correlation by choosing the appropriate statistics to be matched. The approach we adopt towards modeling this correlation is quite intuitive and is directly related to the actual structure of the superposition process.

The correlation that we sought to model results from the fact that changes in the aggregate arrival rate (i.e., the number of voice sources in talkspurt) occur over relatively long time scales compared to the interarrival time of a packet in the superimposed arrival process. Thus, for example, if there were a large number of active voice sources, these sources would be likely to remain active for a relatively long period of time. We conjectured that at least in the case of a large buffer size, it would be these extended periods of congestion that would contribute primarily to packet loss. Thus, our goal was to identify a set of "arrival state(s)" of the superposition process which contribute significantly to packet loss, and to then determine parameters of the MMPP based on statistics associated with these arrival states.

We adopt a division of the arrival rates similar to that in Li [17]. In [17], the arrival process is considered to consist of an underload period and an overload period. An overload state occurs when the number of active sources exceeds the capacity of the system or alternatively the arrival rate exceeds the service rate. In the following, we will match the average traffic rates in both the overload and underload states. Higher moments of the arrival process will be considered depending on the buffer size. For moderate to large buffer sizes, the variance in the arrival process during the overload states will be considered, since (as mentioned above) packet loss is expected to occur in the overload states only; we will refer to the set of MMPP parameters derived when the variance in the overload states is explicitly considered as **MMPP Model II.** For small buffers we will match the variance in the underload states **(MMPP Model III)**. No analytical technique is presented to determine which range of buffer sizes would match best with which particular model. However, we are able to offer some intuitive insight. In general, the buffer size for which the underload period causes packet loss depends on the individual source's packet generation process in talkspurts (this is intuitive since loss in the underload period is due to the stochastics of the packet generation process rather than correlation). The less bursty this process, the smaller is the buffer size for which the underload period is significant. Hence in the case of the superposition of voice sources, (deterministic arrivals in talkspurts) the overload model was found sufficient for most buffer sizes. However, for the superposition of bursty sources (Poisson arrivals in talkspurts) the underload period had to be modeled for buffer sizes smaller than 30.

In the following, we present the MMPP models derived from these matchings. In order to study the effects of various assumptions on the quality of the loss approximations, we will use the MMPP model to approximate the superposition of *two different source models* (separately). Source model 2 (Source2) is the voice source model outlined in section 2. Source model 1 (Source1) will be quite similar, with the difference being that the packet arrival process in talkspurts will be Poisson

instead of deterministic. Source1 is also referred to as the Interrupted Poisson Process (IPP) in the literature. Source1 serves as the model for a overflow process in [14].

The mathematical details (expressions) for the new proposed statistics are presented in appendix B. Here, we overview the analysis and present some mathematical details. As in all previous models, we consider the counting process for the actual superimposed arrival process and the MMPP and attempt to match certain moments of this process. Let us first define some random variables:

$N^S(0,t)$ - The number of arrivals in time t for the superposition (S) process. We take $t = 0$ to be an arbitrary (random) time instant.

$N_O^S(0,t)$ - The number of arrivals in time t for the superposition process, given that the voice sources are in an overload (O) state at $t = 0$.

$N_U^S(0,t)$ - The number of arrivals in time t for the superposition process, given that the voice sources are in an underload (U) state at $t = 0$.

In our two-state MMPP models, one of the two states will correspond to the "high arrival rate" state, and the other will correspond to the "low arrival rate" state. Informally, the MMPP being in the high arrival rate state will model the event that the voice sources are in an overload state. For the MMPP we thus similarly define:

$N^M(0,t)$ - The number of arrivals in time t for the MMPP (M).

$N_H^M(0,t)$ - The number of arrivals in time t for the MMPP, given the MMPP is in the high arrival (H) state at $t = 0$.

$N_L^M(0,t)$ - The number of arrivals in time t for the MMPP, given the MMPP is in the low arrival (L) state at $t = 0$.

We will formulate the two MMPP models below. In order to match statistical properties of the MMPP's to those of the superimposed voice sources, we will match $E[N_O^S(0,t)]$ and $E[N_H^M(0,t)]$ for all time t. Three statistics are necessary to match the averages for all t since the expression for the average (see equation (38) in appendix B) has three unknowns. Three statistics common to both our models are:

1. Value of $E[N_O^S(0,t)]/t$ and $E[N_H^M(0,t)]/t$ at $t = 0$.
2. Value of $E[N_O^S(0,t)]/t$ and $E[N_H^M(0,t)]/t$ at $t = \infty$. Note that evaluating these expressions at $t = \infty$ matches the long-term average arrival rate of the MMPP to that of the superimposed voice sources.
3. Derivative of $E[N_O^S(0,t)]/t$ and $E[N_H^M(0,t)]/t$ at $t = 0$.

We note that the parameter set used in [12] to accurately predict the average delay uses the first three parameters of MMPP Model I and a slightly different variation of the first parameter above (average arrival rate in the overload period). This further confirms our intuition on the strong role of the overload period in determining the multiplexer performance. The fourth statistic used in the matching differs in our two MMPP models and is discussed below.

4.3.1 MMPP Model II

We use this model for predicting packet loss in a multiplexer with moderate to large buffer sizes. For these buffer sizes we expect loss to occur only in the overload states. Hence, in addition to the above three statistics for the averages, we consider a higher moment of the arrival process which is specifically related to the overload states:

Model II

1. 1., 2. and 3. as above and,
2. Value of VAR($N_O^S(0,t)$) and VAR($N_H^M(0,t)$) at $t = t_m$.

The value of t_m is chosen so that VAR($N_O^S(0,t)$) and VAR($N_H^M(0,t)$) match well *over a time period* of one second. We briefly discuss this choice of general time scale for the matching and then discuss our specific choice of t_m.

We choose a time period of one second as it is the sum of the average talkspurt and silence period durations. Hence, it is easy to calculate given the source parameters. This choice has been suggested in a number of earlier papers. Weiss [23] refers to it as the relaxation time. He takes it to be the memory in the source for the initial condition. Ramaswamy [18] recommends one second as an appropriate time scale to model the superposition, based on the time period over which the peaks in the serial correlation of counts for the superposition of voice sources last (the peaks being caused by the correlation). Hence, a time period of one second may be thought of as the period of "significant" correlation in the voice source. The results of our investigation provide further justification to the appropriateness of the choice.

The process of choosing a value of t_m so that VAR($N_O^S(0,t)$) and VAR($N_H^M(0,t)$) matched exactly at t_m and also matched "well" over the entire one second time period was a bit involved. Given that we match the variances at some time t_m (say one second), we found that the MMPP displayed a higher variance for all time $t > t_m$ and lower variance for all time $t < t_m$. Given that we wish to match the variances over a time period of one second, it seemed reasonable to match the two curves (variances) at $t_m = 0.5$ *sec.* in the hope that the effects of higher variance for $t_m < 0.5$ would cancel out the effects of the lower variance in the interval $0.5 \leq t_m \leq 1.0$. We found that we could carry out this matching (at $t_m = 0.5$ *sec.*) for a superposition of sources greater than 110 but not for 110 sources (the solution of non-linear equations did not converge). In this case (110 sources), we simply tested values of t_m starting at 0.5 *sec.* in increments of 0.1 *sec.* until a

Sources	t_m (secs)	K=30		K=60		K=90	
		Analysis	Simulation	Analysis	Simulation	Analysis	Simulation
110	1.0	7.79e-04	5.26e-04	3.93e-04	1.89e-04	-	-
120	0.5	2.14e-03	4.46e-03	1.19e-03	2.47e-03	6.87e-04	1.49e-03
125	0.5	1.19e-02	1.003e-02	7.82e-03	6.64e-03	5.22e-03	4.56e-03
130	0.5	2.58e-02	1.93e-02	1.93e-02	1.46e-02	1.48e-02	1.16e-02

Table 2: Loss probabilities for a superposition of voice sources (Source2)

Sources	t_m (secs)	K=60	
		Analysis	Simulation
110	1.0	3.93e-04	1.89e-04
120	1.0	5.91e-03	2.47e-03
125	1.0	1.21e-02	6.64e-03
130	1.0	2.23e-02	1.46e-02

Table 3: Loss probabilities for a superposition of voice sources (Source2)

solution was obtained. Table 2 shows the values of time at which curves were matched and also the packet loss probabilities predicted by the MMPP Model II. As shown by Tables 2 and 3, matching at $t_m = 1$ *sec.* provided slightly larger loss values than matching at a smaller value for t_m, since in this case the MMPP had a higher variance than the superposition for the entire duration of the pertinent time period. However, the loss values computed using the two values of t_m do not differ significantly. Thus, since $t_m = 1$ *sec.* always provides a solution, this value might be preferable when the accuracy of the prediction is not of tantamount concern.

4.3.2 MMPP Model III

We use this model for predicting packet loss in a multiplexer with a small buffer size. For these buffer sizes we expect loss to occur in the underload states in addition to the overload states. Hence, we consider higher moments related to the underload states. In **Model III** we match,

1. 1., 2., and 3. as defined earlier and
2. Value of VAR$(N_U^S(0,t))$ and VAR$(N_L^M(0,t))$ at $t = t_m$.

In this case a choice of $t_m = 1$ *sec* was adequate for matching VAR$(N_U^S(0,t))$ and VAR$(N_L^M(0,t))$ over a time period of one-second. The results are shown in Table 4, and once again show good agreement between the simulation results and those predicted by the MMPP model.

4.4 Discussion

In order to test the robustness of the performance predictions generated by our approach towards matching the MMPP parameters with those of the actual source, we used the MMPP to model the

Sources	t_m (secs)	K=10	
		Analysis	Simulation
110	1.0	1.81e-02	1.15e-02
120	1.0	3.07e-02	2.52e-02
125	1.0	3.98e-02	3.59e-02
130	1.0	5.13e-02	4.87e-02

Table 4: Loss probabilities for a superposition of bursty sources (Source1)

Sources	t_m (secs)	K=60	
		Analysis	Simulation
110	0.9	3.52e-04	4.59e-04
120	0.5	3.24e-03	4.34e-03
125	0.5	1.08e-02	9.67e-03
130	0.5	2.30e-02	1.84e-02

Table 5: Loss probabilities for a superposition of bursty sources (Source1)

superposition under an alternate source model (Source1) as well. Matching between the MMPP and the sources was then again performed as discussed above. Table 5 shows the time points at which the variance curves were matched and the resulting loss probabilities. Once again, good agreement is obtained between simulation and analysis. We also see that the predictions are not as accurate for the case of deterministic arrivals as for the case of Poisson arrivals in talkspurts. This is due to the fact that the variance curve is matched better over the one second time period in the case of Source1 (Poisson arrivals). We also note that a comparison of the *simulation* values of packet loss under source models Source1 and Source2 (Tables 2 and 5) indicates that the loss under source model Source1 is higher than the loss under model Source2 by half an order of magnitude. The higher burstiness of the Poisson source (as compared to the deterministic source) is responsible for the higher loss. Thus, the short term stochastics of the arrival process will also need to be considered (in addition to long term correlation) when particularly accurate models of such systems are required.

Further, we tested our MMPP model with a different link capacity than the T1-rate link considered so far. The simulation and modeling results for a link capacity of 2.048 Mbps and a buffer capacity of 60 are shown in Table 6. MMPP model II is used to predict the loss due to the reasonably large buffer size. Since $t_m = 1$ *sec.*, the predicted loss values are seen to be larger than the actual values, which is as expected. However, the predicted values are reasonably accurate. For greater accuracy, a better choice for t_m may be made as discussed earlier.

In summary, we would note that as shown in Figure 7, the MMPP models developed here provide better performance predictions than both the renewal approach of the previous section as well as the MMPP approach using the parameters suggested in [7]. Our results also provide insight into how the issue of correlation affects the modeling of bursty sources.

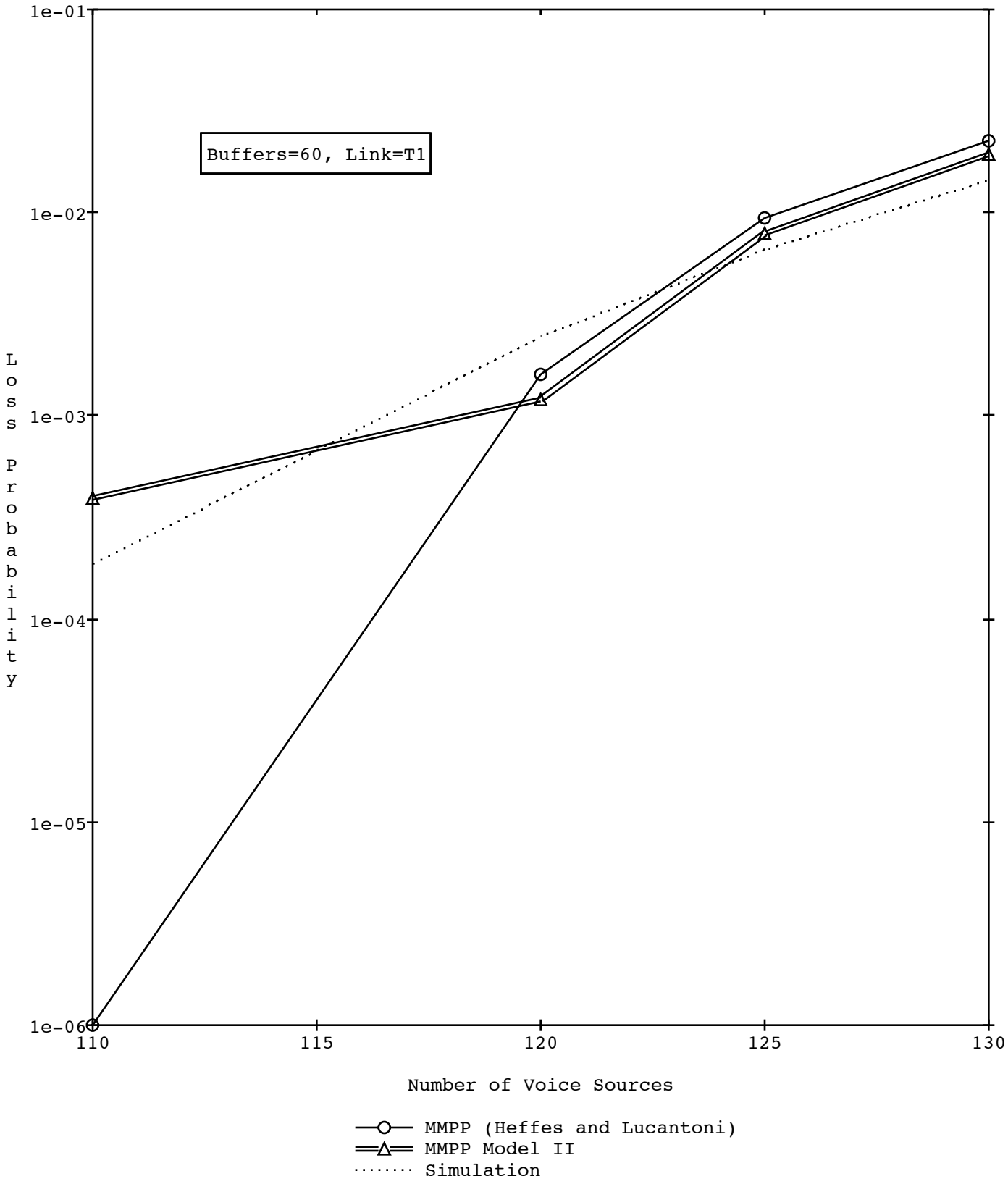

Figure 7: Loss in a Packet Voice Multiplexer (MMPP models - Comparison)

Sources	t_m (secs)	K=60	
		Analysis	Simulation
150	1.0	6.3e-04	4.85e-04
155	1.0	2.21e-03	1.02e-03
160	1.0	4.69e-03	2.77e-03
165	1.0	8.58e-03	5.82e-03

Table 6: Loss probabilities for a superposition of bursty sources (Source1) over a 2.048 Mbps link

5 Fluid Flow Approximation

In this section, we examine the suitability of a fluid flow model, as developed in [1], for predicting packet loss in a finite-buffered voice multiplexer. A thorough analysis of this model is presented in Anick, Mitra and Sondhi [1] and Mitra [15]. Mitra [15] considers multiple sources and multiple servers coupled by a finite buffer. The suitability of a fluid flow approximation for a superposition of voice sources is already examined by Tucker [22] and hence our treatment of the approximation is abbreviated.

Once again, the packet generation process is taken to be a collection of sources alternating between active and inactive states, as in our single voice source model in section 2. The fundamental assumptions underlying the fluid model are that each active source generates information at a uniform rate of one unit of information per unit of time and that the server removes information from the buffer at a uniform rate of C units per unit of time. The unit of time and information can be defined in any convenient fashion. In [1], it is assumed that the active source generates one unit of information in an active period and the average duration of the active period is taken to be the unit of time. For the voice source of interest here, one unit of information would thus correspond to $\frac{1}{\alpha T}$ packets where T is the (deterministic) interarrival time of packets during a talkspurt and α^{-1} is the average duration of the talkspurt.

Let $F_i(x)$ be the steady state probability of i sources being active and the buffer content being less than or equal to x units. Then we have [1, equation (7)],

$$\begin{aligned}(i-c)\frac{dF_i}{dx} &= (N-i+1)\lambda F_{i-1} - ((N-i)\lambda + i)F_i \\ &\quad + (i+1)F_{i+1} \quad 0 \le i \le N \end{aligned} \tag{28}$$

where N is the total number of data sources and λ^{-1} is the average time units that the data source spends in an inactive state. (Note that the definition of λ above differs from our previous definition; we use the notation above for compatibility with [1]). [1] is primarily concerned with the solution of this system of differential equations. Due to inherent instabilities in the system [1, 8] any numerical solution technique for differential equations cannot be used without exercising care that the solution does not grow (unbounded) exponentially. Instead the problem is formulated as an eigenvalue problem and a simple analysis yields all the eigenvalues and eigenvectors. The

Sources	K=60	
	Analysis	Simulation
110	1.4e-04	1.72e-04
120	2.33e-03	2.44e-03
125	4.87e-03	6.39e-03
130	1.1e-02	1.45e-02

Table 7: Loss probabilities for a superposition of voice sources (Fluid model)

solution to this set of differential equations can be written [1]:

$$\mathbf{F}(x) = \Sigma_{i=0}^{N} e^{z_i x} a_i \phi_i \tag{29}$$

where z_i are the eigenvalues, and ϕ_i are the eigenvectors, and $a_i's$ are the coefficients which are determined from certain boundary conditions [15]. The blocking state probability, P_l, is then given by [15, equation 6.11ii],

$$P_l = \Sigma_{i=C+1}^{M} \pi_i - F_i(K). \tag{30}$$

Appendix C briefly outlines the procedure to obtain the eigenvalues and eigenvectors.

The results of the fluid approximation for computing packet loss probabilities in the voice multiplexer are shown in Table 7. The computed loss probabilities in Table 7 are the time averages for a full buffer and not the loss probabilities seen by voice packets (arrivals). We note that the results are quite accurate - yielding performance prediction as accurate as any of the other two methods studied.

Note that the fluid flow model does not incorporate the higher moments of the arrival process during talkspurts. Thus, the model would yield the same loss predictions shown in Table 7 if the interarrival times during talkspurt were exponentially distributed, i.e., it predicts the same loss under source models Source1 and Source2. A comparison of the analytic and simulation results from Table 5 (packet loss under source model Source1) with the analytic results in Table 7 show that, in this case, the MMPP predictions are closer to the simulation results than the fluid flow results. The suitability of a fluid flow versus MMPP approach towards modeling packet loss will thus depend on the nature of the packet arrival process. Also, the fluid flow model does not consider the effect of the underload states (i.e., the queue length does not grow in the underload states in the fluid approximation). Loss could occur in the underload states with small buffer sizes and needs to be modeled as in our MMPP Model III. Table 8 shows the inaccuracy in the fluid model as a result of ignoring the effect of the underload states whereas our MMPP Model III predicts loss relatively accurately.

6 Discussion

In Figure 8 we compare the performance of our models for the case of voice sources (Source2) multiplexed onto a T1-rate link with a buffer capacity of 60. We see from Figure 8 that the

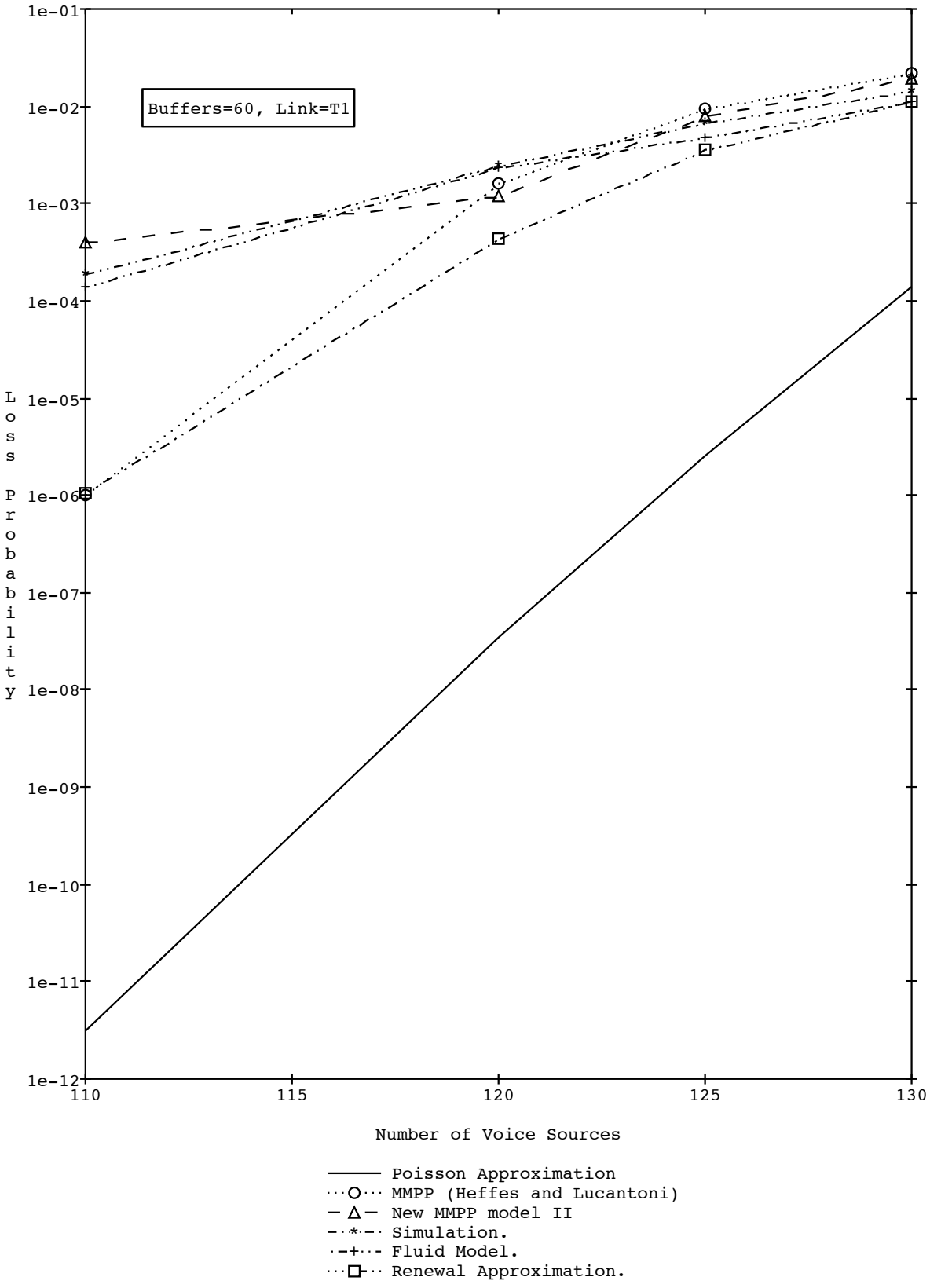

Figure 8: Loss in a Packet Voice Multiplexer (Approximate models - Comparison)

Sources	K=5		
	MMPP Model III	Fluid Model	Simulation
110	4.71e-02	3.99e-03	3.135e-02
120	6.80e-02	2.81e-02	5.125e-02
125	8.03e-02	5.20e-02	6.30e-02
130	9.37e-02	1.44e-01	7.73e-02

Table 8: Loss probabilities for a superposition of voice sources (Comparison)

renewal model and MMPP model using the parameter matching technique of [7] are less accurate in predicting packet loss. The fluid flow model and the MMPP model using the matching technique suggested in this paper perform comparably.

The fluid model accurately captures the *correlational properties* of the superposition but fails to account for the *stochastics in the flows*. We conjecture that the correlational aspects of the sources dominate the performance metrics, but for particularly accurate modeling it is necessary to account for the stochastics. This is also the reason that the fluid flow model fails for small buffers wherein the stochastics are more significant than the correlation. We have also shown that the fluid model is not very accurate for a source model with Poisson arrivals in talkspurts since again the fluid approximation only models the correlation and not the stochastics. Our MMPP model accounts for the correlational aspects but, in addition, also attempts to model the stochastics (using second moment information). In this context, the MMPP model was shown to provide reasonably good results for two different source models as well as for a range of buffer sizes.

The complexity of the solution techniques should also be considered in the choice of a particular model. There are two steps in the solution of our MMPP model. First, a set of non-linear equations must be solved to determine the parameters of the MMPP. Next, the associated queueing model must be solved. The total time for proceeding through the two stages is slightly greater than the time taken to solve the fluid model for the parameter ranges considered here. On the other hand, the fluid model potentially suffers from numerical problems which may cause it to break down for large systems; the reader may refer to [23] for examples of such systems (the problem is alleviated to a extent in [22] by truncation of the state space).

The renewal approximation is the simplest and most versatile of all the models. However, it has limited accuracy. It may be used in heavy traffic situations with reasonable accuracy. The versatility of the renewal process approximation has been demonstrated by its application in a network setting. Further, the applicability of the considered traffic models seems limited to off-line buffer dimensioning (due to their computational complexity) rather than on-line performance evaluation (for e.g., in call 'admission control' for determining if a specified grade of service can be satisfied). However, the renewal process approximation seems promising in this respect.

A number of interesting open problems still remain. The applicability of the stochastic models in this paper for modeling the superposition of nonhomogenous sources remains to be studied. The multiplexer performance in a *network* setting needs to be further investigated. More accurate

stochastic models for the intra-network traffic are necessary in order to predict the performance of the multiplexer in a network setting accurately.

References

[1] D. Anick, D. Mitra, and M. M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *Bell System Technical Journal*, 61:1871–1894, 1982.

[2] Robert B.Cooper. *Introduction to Queueing Theory: Second Edition.* North Holland, 1981.

[3] Amit Bhargava et al. A performance comparison of error control protocols in high speed communication networks. In *IEEE INFOCOM'88*, March 1988.

[4] C. Blondia. The N/G/1 finite capacity queue. *Communications in Statistics - Stochastic Models*, 5:273–293, 1989.

[5] John N. Daigle and Joseph D. Langford. Models for analysis of packet voice communications systems. *IEEE J.Select.Areas Commun.*, SAC-6:847–855, 1986.

[6] W. K. Grassmann. The GI/PH/1 queue: A method to find the transition matrix. *Infor*, 20:144–156, 1982.

[7] Harry Heffes and David Lucantoni. A markov modulated characterization of voice and data traffic and related statistical multiplexer performance. *IEEE J.Select.Areas Commun.*, SAC-4:856–867, September 1986.

[8] L. Kosten. Stochastic theory of a multi-entry buffer. *Delft Progress Report*, pages 10–18, 1974.

[9] K.T.Marshall. Some inequalities in queueing. *Operations Research*, May 1968.

[10] Paul Kuehn. Approximate analysis of general queueing networks by decomposition. *IEEE Transactions on Communications*, COM-27:113–125, January 1979.

[11] Stephen Lavenberg. *Computer Performance Modeling Handbook.* Academic Press, 1983.

[12] Ke-Qiang Liao and Lorne G. Mason. A discrete-time single server queue with a two level modulated input and its applications. In *IEEE GLOBECOM'89*, November 1989.

[13] David M. Lucantoni and V. Ramaswamy. Efficient algorithms for solving the non-linear matrix equations arising in phase type queues. *Communications in Statistics - Stochastic Models*, 1:29–51, 1985.

[14] Kathleen S. Meier-Hellstern. The analysis of a queue arising in overflow models. *IEEE Transactions on Communications*, 37:367–372, April 1989.

[15] Debasis Mitra. Stochastic theory of a fluid model of producers and consumers coupled by a buffer. *Advances in Applied Probability*, 20:646–676, 1988.

[16] Marcel F. Neuts. A versatile markovian point process. *Journal of Applied Probability*, 16:764–779, December 1979.

[17] San qi Li. Study of packet loss in a packet switched voice system. In *ICC'88*, pages 47.1.1–47.1.8, June 1988.

[18] V. Ramaswamy. Traffic performance modeling for packet communication whence, where and wither. In *Third Australian Teletraffic Seminar*, November 1988. Keynote Address.

[19] Andrew Reibman and Kishore Trivedi. Numerical transient analysis of markov models. *Comput. Opns. Res.*, 15:19–36, 1988.

[20] Sheldon M. Ross. *Applied Probability Models.* Academic Press, fourth edition, 1989.

[21] Kotikalapudi Sriram and Ward Whitt. Characterizing superposition arrival processes in packet multiplexers for voice and data. *IEEE J.Select.Areas Commun.*, SAC-4:833–846, September 1986.

[22] Roger C. F. Tucker. Accurate method for analysis of a packet-speech multiplexer with limited delay. *IEEE Transactions on Communications*, 36:479–483, April 1988.

[23] Alan Weiss. A new technique for analyzing large traffic systems. *Journal of Applied Probability*, pages 506–532, 1985.

[24] Ward Whitt. The queueing network analyzer. *Bell System Technical Journal*, pages 2779–2813, November 1983.

[25] Chin Yuan and John A. Silvester. Queueing analysis of delay constrained voice traffic in a packet switching system. *IEEE J.Select.Areas Commun.*, 7:729–738, June 1989.

Appendix A

Here we look at the algebra applied to the two parameter characterizations to handle the operations of superposition, splitting and flow through a queue resulting from the decomposition approach.

.1 Superposition

In general, the superposition of renewal processes is a non-renewal process. The approximation in [24] is used. The squared coefficient of variation and mean of the interarrival time distribution in the complex superposition process are to be obtained and these are taken to be the moments of the approximating renewal process. The mean arrival rate is simply the sum of the mean arrival rates of the component processes. To obtain the squared coeffecient of variation, we use equations

(29), (31) and (33) in [24]. Equation (33) basically uses the Asymptotic Approximation procedure along with a weighting function as shown here.

$$c_H^2 = wc_A^2 + (1 - w) \tag{31}$$

where c_H^2 is the squared coeffecient of variation of the renewal interval c.d.f in the approximating renewal process and c_A^2 is the squared coefficient of variation of the renewal interval c.d.f as obtained by the Asymptotic Approximation method. w is the weighting function. w is a function of the number of component processes being superposed and the traffic intensity.

$$w = [1 + 4(1 - \rho)^2(v - 1)]^{-1} \tag{32}$$

where ρ is the traffic intensity and v is the number of component processes (actually an equivalent number) and is obtained as in [24, eq. 35].

.2 Splitting

A renewal process split according to independent probabilities is again a renewal process [24]. So no approximations are needed in this case. Equation (36) of [24] is used in this case. Equation (36) is derived in [10].

$$c_i^2 = p_i c_d^2 + (1 - p_i) \tag{33}$$

where c_i^2 is the squared coeffecient of variation of the ith process obtained from the splitting and c_d^2 is the squared coeffecient of variation of the process being split. p_i is the probability of the ith stream. The mean rate of the split processes is easily obtained from the rate of the original process and the splitting probabilities as

$$\lambda_i = \lambda_d p_i \tag{34}$$

λ_i and λ_d are the mean rates of the ith split process and the original stream respectively.

.3 Flow through a queue

The Stationary interval method is used here as the Asymptotic Approximation method yields an elementary approximation [24]. In fact, the Asymptotic method approximation for the squared coeffecient of variation of the departure process is just the squared coefficient of variation of the arrival process. Equation (42) of [24] is used here. Equation (42) is a modified version of equation (37) in [24] which was earlier obtained by Marshall[9]. Equation (37) gives the squared coefficient of variation of the departure process given the two parameter characterization of the arrival and service processes. In addition, it requires the mean waiting time which is obtained from equation (2) of [24] with g=1. Equations (2) and (37) put together yield equation (39) in [24]. However, it was seen that equation (39) overestimated the decrease in variability of the departure process due to a deterministic service process. Equation (42) overcomes the inaccuracy in equation (39) and is

the one used in our analysis. The mean rate of the departure process is the same as that of the arrival process. Equation (42) is reproduced here.

$$c_d^2 = 1 + (1-\rho^2)(c_a^2 - 1) + \frac{\rho^2}{\sqrt{m}}(max(c_s^2, 0.2) - 1) \tag{35}$$

where c_d^2, c_a^2 and c_s^2 are the squared coefficients of variation of the departure, arrival and service processes respectively. m is the number of servers and ρ is the traffic intensity.

Appendix B

In this appendix we derive the various expressions needed to fit the parameters of the MMPP to the two source models, Source1 and Source2. The random variables used in this appendix are defined in section 4.3. We take T to the deterministic packet interarrival time during talkspurts for the voice source. Also, α^{-1} and β^{-1} are taken to be the mean talkspurt and silence period durations for both source models. For the two-state MMPP, we define λ_H and λ_L as the arrival rates and r_H^{-1} and r_L^{-1} to be the mean sojourn times. The subscripts L and H are used to denote the low and high arrival states respectively.

First let us derive expressions for various statistics associated with the two-state MMPP. The generating function, $G(z,s)$, for the number of arrivals in time t for the MMPP is given by [7, equation (A1)]:

$$G(z,s) = \boldsymbol{\pi}[s\boldsymbol{I} - \boldsymbol{R} - (z-1)\boldsymbol{\Lambda}]^{-1}\boldsymbol{e} \tag{36}$$

where $\boldsymbol{\pi}$ is the vector of steady state probabilities for the MMPP states, $\boldsymbol{R}$ is the infinitesimal generator for the Markov chain of state transitions, $\boldsymbol{\Lambda}$ is the diagonal matrix of arrival rates for the MMPP, and $\boldsymbol{e}$ is the unit vector. To account for the condition on the initial state, we have to consider a modified $\boldsymbol{\pi}$ matrix. Since we are interested in the mean and variance of $N_H^M(t)$, we set $\boldsymbol{\pi} = [0\ 1]$.

Let us first derive an expression for the expected number of arrivals in time t assuming that the MMPP modeling the arrival process is initially in the high arrival state. Since the average is just the Laplace inverse of the partial derivative of the Laplace transform of the generating function, we have

$$\begin{aligned} E[N_H^M(0,t)] &= \mathcal{L}^{-1}[\frac{\partial}{\partial z}G(z,s)|_{z=1}] \\ &= \mathcal{L}^{-1}[\frac{\boldsymbol{\pi}}{s}(s\boldsymbol{I} - \boldsymbol{R})^{-1}\boldsymbol{\Lambda}\boldsymbol{e}]. \end{aligned} \tag{37}$$

After carrying out the above operation and a few manipulations we obtain

$$E[N_H^M(0,t)] = At + B(1 - e^{-Ct}) \tag{38}$$

where

$$A = \frac{\lambda_H r_L + \lambda_L r_H}{r_L + r_H},$$

$$
\begin{aligned}
B &= \frac{r_H(\lambda_H - \lambda_L)}{(r_L + r_H)^2}, \\
C &= r_H + r_L.
\end{aligned} \tag{39}
$$

For the second moment of the number of arrivals for the MMPP, we similarly have:

$$
E[(N_H^M(0,t))^2] = \mathcal{L}^{-1}[\frac{\partial^2}{\partial^2 z} G(z,s)|_{z=1}] + \mathcal{L}^{-1}[\frac{\partial}{\partial z} G(z,s)|_{z=1}] \tag{40}
$$

where the expression for the second term on the right is given by equation (38). After some manipulation, we get for the second term:

$$
\begin{aligned}
\mathcal{L}^{-1}[\frac{\partial^2}{\partial^2 z} G(z,s)|_{z=1}] &= \mathcal{L}^{-1}[2\frac{\boldsymbol{\pi}}{s}((s\boldsymbol{I} - \boldsymbol{R})^{-1}\boldsymbol{\Lambda})^2 \boldsymbol{e}] \\
&= 2(\frac{A^2}{2}t^2 + Gt + K(1 - e^{-(r_L + r_H)t}) + Rte^{-(r_L + r_H)t})
\end{aligned} \tag{41}
$$

where

$$
\begin{aligned}
G &= \frac{r_H r_L(\lambda_H - \lambda_L)^2 + r_H(\lambda_H - \lambda_L)(\lambda_H r_L + \lambda_L r_H)}{(r_L + r_H)^3}, \\
K &= \frac{r_H(\lambda_H - \lambda_L)^2(r_H - 2r_L)}{(r_L + r_H)^4}, \\
R &= \frac{(\lambda_L - \lambda_H) r_H(\lambda_H r_H + \lambda_L r_L)}{(r_L + r_H)^3}.
\end{aligned} \tag{42}
$$

The variance in the number of arrivals is then, by definition,

$$
\mathrm{VAR}(N_H^M(0,t)) = E[(N_H^M(0,t))^2] - E^2[N_H^M(0,t)]. \tag{43}
$$

Let us next derive expressions for various statistics associated with the superposition of the traffic sources, assuming that each source behaves as a Poisson process during a talkspurt (source model Source1). Define M as the total number of sources and C such that when C sources are actively in talkspurt, the overall arrival rate of packets equals the service rate of the multiplexer,

$$
C = min\{n : n/T \geq 1/d\} \;\; 1 \leq n \leq M, \tag{44}
$$

where d is the deterministic packet service time. C thus serves to delineate the high arrival states of the sources (when the number of sources in talkspurt exceeds C) from the low arrival states.

To obtain the mean and variance for the superposition of bursty sources (Source1), we note that each individual source behaves as a two-state MMPP with one of the arrival rates set to zero. Specifically, we can model each source as a two-state MMPP with

$$
\begin{aligned}
\lambda_L &= 0, \quad \lambda_H = 1/T \\
r_H &= \alpha, \quad r_L = \beta
\end{aligned} \tag{45}
$$

and use these value of $\lambda_l, \lambda_H, r_H$ and r_L when evaluating the expressions involving $N_H^M(0,t)$ and $N_L^M(0,t)$ below.

For the mean number of arrivals we have:

$$E[N_O^S(t)] = \Sigma_{i=C+1}^{M} \frac{p_i}{\Sigma_{j=C+1}^{M} p_j} (iE[N_H^M(t)] + (M-i)E[N_L^M(t)]) \tag{46}$$

where p_i is the steady state probability of exactly i sources being in talkspurt. From [17, equation (2)], we have:

$$p_i = \begin{pmatrix} M \\ i \end{pmatrix} f^i (1-f)^{M-i} \tag{47}$$

f is the activity factor for a single source given by

$$f = \frac{\beta}{\alpha + \beta}. \tag{48}$$

For the variance we have,

$$\text{VAR}(N_O^S(t)) = \Sigma_{i=C+1}^{M} \frac{p_i}{\Sigma_{j=C+1}^{M} p_j} (i\text{VAR}(N_H^M(0,t)) + (M-i)\text{VAR}(N_L^M(0,t)). \tag{49}$$

Finally, let us derive expressions for various statistics associated with the superposition of the traffic sources, assuming that each source behaves as a true voice source during talkspurt (source model Source2). For a superposition of voice sources, the expressions for the averages (equation (46)) remain the same. For the variance, we again use equation (49) suitably replacing the expressions for the variances of the MMPP by the variances for the voice source. Just as in the case of the MMPP, we first derive the generating function of the number of arrivals in time t for the voice source.

Let $P_0(n,t)$ be the probability of n arrivals in time t for the voice source given that the source is in state 0 (inactive) at $t = 0$. Similarly, let $P_1(n,t)$ be the probability of n arrivals in time t for the voice source given that the source is in state 1 (active) at $t = 0$. Before we derive expressions for $P_0(n,t)$ and $P_1(n,t)$ and their transforms, we first derive the generating function for the number of arrivals in a deterministic process with interarrival time T; this expression will be useful when we derive expressions for $P_0(n,t)$ and $P_1(n,t)$.

Let $P_D(n,t)$ be the probability of n arrivals in time t for the deterministic process starting from a random point in the stationary process, $t = 0$. From first principles, we have:

$$P_D(n,t) = \begin{cases} \frac{1}{T}\int_0^T \int_{(n-1)T+x}^{nT+x} \delta(t-u)dudx & \text{if } n > 0, \\ 1 - \frac{t}{T}, & \text{if } n = 0 \text{ and } 0 \le t \le T, \\ 0, & \text{otherwise.} \end{cases} \tag{50}$$

The first term on the right is for the case that there is at least one arrival and that this first arrival occurs at some time x sampled from a uniform distribution on $(0,T)$. The other two terms account

for cases in which there are no arrivals. Taking Z and Laplace transforms, we obtain, after a few manipulations:

$$P_D(z,s) = \frac{(1-e^{-sT})}{s^2T}\left(\frac{z(1-e^{-sT})}{1-ze^{-sT}} + \frac{sT}{1-e^{-sT}} - 1\right). \tag{51}$$

We shall use the above expression when evaluating $P_0(n,t)$ and $P_1(n,t)$. First let us evaluate $P_0(n,t)$. The transition rate from state 0 (inactive) is taken to be r_0 and the transition rate from state 1 is taken to be r_1. We have,

$$P_0(n,t) = e^{-r_0 t}\delta_{n0} + \int_0^t r_0 e^{-r_0 u} P_1(n, t-u)du. \tag{52}$$

This equation results from considering the cases in which there are no state transitions in time t (giving rise to the first term) and in which there is at least one state transition (to state 2). Similarly, for $P_1(n,t)$ we have,

$$P_1(n,t) = e^{-r_1 t}P_D(n,t) + \Sigma_{i=0}^{n}\int_0^t r_1 e^{-r_1 u} P_D(i,u)P_0(n-i, t-u)du. \tag{53}$$

Taking transforms, we obtain after some manipulation:

$$\begin{aligned} P_0(z,s) &= \frac{1}{s+r_0} + \frac{r_0}{s+r_0}P_1(z,s) && (54)\\ P_1(z,s) &= P_D(z, s+r_1) + r_1 P_D(z, s+r_1)P_0(z,s) && (55) \end{aligned}$$

Solving simultaneously we obtain,

$$\begin{aligned} P_0(z,s) &= \frac{1 + r_0 P_D(z,s)}{(s+r_0) - r_0 r_1 P_D(z,s)} && (56)\\ P_1(z,s) &= \frac{P_D(z, s+r_1)(1 + \frac{r_1}{s+r_0})}{1 - \frac{r_0 r_1}{s+r_0}P_D(z, s+r_1)}. && (57) \end{aligned}$$

Following the same procedure we used when deriving expressions for the MMPP, we can obtain the means and variances for the single voice source from the generating functions derived above. Define $N(0,t)$ as the number of arrivals in time t for the voice source. Conditioning this random variable on the state of the voice source at $t = 0$ (active or inactive), we obtain for the variance of $N(0,t)$:

$$\text{VAR(N(0, t) | initially inactive)} = \mathcal{L}^{-1}\left(\frac{d^2}{d^2 z}P_0(z,s) + \frac{d}{dz}P_0(z,s)\right) - \mathcal{L}^{-1}\left(\frac{d}{dz}P_0(z,s)\right)^2 \tag{58}$$

and

$$\text{VAR}(N(0,t) \mid \text{initially active}) = \mathcal{L}^{-1}\left(\frac{d^2}{d^2 z}P_1(z,s) + \frac{d}{dz}P_1(z,s)\right) - \mathcal{L}^{-1}\left(\frac{d}{dz}P_1(z,s)\right)^2. \tag{59}$$

In the case of the MMPP, we were able to analytically obtain the Laplace inverse. In this case, we were unable to do so and we numerically invert the Laplace transform to evaluate the variance. The variance of the number of arrivals generated by the superposition of voice sources is given by equation (49) with the variances for the MMPP replaced by the above equations for the variances of the voice source.

Appendix C

We briefly outline the procedure to obtain the eigenvalues and eigenvectors for the system of differential equations formulated in section 5.

The corresponding eigenvalue problem for the system of differential equations can be formulated as [1, equation (11)],

$$z\boldsymbol{D}\phi = \boldsymbol{M}\phi \tag{60}$$

where $\boldsymbol{D} = diag\{-c, 1-c, 2-c, ..., N-c\}$ and $\boldsymbol{M}$ is the infinitesimal generator for the birth-death process of the number of active sources. The eigenvalues for this system of equations are then obtained as the solution of a set of N quadratics [1, equation (20)]

$$A(k)z^2 + B(k)z + C(k) = 0 \quad k = 0, 1, \cdots, N. \tag{61}$$

The eigenvectors are then obtained by solving a recursive system of equations [15],

$$\begin{aligned} z(i-c)\phi_i &= \lambda(N+1-i)\phi_{i-1} - ((N-i)\lambda + i)\phi_i + (i+1)\phi_{i+1} \quad 0 \le i \le N, \\ \phi_N &= 1. \end{aligned} \tag{62}$$

The solution then follows as outlined in section 5.