Improved Loss Calculations at an ATM Multiplexer

Ness B. Shroff, *Member, IEEE*, and Mischa Schwartz, *Life Fellow, IEEE*

***Abstract*— In this paper we develop a simple and accurate analytical technique to determine the loss probability at an access node to an asynchronous transfer mode (ATM) network. This is an important problem from the point of view of admission control and network design. The arrival processes we analyze are the Markov-modulated Poisson process (MMPP) and the Markov-modulated fluid (MMF) process. These arrival processes have been shown to model various traffic types, such as voice, video, and still images, that are expected to be transmitted by ATM networks. Our *hybrid analytical technique* combines results from large buffer theories and quasi-stationary approaches to analyze the loss probability of a *finite-buffer* queue being fed by Markov-modulated sources such as the MMPP and MMF. Our technique is shown to be valid for both *heterogeneous* and *homogeneous sources*. We also show that capacity allocation based on the popular *effective-bandwidth* scheme can lead to considerable underutilization of the network and that allocating bandwidth based on our model can improve the utilization significantly. We provide numerical results for different types of traffic and validate our model via simulations.**

***Index Terms*—Admission control, bandwidth allocation, effective bandwidth, large deviations, loss probability, Markov models, MMPP, queueing analysis.**

I. INTRODUCTION

THERE IS a large-scale effort being undertaken in both the industrial and academic environments to design and build high-speed asynchronous transfer mode (ATM) networks. These networks are envisioned to support high-speed real-time applications that will significantly impact not only the scientific and engineering communities but also the general population at large. These applications will generate a variety of traffic types such as video, voice, still images, and data, each with their own quality-of-service (QoS) objectives that need to be met by the network. It will be the responsibility of the network admission controller to allow calls to enter the network only if it can provide them their negotiated QoS and not violate the QoS guarantees of existing calls in the network. Accurate traffic modeling and analysis of the QoS parameters in the ATM environment will enable the admission controller to make decisions that ensure the integrity of the traffic sources and are efficient to the network. An important QoS measure that we will study in this paper is the *loss probability.*[1]

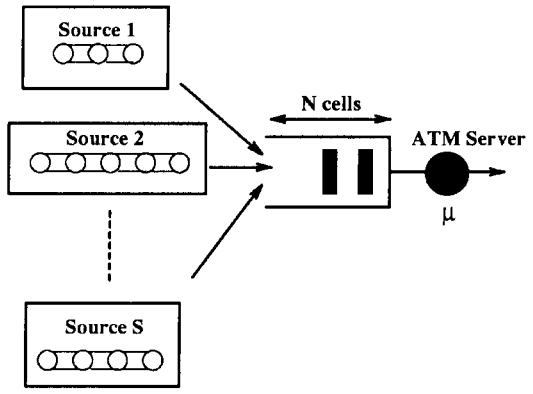


Fig. 1. Sources arriving at a multiplexer.

We will develop a simple yet accurate hybrid model to determine the loss probability at an ATM multiplexer. The model under consideration consists of Markov-modulated sources being served by an ATM multiplexer with (finite) buffer capacity N and link capacity μ, as shown in Fig. 1. Markov-modulated arrival processes find many applications in computer and communication systems. For example, special cases of this model are used to model voice, data, and video traffic sources, and appear in [1], [3], [4], [7], [9], [10], [13], [16], [17], [21], [25], and [32]–[36], to cite but a few. Unfortunately, the *curse of dimensionality* so appropriately named by Bellman in his book *Dynamic Programming* [5] weighs heavily on many branches of applied probability, and queueing theory is no exception. Hence, even for special cases of this arrival process, solving for the loss probability is computationally intensive and impractical, especially when the state space of the aggregate arrival process is large [8], [11]. This is usually the case in high-speed ATM networks since we may expect to have a large number of different types of sources (especially the ones that individually consume low bandwidth) being served by an ATM multiplexer. Thus, improving the computational complexity in such systems is a topic of active research.

Our approach to solving this problem stems from our extensive work in video modeling [26]–[28]. We have been able to determine the probability of loss in an ATM environment for highly correlated traffic sources such as JPEG-encoded video. A quasi-stationary approximation, called the histogram model [31] (or generalized histogram model [26], [27]), was found to be valid because the probability of loss did not significantly

Manuscript received June 28, 1996; revised February 10, 1998; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor D. Mitra. This work was supported in part by the Office of Naval Research under Grant N00014-90-J-1189, and in part by the National Science Foundation under Grant EEC-88-11111, Grant NCR-9624525, and Grant CDA-94-22250.

N. B. Shroff is with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907-1285 USA (e-mail: shroff@ecn.purdue.edu).

M. Schwartz is with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA (e-mail: schwartz@ctr.columbia.edu).

Publisher Item Identifier S 1063-6692(98)05706-9.

[1] We will use the terms *loss probability* and *loss rate* interchangeably in the paper. They basically refer to the long term fraction of cells lost.

decrease when the buffer size was increased beyond a certain range (called the *cell region*). When applying the histogram model to other less-correlated sources, we found that in the cell region the model continued to predict the loss behavior well; however, for large buffer sizes it was not accurate anymore. To overcome this difficulty we have combined the quasi-stationary analysis and results from large buffer theories into one simple model called the *hybrid model*. Using this model we will show that the loss probability at a finite-buffer ATM multiplexer can be efficiently and accurately determined.

The paper is organized as follows. Section II provides a literature overview and the motivation to study our problem. In Section III we describe our problem in detail and define a few parameters to be used in the paper. In Section IV we develop our simple hybrid model to efficiently calculate the loss probability at an ATM multiplexer for Markov-modulated arrival sources. Specifically, we show how to apply this technique for arrival sources such as the Markov-modulated Poisson process (MMPP) and the Markov-modulated fluid (MMF). Finally, in Section V, examples of the applicability of the model are demonstrated and the limitations of the classical effective bandwidth approximations are discussed.

II. Overview

In the literature, the loss probability P_L in a finite buffer queueing system (with buffer size N) is often approximated by $P(Q \geq N)$, the tail of the queue length distribution, in the corresponding infinite buffer queueing system.

For infinite buffer queueing systems, it has been shown in considerable generality that $P(Q \geq N)$ is asymptotically exponential, i.e.,

$$P(Q \geq N) \sim Ae^{-\eta N} \quad \text{as } N \to \infty. \tag{1}$$

Here η is a positive constant called the *asymptotic decay rate*, A is a positive constant called the *asymptotic constant*, and $f(x) \sim g(x)$ means that $\lim_{x\to\infty} f(x)/g(x) = 1$. Moreover, the asymptotic decay rate in a *finite buffer* queueing system is the same as η in the equivalent *infinite buffer* system (which is one more reason why the tail is often used to approximate the loss).

However, it is usually computationally intensive to determine A, the asymptotic constant in (1); hence, as we see in the next section, it is sometimes ignored.

A. Effective Bandwidth Approximation

The classical effective bandwidth approximation assumes that the constant A in (1) is 1, i.e., if we write (1) as

$$P(Q \geq N) \sim e^{-\eta N + \log(A)} \quad \text{as } N \to \infty$$

then

$$P_L \approx P(Q \geq N) \approx e^{-\eta N} \quad \text{as } N \to \infty. \tag{2}$$

Although this approximation is logarithmically similar to the exact loss probability [14], often the constant A can be a fairly small multiplicative factor such as 10^{-7}. In this case the approximation would certainly not be useful for any practical loss probabilities. One of the main reasons why effective bandwidth has become very popular in the literature is that it provides an easy way of allocating bandwidth independent of the number of sources being multiplexed [12], [15], [18], [19]. For example, consider the following bandwidth allocation problem.

For each source i, a constraint on the probability of loss P_L^i [as defined by (2)] is given by

$$P_L^i \leq \epsilon. \tag{3}$$

Then for a total capacity μ, how should the bandwidth be allocated such that the above constraint is met for each source?

Effective bandwidth provides a relatively simple answer to this question. It says that we can look at each source in isolation being fed by a queue with a certain capacity (that can be varied). Then, for source i, we find the minimum capacity C_i such that the constraint given by (3) is met. The effective bandwidth approximation further states that as long as

$$\sum_i C_i \leq \mu \tag{4}$$

none of the constraints $P_L^i \leq \epsilon$ will be violated. In other words given that the bandwidth required to meet the needs of each source is C_i, the *effective bandwidth* of all the sources combined is simply $\sum_i C_i \triangleq C_{\text{eff}}$. Therefore, in the bandwidth allocation problem, the only computationally intensive part is to determine the capacity C_i for each individual source. This is why effective bandwidth appears to be such a tempting method to use for bandwidth allocation. However, it is the simplicity of (4) which also exposes its main weakness, i.e., that since the effective capacities add, it means that statistical multiplexing gain, which is so important to the success of these high-speed ATM networks, is not exploited.

A little thinking will convince the reader that the only sources for which the capacities add as in (4) are Poisson sources. For all sources that are more variable than the Poisson source (most Markov-modulated sources will fit into this category), adding the capacities in this way will prove to be conservative and wasteful of network resources since the aggregate source will be "smoother" than the original sources. Moreover, if the sources are less variable than Poisson, adding the capacities would result in the system being overutilized and lead to violations of the negotiated QoS parameters [6], [26].

In Section V, using numerical results, we will demonstrate just how poorly effective bandwidth can perform. We next describe Markov-modulated models in more detail, in particular the MMPP and the MMF models, and provide the motivation for studying the loss behavior for these arrival processes. *It should be noted here that although most works in the literature have focused on using the tail probability (in an infinite buffer system) to approximate the loss in a finite buffer system, such an approach could result in significant overestimation of the loss probability.* To avoid this problem, our method will be to directly estimate the loss probability for Markov-modulated arrival processes.

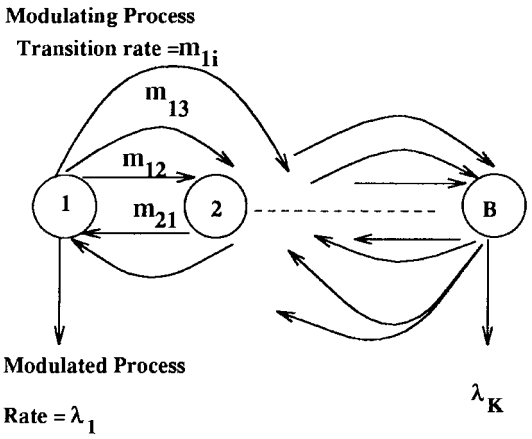


Fig. 2. A Markov-modulated source having B states. m_{ij} correspond to the transition rates from state i to state j.

B. Markov-Modulated Models

A Markov-modulated source is governed by an underlying continuous Markov chain, with state space Σ, which determines the current state of the source. An example of such a source is shown in Fig. 2. In each state i, $i \in \Sigma$, the source transmits information at a rate λ_i according to a stochastic process which we will call the *modulated process*. The sojourn times for each state i are exponentially distributed with mean $-1/m_{ii}$. When each sojourn time is over, the Markov chain moves to a state $j \neq i$ with probability $-m_{ij}/m_{ii}$. Hence, m_{ij} is often called the transition rate from state i to state j.

We next describe one of the most well studied Markov-modulated processes, the MMPP, in which the modulated process is Poisson.

1) The MMPP: The MMPP has been widely used to characterize different types of traffic such as voice, video, and still images [4], [6], [7], [11], [12], [16], [26]. In this model cells arrive at a multiplexer according to a Poisson process whose intensity depends on the state of a Markov process. The fact that the modulated process is assumed to be Poisson allows for mathematical tractability. Moreover, the Poisson approximation for the modulated process of the MMPP is fairly good when a large number of Markov-modulated sources are multiplexed, as is expected to be the case in an ATM environment.[2]

Heffes and Lucantoni [16] have studied the performance of a statistical multiplexer whose inputs consist of packetized voice sources and data and the server is allowed to be general (MMPP/G/1 system). The superposition is approximated by a correlated MMPP which is chosen such that several of its statistical characteristics identically match those of the superposed process, and matrix analytic methods are used to evaluate system performance measures. This technique is shown to have good results in determining the mean delay in the queue but is not as accurate in determining the queue delay distribution. In [4] Baiocchi *et al.* provide a technique which studies multiplexed ON–OFF sources by approximating the aggregate input process by means of a suitably chosen two-state MMPP. Their model provides good insight but is unfortunately limited to the multiplexed ON–OFF case. In [6], Choudhury, Lucantoni, and Whitt provide an interesting *three-term approximation* to determine the tail probability of the waiting time W for independently identically distributed (i.i.d.) ON–OFF sources, i.e.,

$$P(W \geq x) \approx \alpha_1 e^{-\eta_1 x} + \alpha_2 e^{-\eta_2 x} + \alpha_3 e^{-\eta_3 x} \qquad (5)$$

where α_1 and η_1 are the asymptotic constant and the asymptotic decay rate, respectively, in (1), while the other parameters are chosen to match the different parameters and moments of the distribution $P(W \geq x)$. This approximation for the tail of the MMPP can be computed much more easily than the exact tail, but it still suffers from the computational complexity involved in determining the asymptotic constant α_1.

2) MMF: Another Markov-modulated source that has been extensively used to model various types of traffic is the MMF source. In the MMF information is generated and processed as a continuous flow (fluid) at a rate which depends on the state of the Markov process. The model gained widespread popularity as a result of the pioneering work by Anick, Mitra, and Sondhi [2]. The advantage of this model over traditional queueing models is that the numerical complexity is independent of the buffer size [11]. However, unlike the case of the MMPP, where the discrete nature of the cells is preserved, the fluid is unable to capture the effect of cell variability. Hence, the fluid model is usually inaccurate for small buffer sizes and can be viewed as an approximation to the MMPP for large buffer sizes. Furthermore, as in the case of the MMPP, when a large number of sources are being multiplexed, the computational complexity to estimate either the tail or loss probability can become prohibitive. In [9] the authors provide a method to estimate the asymptotic constant in (1) based on the Chernoff bound. This method, called the CDE approximation, could potentially be applied to both homogeneous and heterogeneous sources. However, in [9] numerical results are only provided for identically distributed sources.

III. Problem Definition

The Markov-modulated processes such as the MMPP and the MMF provide a rich stochastic framework to model many different types of traffic sources. However, as we have seen, simpler approximate techniques need to be developed in order to reduce the computational burden.

For analyzing the loss probability, the system we consider consists of S Markov-modulated sources being served by an ATM multiplexer with buffer size N and link capacity μ, as shown in Fig. 1 earlier. We next define a few parameters to be used in our analyses:

σ^s $\triangleq$ state space of source s;
$B^{(s)}$ $\triangleq$ total number of states corresponding to source s;
$\lambda_j^{(s)}$ $\triangleq$ arrival rate generated by source s when source s is in state j;

[2] It should be emphasized here that assuming that the entire arrival stream is Poisson is a *terrible* approximation for such highly variable sources [26]. However, approximating only the modulated process as Poisson turns out to provide good results, especially when there are a large number of sources being multiplexed.

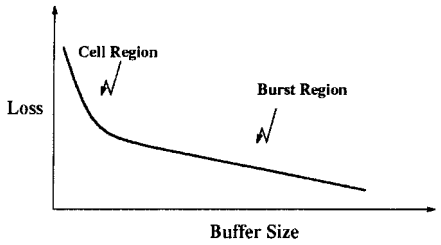


Fig. 3. Cell and burst regions. The loss plotted on the y-axis is on a log scale while the buffer size plotted on the x-axis is on a linear scale.

- $\vec{\lambda}^{(s)} \triangleq \{\lambda_j^{(s)} | j \in \sigma^s\}$—arrival rate vector of source s;
- $m_{ij}^{(s)} \triangleq$ transition rate from state i to state j for source s;
- $M^{(s)} \triangleq$ *infinitesimal generator* of the underlying Markov chain of source s;
- $P_i^{(s)} \triangleq$ stationary probability of source s being in state i;
- $\vec{P}^{(s)} \triangleq \{P_i^{(s)} | i \in \sigma^s\}$, the stationary probability vector corresponding to state s;
- $\lambda_i \triangleq$ arrival rate corresponding to state i of the histogram of the *aggregate* source;
- $P_i \triangleq$ probability of being in state i corresponding to the histogram of the *aggregate* source;
- $B \triangleq$ total number of states of the histogram of the aggregate source;
- $E(\lambda) \triangleq$ (average) arrival rate of the aggregate source.

We will develop a simple analytical technique to determine the probability of loss at a *finite-buffer* ATM multiplexer. The sources that we will consider in this paper are the MMPP and the MMF. Our goal is to be able to efficiently handle *homogeneous* (i.i.d.) as well as *heterogeneous* (independent, but not identically distributed) sources being multiplexed. We next describe our hybrid model.

IV. Hybrid Model

It is well known that for Markov-modulated arrival processes and deterministic servers there are typically two main regions in which increasing the buffer size reduces the cell loss—the "cell region" and the "burst region," as depicted in Fig. 3. In the cell region the main component of the loss rate is the cell variability within the modulated process. As the buffer size is increased, this variability gets absorbed and the loss rapidly decreases. In the burst region the loss due to cell variability is negligible and the loss is caused mainly due to the fact that the rate in one or more of the states is greater than the link capacity μ (in other words, the loss is caused by the overload states). Let δ be the asymptotic slope of the burst region; then it follows that δ is simply the negative of the *asymptotic decay rate* η shown in (1), i.e., $\delta = -\eta$. For the purpose of our hybrid model we assume that the slope of the entire burst region is δ, i.e., the loss in the burst region can be modeled by a single negative exponential with parameter δ [similar to (1), but the constant multiple may be different]. In Section IV-C we will see that δ corresponds to the dominant eigenvalue of a matrix.

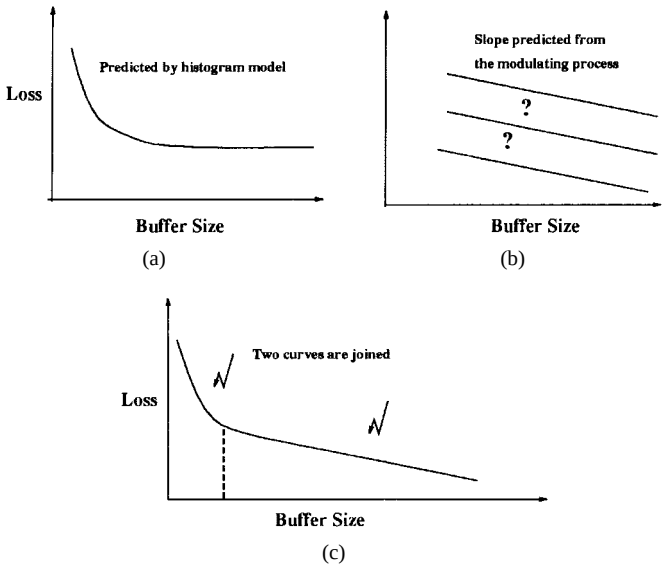


Fig. 4. (a) Determine the loss using the histogram model. It will be accurate only in the cell region. (b) Determine the slope of the burst region which will be determined by the modulating process. (c) Connect the two curves at the point at which the negative of the slope of the cell region is smaller than that of the burst region.

Our approach to solving for the loss probability takes advantage of the fact that the loss in the cell region and the slope of the loss in the burst region can be easily determined. We then simply determine the point of transition and connect the two curves. We describe our methodology in three different steps. The steps are also graphically illustrated in Fig. 4.

Step 1: In this step we determine $P_{L_{\text{cell}}}(N)$, the probability of loss in the cell region for buffer size N. The way to solve for the loss probability in the cell region is to let the time spent in each state go to infinity, i.e., let the transition rates of each source s go to zero ($m_{ij}^{(s)} \to 0$ for all $i, j \in \sigma^s$). In this way, if the process is in an overload state, it will continue to remain in that state forever, and increasing the buffer size will not significantly impact the probability of loss. In this situation the probability of loss can be determined by a well-known quasi-stationary approach called the histogram decomposition approximation [31], or the generalized histogram model (GHM) [26], [27]. The idea is as follows. A rate histogram is determined for each Markov-modulated source s. This histogram can be represented by the doublet $(\vec{\lambda}^{(s)}, \vec{P}^{(s)})$ defined in Section III. The x-axis of the histogram corresponds to the arrival rate in state i while the y-axis corresponds to the stationary probability of being in state i ($i \in \sigma_s$). Since the sources are independent, the histogram of the multiplexed (or aggregated) source, with average rate $E(\lambda)$, can be determined as an S-fold convolution of the respective histograms. (If the number of sources being multiplexed is large, by the central limit theorem we can use the Gaussian approximation and then need only to add the mean and variances of the individual histogram to perform the equivalent convolution. The aggregate histogram can then be drawn from the Gaussian distribution.) The states of the aggregate source are combined such that the total number of states of the aggregate histogram (λ_i, P_i) is no greater

than some maximum number B. After extensive simulations we have found that 20 states are sufficient to model the aggregate process, and further increasing the number of states does not result in a significant change in the probability of loss calculations. This is consistent with the results in [27] and [31].

Using the aggregate histogram and applying the loss formula from [27], we have

$$P_{L_{\text{cell}}}(N) = \frac{1}{E(\lambda)} \sum_{i=1}^{B} P_{L_i}(N) P_i \lambda_i \tag{6}$$

where $P_{L_i}(N)$ is the probability of loss conditioned on the arrival being in state i.

The probability of loss $P_{L_{\text{cell}}}(N)$, as predicted using the histogram model, will level off with increasing N [Fig. 4(a)] for any general Markov-modulated source, as long as the peak rate of this source is greater than the capacity (which is the situation we are interested in).[3] In other words, as long as

$$\max_i \lambda_i > \mu$$

$P_{L_{\text{cell}}}(N)$ asymptotically approaches a value which is constant in N. The proof of this property is shown in [29]. We will make use of this property to determine the entire loss-buffer curve.

Step 2: The next step is to determine the buffer size N_0, shown in Fig. 4(c), which is the point at which the transition takes place between the cell and burst regions. N_0 is the maximum buffer size at which the slope of the cell region is still steeper than the slope of the burst region. Hence, since the slope of the cell region at some buffer size x is (approximately) $(\log[P_{L_{\text{cell}}}(x)] - \log[P_{L_{\text{cell}}}(x-1)])/[x-(x-1)]$

$$N_0 = \max N$$

such that

$$\log[P_{L_{\text{cell}}}(N)] - \log[P_{L_{\text{cell}}}(N+1)] \geq \eta.$$

Here $\eta = -\delta$ is the asymptotic decay rate.

Step 3: In our model $P_{L_{burst}}(N)$ the loss probability in the burst region as a function of the buffer size N is of the form of a negative exponential with parameter (slope) δ. Also from Step 2 we can determine the buffer size N_0 and the probability of loss $P_{L_{\text{cell}}}(N_0)$ at which the transition from the cell region to the burst region takes place. Therefore

$$P_{L_{burst}}(N) = P_{L_{\text{cell}}}(N_0) e^{\delta(N-N_0)}.$$

To determine the entire loss curve, all we now need to find is $P_{L_i}(N)$, the probability of loss conditioned on the arrival being in state i, and δ, the slope of the burst region.

A. Determining $P_{L_i}(N)$ for MMPP and MMF Models

We assume that the arrival rate in state i of the aggregate source histogram is λ_i, the probability of being in state i is P_i, and the total number of states is B. We next determine $P_{L_i}(N)$, the probability of loss conditioned on state i of the aggregate histogram, for the MMF.

[3]This condition is often considered the nontrivial condition.

The fluid model can be thought of as approximating a Markov-modulated deterministic process (MMDP), where the cells corresponding to a given state arrive equispaced. Hence, $P_{L_i}(N)$ is the loss probability in a $D/D/1/N$ system with arrival rate λ_i and service rate μ. It can be readily shown that for such a system, the probability of loss $P_{L_i}(N)$ is given by

$$P_{L_i}(N) = \begin{cases} 1 - \dfrac{\mu}{\lambda_i}, & \text{if } \dfrac{\lambda_i}{\mu} > 1 \\ 0, & \text{otherwise} \end{cases} \qquad \text{for all } N. \tag{7}$$

Since $P_{L_i}(N)$ can be determined so compactly, we apply (6) to obtain

$$P_{L_{\text{cell}}}(N) = \frac{1}{E(\lambda)} \sum_{\substack{i=1 \\ \lambda_i > \mu}}^{B} \lambda_i P_i \left(1 - \frac{1}{\rho_i}\right) \tag{8}$$

where $\rho_i = \lambda_i/\mu$. Also, since the fluid source assumes a constant arrival rate, the cell region is only a point on the y-axis, and the cutoff point is given by $N_0 = 0$. Hence, the probability of loss in the burst region (which in this case turns out to be the overall cell loss) is given by

$$P_L = P_{L_{burst}}(N) = \frac{1}{E(\lambda)} \sum_{\substack{i=1 \\ \lambda_i > \mu}}^{B} \lambda_i P_i \left(1 - \frac{1}{\rho_i}\right) e^{-\delta N}. \tag{9}$$

Here, it is instructive to compare (9) to the CDE approximation in [9], since both of them are single-exponential approximations. The exponential decay rate is the same in both cases. The difference in the two approaches is in evaluating the probability when $N_0 = 0$ (i.e., the terms preceding the exponential). In the CDE approach the authors propose to compute this y-intercept, what they call the "loss in the bufferless system," by determining the stationary probability that the arrival rate of the aggregate source exceeds the capacity μ. They estimate this probability[4] using the Chernoff theorem, and provide a refinement to approximate this probability when $\mu \to \infty$. A detailed comparison of our scheme with the CDE approximation has been conducted at Lucent Bell Laboratories by Pancha *et al.* [22]. They have found that the hybrid model is computationally simpler and typically provides more accurate estimates of the loss probability than the CDE approximation. Numerical studies on our part have resulted in the same conclusion, which is to be expected since the CDE approximation focuses on estimating the tail probability, while our approximation focuses on estimating the loss probability in a finite-buffer system.

In the case of the MMPP the arrival process corresponding to each state of the aggregate histogram is Poisson. Hence, $P_{L_i}(N)$ can be determined by solving for the loss in an $M/D/1/N$ system with arrival rate λ_i and service rate μ. This can be accomplished by using the standard technique of approximating the service rate by an Erlang distribution, and solving the corresponding equations derived from the continuous-time Markov chain representing the number of

[4]It should be noted that the "loss in the bufferless system" in [9] can be directly estimated by using our approach by setting $E(\lambda) = 1$ and $\lambda_i = 1$ in the prefactor term $\sum_{\substack{i=1 \\ \lambda_i > \mu}}^{B} \lambda_i P_i (1 - 1/\rho_i)$ of (9).

packets in the system. We have also derived a compact exact solution for the loss in an $M/D/1/N$ system [23], as in (9a), shown at the bottom of the page.

Once $P_{L_i}(N)$ is determined, we can follow Steps 2 and 3 to get P_L, the overall loss in the queueing system. Since the loss in the $M/D/1/N$ case is more computationally complex than the $D/D/1/N$ case, solving for the loss in the MMPP case requires more computations than in the MMF case. However, the buffer size(s) for which we need to calculate this value of loss will be relatively small (only in the cell region) and, therefore, this increase in complexity over the fluid case will not be that significant. Moreover, if necessary, one could use a fast approximate technique for the loss in an $M/D/1/N$ system (such as the one developed in [26] which has provided a good match with simulations over a broad spectrum of traffic parameters).

Also note here that the limiting value for $P_{L_{\text{cell}}}(N)$, under fairly general conditions, equals the value of $P_{L_{\text{cell}}}(N)$ for the MMF case given by (8). This is shown in [29].

B. A Brief Discussion on the Convexity of $\log P_{L_{\text{cell}}}(N)$

In the context of determining N_0 in the MMPP case, it is instructive to study whether $f(N) \triangleq \log P_{L_{\text{cell}}}(N-1) - \log P_{L_{\text{cell}}}(N)$ is a decreasing function of N (or, equivalently, whether the values of $\log P_{L_{\text{cell}}}(N)$ fall on the graph of a convex function of N). We conducted extensive experiments in which we used different parameters for the MMPP and found that $f(N)$ was, in fact, a decreasing function of N in all of the cases tested. However, it was difficult to do a completely systematic study in this case since there were many parameters one could vary, e.g., the number of states of the MMPP, the state transition rates, the utilization, etc. Hence, we also tried a different approach. Note that for an MMPP arrival process, $P_{L_i}(N)$ for any i is obtained by calculating the loss probability in an $M/D/1/N$ system. It can easily be shown that if the values of $\log P_{L_i}(N)$ fall on the graph of a convex function of N, then the values of $\log P_{L_{\text{cell}}}(N)$ also fall on the graph of a convex function of N [20]. This means that we only need to check whether the loss probability in an $M/D/1/N$ system is a convex function of N. For experimental verification, we varied ρ from 0.3 to 0.99 in 0.01 increments, and for each ρ we varied the buffer size from five cells to 1000 cells. We found that in almost all of the cases (except for a few cases at very low utilization) the log of the loss probability fell on a convex curve. Further, even for those cases, the points were very close to falling on a convex curve. This implies that inaccuracies in the hybrid model will not be due to $\log P_{L_{\text{cell}}}(N)$ not being a convex function of N.

C. Determining the Slope of the Burst Region δ

The (asymptotic) slope of the burst region δ is independent of the modulated process; hence, the following analysis is valid for any Markov-modulated source. Determining δ (or equivalently η, the asymptotic decay rate) has received considerable attention in the literature [2], [11], [12], [14]. Here we present an outline based on the work in [11 and [12].

Consider again the system of Fig. 1, where the statistical multiplexing system consists of a buffer which is supplied by various statistically independent Markov-modulated sources served by a channel of constant capacity μ. Let the aggregate Markov-modulated source with state space Σ generate information at a rate λ_i in state $i (i \in \Sigma)$. Further, let $\vec{\lambda} \triangleq \{\lambda_s | s \in \Sigma\}$ and the rate matrix $\Lambda \triangleq \text{diag}(\vec{\lambda})$. Let M denote the irreducible generator of the aggregate source. Then the aggregate source is characterized by $(M, \vec{\lambda})$. Let I denote the identity matrix; then it can be shown that δ is simply the smallest negative eigenvalue of the matrix $M(\Lambda - \mu I)^{-1}$ [11]. However, this matrix could easily be composed of hundreds to tens of thousands of rows depending on the number of sources multiplexed. Fortunately, the complexity of the calculation can be substantially reduced if we do not lump all of the sources together [11], [12].

Thus, suppose that there are S sources characterized by $(M^{(k)}, \vec{\lambda}^{(k)})$, $k = 1, \cdots, S$ (where $M^{(k)}$ and $\vec{\lambda}^{(k)}$ are defined as in Section III). Let $\Lambda^{(k)} = \text{diag}(\vec{\lambda}^{(k)})$. It can then be shown that δ can be found by solving for the root of the equation (see Elwalid [11]),

$$g_1(\delta) + g_2(\delta) + \cdots + g_S(\delta) = \mu, \tag{10}$$

where $g_k(\delta)$, the eigenvalue with the greatest real part, is found by solving the inverse eigenvalue problem

$$g_k(\delta)\phi^{(k)}(\delta) = \phi^{(k)}(\delta)\left(\Lambda^{(k)} - M^{(k)}/\delta\right), \qquad k = 1, \cdots, S. \tag{11}$$

Note that $g_k(\delta)$ are strictly decreasing functions with values between the mean and peak rates of source k [12], thus assuring a unique solution to (10). We can solve for (10) by using standard iterative root-finding techniques (such as the Newton's method), which have been found to work very efficiently. Next we consider two special cases.

1) Homogeneous Sources: If all of the sources are homogeneous, then $g_1(\delta) = g_2(\delta) = \cdots = g_S(\delta)$ and, therefore, (10) reduces to

$$g_1(\delta) = \frac{\mu}{S}.$$

$$P_{L_i}(N) = \begin{cases} \dfrac{\rho_i}{1+\rho_i}, & \text{if } N = 1 \\ 1 - \dfrac{1 - \left(1 + \rho_i \left(\displaystyle\sum_{k=1}^{-1+\mathrm{N}} \frac{(-1)^{-1-k+\mathrm{N}} e^{k\rho_i} (k\rho_i)^{-1-k+\mathrm{N}}}{(-1-k+\mathrm{N})!}\right)\right)^{-1}}{\rho_i}, & \text{if } N > 1 \end{cases} \tag{9a}$$

Working backward, and substituting the above equation in (11), we get

$$\frac{\mu}{S}\phi^{(1)}(\delta) = \phi^{(1)}(\delta)(\Lambda^{(1)} - M^{(1)}/\delta)$$
$$\phi^{(1)}(\delta)\left(\Lambda^{(1)} - \frac{\mu}{S}I\right) = \phi^{(1)}(\delta)M^{(1)}/\delta$$
$$\phi^{(1)}\delta = M^{(1)}\left[\Lambda^{(1)} - \frac{\mu}{S}I\right]^{-1}. \qquad (12)$$

Therefore, δ can be found directly by solving for the dominant eigenvalue of $M^{(1)}[\Lambda^{(1)} - \mu/SI]^{-1}$. In other words, if there are S Markov-modulated sources being served by an ATM multiplexer, we can find δ of the multiplexed system by solving for the slope of one of the sources in isolation with the capacity decreased by a factor of S.

2) Groups of Homogeneous Sources: In many cases the sources that are fed to the multiplexer may not be all homogeneous or heterogeneous but something in between, i.e., there may be, for example, K heterogeneous groups each containing $m_i (i = 1, \cdots, K)$ homogeneous sources, such that $\sum_{i=1}^{K} m_i = S$. In this case too we can reduce the complexity in determining δ over the completely nonhomogeneous case by rewriting (10) as

$$m_1 g_1(\delta) + m_2 g_2(\delta) + \cdots + m_K g_K(\delta) = \mu \qquad (13)$$

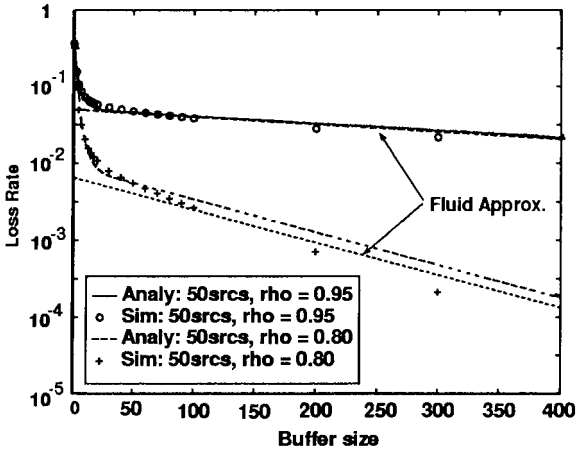
and appropriately modifying (11). Then δ can be found by applying Newton's method (or some other root finding technique) to (13).

V. Discussion and Results

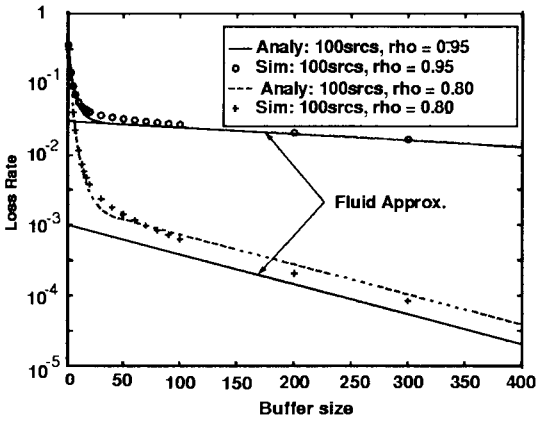
We will now illustrate the salient features of the hybrid approximation. The experiments involve a number of sources that can be analytically modeled by an MMPP that are served by an ATM multiplexer as in Fig. 1. The fluid model is identical to the MMPP except that it is unable to capture the cell variability due to multiplexing in the *cell region*. Hence, to not overly clutter the figures, we mostly focus on analytical comparisons here with the MMPP sources. However, since large buffer behavior is important for many applications, we also show some examples to illustrate the accuracy of the simple fluid approximation given by (9).

Since simulation results cannot validate our model at very low loss probabilities, we use the experimental parameters in [6] (where they have computed the exact tail) for such validation. We show that the hybrid model captures the loss probability quite well. Further, we show that the asymptotic approximation from (1) may provide a poor lower bound to the loss under some circumstances. We also show that the effective bandwidth approximation is quite conservative and we develop a simple *rule of thumb* for determining how good it will be for a particular set of traffic parameters.

Example 1: In Fig. 5(a) and (b) we show analysis versus simulation results for multiplexed voice sources. For our simulation results, each voice source is modeled by a two-state Markov-modulated ON–OFF process, in which the amount of time spent in the ON and OFF states is exponentially distributed with mean 0.4 and 0.6 s, respectively. With normal speech encoding, the voice process is sampled every 125 μs, with samples encoded into 8 b [24]. Hence, in the ON-state ATM cells are generated periodically at a rate of 170 cells/s. For the analytical model we consider each voice source as a two-state MMPP with the same parameters as the simulation. The loss probability is then determined via our hybrid technique. In Fig. 5(a) and (b) we show results for 50 and 100 multiplexed voice sources, respectively. It can be observed that the simulation and analytical results match quite well. Notice that, for the same utilization (denoted by "rho" in the figure), the loss–buffer curve in the burst region for 50 sources is parallel to the equivalent curve for 100 sources. The reason for this behavior is that the sources are homogeneous; hence, at the same utilization, the slope δ should be constant with respect to the number of sources multiplexed. Of course, as can be observed, statistical multiplexing does help provide a lower loss–buffer curve for the "100"-source case over the "50"-source case. Also observe that the fluid approximation, predicted using the hybrid analytical technique, is quite good except for very small buffers. In fact, we have tested the fluid approximation for many other sources as well, and have found it to be usually quite accurate for buffer sizes larger than about 100 cells. Also note that, as would be expected, the analytical fluid and MMPP approximations are closer when the slope of the burst region is shallower.



(a)



(b)

Fig. 5. (a) 50 multiplexed voice sources and (b) 100 multiplexed voice sources. In the figures "rho" corresponds to the link utilization.

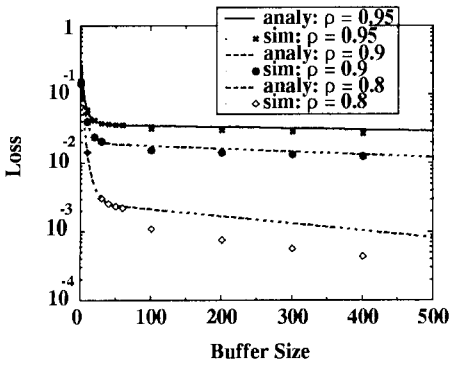


Fig. 6. Loss versus buffer size for JPEG video sources that are randomly smoothed and modeled as MMPP sources. These sources are so chosen that they have a relatively fast change of rates.

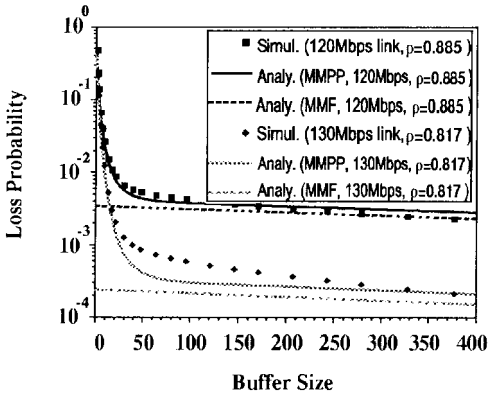


Fig. 7. A large multiplexed system: 1100 voice sources and 45 video sources. The video sources are divided into three heterogeneous groups. The arrival rate for each voice source is 170 cells/s and the arrival rate for each video source is 3538 cells/s (1.5 Mb/s).

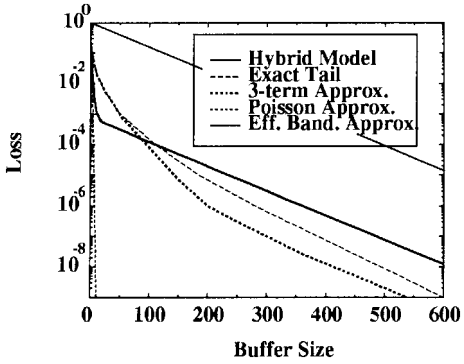


Fig. 8. MMPP source—tail and loss probability methods. Comparing our analytical *hybrid model* for determining loss to the *exact tail distribution*, the *three-term approximation model*, the *Poisson* model, and the *effective bandwidth model* for a system with 24 multiplexed ON–OFF sources.

Example 2: In Fig. 6 we consider intraframe JPEG-encoded video sources being statistically multiplexed at an ATM link. *Actual traces of the encoded movie Star Wars have been used for the simulations.* Five video sources were specifically chosen with the shortest correlation intervals so that one could observe a decrease in the burst region with increasing buffer sizes.[5] Each source is *randomly smoothed* over one frame before being multiplexed, i.e., the cells in each frame are spread over that frame using a uniform distribution [31]. For the simulation, each source is obtained from different parts of the movie and, hence, are not identically distributed.[6] For the analysis, each source is modeled by a 20-state MMPP, and the generator matrix, rate, and the probability vector are determined directly from the real sources, as in [30]. We use our hybrid technique to model the resultant system that is formed by multiplexing heterogeneous MMPP sources. In this case the total number of states of the resultant aggregate Markov chain could be as high as 20^5 (3.2 million states!). However, we combine the states as was described in Section IV-A into a total of 20 states. Once again, the results indicate a fairly close match between the analysis and simulations.

Example 3: In the next example, shown in Fig. 7, we consider an experiment involving a large number of sources being served by an ATM multiplexer. The arrival process consists of 1100 voice and 45 video sources. The voice sources have the same statistics as in Fig. 5. The video sources can be classified into three heterogeneous groups; each group consists of the superposition of 15 homogeneous 20-state MMPP sources. The sources in the three different groups derive their statistics from MPEG-1 sequences corresponding to three different movies following Skelly's matching procedure [30]. For the analysis, the slope of the burst region is determined by using (13). The arrival rate for each voice source is 170 cells/s (64 kb/s), and the arrival rate for each video source is (approximately) 3538 cells/s (1.5 Mb/s). Since the arrival rates are fixed, the utilization in Fig. 7 is varied by changing the link rate μ. The analysis versus simulation curves are plotted for utilizations of 0.89 and 0.82, and match quite well (especially considering that this is a large multiplexed system). Once again, the MMF approximation is quite good, except for fairly small buffers.

Example 4: One of the problems with simulations is that it is currently infeasible using today's high-performance computers to obtain values of loss in the range of 10^{-9} or lower (which is the loss that is often expected for many applications in ATM) without using some form of importance sampling techniques. Hence it is difficult to validate our hybrid model at these low loss probabilities. Therefore, for our next example, in Fig. 8 we consider a scenario from [6] in which the exact tail probabilities have been analytically computed. For examples 4–6, the mean service time of a cell is normalized to 1. Further, cells are generated by each source in the ON period at a peak rate p.

For this experiment, 24 sources are served by an ATM multiplexer. The mean ON time of each source is measured in the mean number of cells (60 cells) that are generated during the ON duration at a peak rate p. Similarly, the mean OFF time duration of each source is measured in the mean number

[5] In [27] we show that JPEG sources typically tend to be so highly correlated that even large increases in buffer sizes have marginal decreases in the probability of loss. This is why it is so convenient to analyze them using the histogram decomposition approximation. Even though care was taken to pick five sources with the shortest correlation, it can be observed from Fig. 6 that the burst region is quite shallow.

[6] This is especially true for variable bit-rate video traffic which is nonstationary.

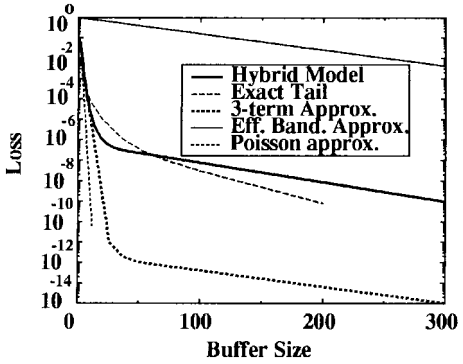


Fig. 9. MMPP source—tail and loss probability methods. Comparing our analytical *hybrid model* for determining loss to the *exact tail distribution*, the *three-term approximation model*, the *Possion* model, and the *effective bandwidth model* for a system with 60 multiplexed ON–OFF sources.

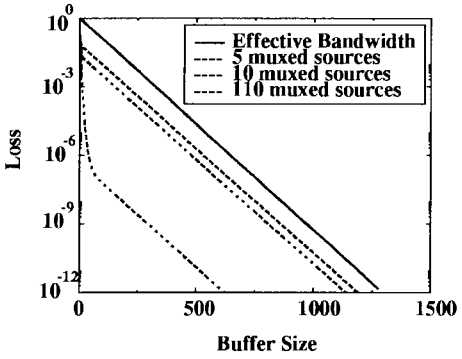


Fig. 10. Multiplexed ON–OFF sources, $\rho = 0.7$. This figure shows that the effective bandwidth approximation can be quite poor even for reasonably high utilizations.

of cells (600 cells) that could have been generated at a rate p during the OFF state duration. The peak rate $p = 0.1375$, and 24 such sources have been multiplexed resulting in a link utilization of 0.3. The *three-term approximation* (5) and the exact tail are reproduced in Fig. 8 directly from [6]. In the "small buffer" region we would expect the tail to be an upper bound to the loss, and the hybrid model predicts just that. In the large buffer regime, for such small utilization, one would expect the loss to be almost the same as the tail; however, here we find that the hybrid model predicts a conservative bound to the tail. The reason for this is that in this experiment the other eigenvalues of the matrix $M(\Lambda - \mu I)^{-1}$ discussed in Section IV-C play a more significant role in determining the slope of the burst region and it requires a much larger buffer for the slope to approach that of the dominant eigenvalue. This effect is not taken into account in our hybrid scheme which is based on the cell-burst region model of Fig. 3. Still, our predicted loss curve is only conservative by at most an order of magnitude, which is considered reasonably accurate. (Remember that at high utilizations the tail approximation is often over an order of magnitude more than the loss.) The *three-term approximation* turns out to be a lower bound but also predicts the tail quite well. The effective bandwidth approximation is off by about four orders of magnitude and the Poisson approximation is completely invalid.

Example 5: In the next example (again taken from [6]), shown in Fig. 9, the number of sources being multiplexed is now increased to 60 while keeping the slope of the burst region and the utilization the same. Therefore, we decrease the peak rate of each source to 0.055 and the link rate is kept the same at one. For the slope of the burst region to be the same, the average number of cells generated in the ON period is still kept at 60 cells and the average number of cells that would have been generated in the OFF period (if the source were transmitting at its peak rate) is still kept at 600 cells. Once again, the *three-term approximation* and the exact tail are reproduced in Fig. 9 from [6]. As expected, because the number of multiplexed sources is increased from the previous example, while the utilization is kept constant, the loss probability for a particular buffer size is smaller in this case. (In other words, the asymptotic constant A of the tail probability in (1) will be smaller in this example than in the previous case.) From Fig. 9 we can also tell that our hybrid model once again predicts the exact tail reasonably well. On the other hand, the *three-term approximation* does not do as well and is a lower bound, off by about four orders of magnitude. The effective bandwidth approximation is extremely conservative, off by six orders of magnitude! It is interesting to note here that the asymptotic constant A of (1) is of the order of 10^{-12} and that using (1) for the tail probability would give us a straight line coinciding with the burst region of the *three-term approximation* [6]. This suggests that the asymptotic approximation in this particular case is also quite bad for the range of probabilities that we are interested in for ATM networks (typically between 10^{-6}–10^{-9}).

Example 6: The effective bandwidth approximation is often considered to be reasonably good at higher utilizations [6], [12]. We have so far only compared it when the link utilization is relatively low. However, with the help of this next example, we hope to eliminate another misconception about this approximation. In this example we keep the link utilization constant at 0.7, while we vary the number of homogeneous sources served by the link. The utilization is kept constant by varying the peak rate of the ON–OFF sources to give different loss versus buffer curves. The average number of cells generated in the ON and OFF time of each source is kept at 100 cells. Remember that the effective bandwidth approximation only depends on the value of δ, the slope of the burst region, which is independent of the number of homogeneous sources being multiplexed. Now, in Fig. 10 we find that when the number of sources being multiplexed at the link is very small (only five sources), the effective bandwidth approximation gives a reasonable upper bound to the loss. However, as the number of multiplexed sources is increased, the effective bandwidth approximation becomes worse. When 110 sources utilize the link, the approximation is off by a factor of seven orders of magnitude. Similar results can also be shown at higher utilizations. The following is an approximate rule that determines the effectiveness of the effective bandwidth approximation.

Rule of Thumb for the Effective Bandwidth Approximation: A sufficient requirement for the effective bandwidth approximation to be bad is that the probability of loss at

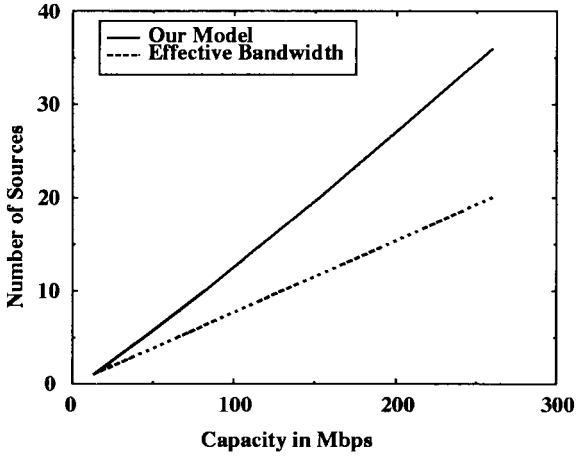


Fig. 11. Number of video calls that can be admitted. $\text{Loss} < 1.0e-6, N = 1000$. Limitation of the effective bandwidth scheme—a demonstration with JPEG-type video sources.

the point of transition between the cell domain and the burst domain must be significantly smaller than one. It is important to note that this probability is approximately given by the right-hand side of (8). The reason is twofold: first, because the probability of loss due to cell variability at the point of transition has so diminished that beyond this buffer size the loss due to the burst dominates, and second, because the probability of loss given by (8) corresponds to the loss when there is no cell level variation in the arrival process. Hence, using (8), depending on how close the value of $[(1/E(\lambda))\sum_{\substack{i=1\\ \lambda_i>\mu}}^{B} \lambda_i P_i(1-1/\rho_i)]$ is to one, will roughly determine the accuracy of the effective bandwidth approximation.

Example 7: In Fig. 11 we compare the number of video sources that can be accommodated into the network using our technique and the effective bandwidth scheme. The constraint on each call is a maximum probability of loss of 10^{-6}, and the size of the multiplexer buffer is 1000 cells. We compare the effective bandwidth results with the loss calculated using our model (a 20-state MMF source is used to describe each JPEG source for the effective bandwidth model). Fig. 11 shows the number of sources that can be accommodated for a particular link capacity. When the link capacity can only accommodate one source, only peak rate allocation is possible for JPEG video [27], [31]. Hence, there is no difference between the two curves at low capacities. However, as the capacity of the link is increased, due to the effect of statistical multiplexing, the difference between the number of calls allowed with our model and the effective bandwidth also increases. For example, when the effective bandwidth scheme can accommodate 20 calls, we find that with our loss estimate we can accommodate 36 calls, almost doubling the throughput!

VI. Conclusion and Future Work

We have provided an effective analytical technique to solve for the probability of loss for a Markov-modulated arrival process. In particular we analyze the MMPP and the fluid flow models in detail, and find that our approximation is both accurate as well as computationally efficient. An important contribution of our work is that its applicability encompasses heterogeneous sources as well.

We have also shown that the effective bandwidth technique that has become very popular because of its simplicity can in fact lead to considerable underutilization of the network. The accuracy of the effective bandwidth approximation depends on the type of sources being multiplexed, the number of sources being multiplexed, and the utilization at which the multiplexing takes place. We also provide a rule of thumb for determining when this approximation is good and when it is not. In most realistic cases we believe this approximation will be quite poor and we have shown many cases to support this belief. The Poisson approximation to the superposition of Markov-modulated sources is mainly invalid (except for some cases for very small buffers).

The hybrid model provides very encouraging results but needs more investigation and theoretical validation. For some types of sources, the effect of the other eigenvalues may not disappear until the buffer size is fairly large, as was shown in Fig. 8. For such sources, the hybrid model provides a conservative estimate, although still more accurate than the effective bandwidth approximation. Also, for future work, we intend to find rigorous conditions under which the hybrid model always predicts an upper bound. The difficult step here is to derive necessary and sufficient conditions for which the (log of the) exact loss probability curve, shown in Fig. 3, is convex.

Although we have only analyzed the MMPP and MMF models here, it should be clear that the approach can be generalized to other Markov-modulated sources. Preliminary results with general doubly stochastic sources that allow for *nonexponential* sojourn time distributions in each state using the hybrid model have also been quite promising.

References

[1] N. Akar and E. Arikan, "Markov modulated periodic arrival process offered to an multiplexer," in *Proc. IEEE GLOBECOM'93*, Houston, TX, 1993, pp. 783–787.

[2] D. Anick, D. Mitra, and M. Sondhi, "Stochastic theory of a data handling system with multiple sources," in *Conf. Rec. 1980 Int. Conf. Communications*, Seattle, WA, 1980, pp. 13.1.1–13.1.5.

[3] ———, "Stochastic theory of a data handling system with multiple sources," *Bell Syst. Tech. J.*, vol. 61, no. 8, pp. 1871–1894, 1982.

[4] A. Baiocchi, N. Melazzi, M. Listani, A. Roveri, and R. Winkler, "Loss performance analysis of an ATM multiplexer loaded with high-speed ON–OFF sources," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 388–393, Apr. 1991.

[5] R. Bellman, *Dynamic Programming.* Princeton, NJ: Princeton University Press, 1957.

[6] G. Choudhary, D. Lucantoni, and W. Whitt, "Squeezing the most out of ATM," *IEEE Trans. Commun.*, vol. 44, pp. 203–217, Feb. 1996; initially appeared as an AT&T internal technical report, 1993.

[7] J. Daigle and J. Langford, "Models for analysis of packet voice communications systems," *IEEE J. Select. Areas Commun.*, vol. SAC-4, pp. 847–855, Sept. 1986.

[8] J. Daigle and D. Lucantoni, "Queueing systems having phase-dependent arrival and service rates," in *Numerical Solution of Markov Chains*. New York: Dekker, 1990, pp. 161–202.

[9] A. Elwalid, D. Heyman, T. Lakshman, D. Mitra, and A. Weiss, "A new approach for allocating buffers and bandwidth to heterogeneous regulated traffic in an ATM node," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1017–1027, Aug. 1995.

[10] A. Elwalid and D. Mitra, "Approximations and admission control of a multi-service multiplexing system with priorities," in *Proc. IEEE INFOCOM'96*, Boston, MA, Apr. 1995, pp. 463–472.
[11] A. I. Elwalid, "Markov modulated rate processes for modeling, analysis and control of communication networks," Ph.D. dissertation, Grad. Sch. Arts Sci., Columbia Univ., New York, 1991.
[12] A. I. Elwalid and D. Mitra, "Effective bandwidth of general Markovian traffic sources and admission control of high speed networks," *IEEE/ACM Trans. Networking*, vol. 1, pp. 329–343, June 1993.
[13] M. Garrett, "Contributions toward real-time services on packet switched networks," Ph.D. dissertation, Grad. Sch. Arts Sci., Columbia Univ., New York, 1993.
[14] P. W. Glynn and W. Whitt, "Logarithmic asymptotics for steady-state tail probabilities in a single-server queue," *Adv. Appl. Prob.*, vol. 26, pp. 131–156, 1994.
[15] R. Guerin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high speed networks," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 968–991, Sept. 1991.
[16] H. Heffes and D. Lucantoni, "A Markov-modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," *IEEE J. Select. Areas Commun.*, vol. SAC-4, pp. 856–868, Sept. 1986.
[17] S.-S. Huang, "Source modeling for packet video," in *Proc. IEEE INFOCOM'88*, New Orleans, LA, 1988, pp. 1262–1267.
[18] F. Kelly, "Effective bandwidth in multiclass queues," *Queueing Syst.*, vol. 9, pp. 5–16, 1991.
[19] G. Kesidis, J. Walrand, and C.-S. Chang, "Effective bandwidths for multiclass Markov fluids and other ATM sources," *IEEE/ACM Trans. Networking*, vol. 1, pp. 424–428, Aug. 1993.
[20] J. Kingman, "A convexity property of positive matrices," *Quart. J. Math. Oxford*, vol. 12, no. 2, pp. 283–284, 1961.
[21] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson, and J. D. Robbins, "Performance models of statistical multiplexing in packet video communications," *IEEE Trans. Commun.*, vol. 36, pp. 834–844, July 1988.
[22] P. Pancha *et al.*, "Source characterization and equivalent bandwidth computation for MPEG video communication," manuscript in preparation/private communication, April 1997.
[23] J. Sairamesh and N. Shroff, "Exact formulas and properties of loss probability in some practical finite buffer queueing systems," in preparation.
[24] M. Schwartz, *Information Transmission, Modulation, and Noise*, 4th ed. New York: McGraw-Hill, 1990.
[25] P. Sen, B. Maglaris, N. Rikli, and D. Anastassiou, "Models for packet switching of variable-bit-rate video sources," *IEEE J. Select. Areas Commun.*, vol. 7, pp. 865–869, June 1989.
[26] N. Shroff, "Traffic modeling and analysis in high speed ATM networks," Ph.D. dissertation, Grad. Sch. Arts Sci., Columbia Univ., New York, 1995.
[27] N. Shroff and M. Schwartz, "Modeling VBR video over networks end-to-end using deterministic smoothing at the source," *Int. J. Commun. Syst.*, vol. 7, pp. 337–348, Dec. 1994.
[28] ——, "Video modeling within networks using deterministic smoothing at the source," in *Proc. IEEE INFOCOM'94*, Toronto, Ont., Canada, June 1994, pp. 342–349.
[29] ——, "Improved loss calculations at an ATM multiplexer," Sch. Elec. Comput. Eng., Purdue Univ., West Lafayette, IN, Tech. Rep., 1998.
[30] P. Skelly, "Characterization and control of variable bit rate video in broadband networks," Ph.D. dissertation, Grad. Sch. Arts Sci., Columbia Univ., New York, 1994.
[31] P. Skelly, M. Schwartz, and S. Dixit, "A histogram-based model for video traffic behavior in an ATM multiplexer," *IEEE/ACM Trans. Networking*, vol. 1, pp. 446–459, Aug. 1993.
[32] K. Sohraby, "On the asymptotic analysis of statistical multiplexers with hyper-bursty arrivals," IBM Watson Res. Center, Yorktown Heights, NY, Tech. Rep., 1993.
[33] ——, "On the theory of general ON/OFF sources with applications in high speed networks," in *Proc. IEEE INFOCOM'93*, San Francisco, CA, Apr. 1993, pp. 401–410.
[34] K. Sriram and W. Whitt, "Characterizing superposition of arrival processes in packet multiplexers for voice and data," *IEEE J. Select. Areas Commun.*, vol. SAC-4, pp. 833–846, Sept. 1986.
[35] T. E. Stern, "A queueing analysis of packet voice," in *Proc. GLOBECOM'83*, San Diego, CA, Nov. 1983, pp. 71–76.
[36] Z. Zhang and A. Acampora, "Effect of ON/OFF distributions of the cell loss probability in ATM networks," in *Proc. IEEE GLOBECOM'92*, Orlando, FL, Dec. 1992, pp. 1533–1539.

Ness B. Shroff (S'91–M'93) received the B.S. degree from the University of Southern California, Los Angeles, the M.S.E. degree from the University of Pennsylvania, Philadelphia, and the M.Phil. and Ph.D. degrees from Columbia University, New York, all in electrical engineering.

He is currently an Assistant Professor with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN. He has also had industrial experience with AT&T Bell Laboratories (1991) and Bell Communications Research (1992) during his doctoral study. His current research interests are in high-speed broad-band and wireless communication networks. He is especially interested in studying issues related to performance modeling, routing, network management, scheduling, and control in such networks. He has received research and equipment grants to conduct fundamental work in broad-band and wireless networks from the National Science Foundation, AT&T, Hewlett Packard, Intel, and the Purdue Research Foundation.

Dr. Shroff received the National Science Foundation CAREER award in 1996.

Mischa Schwartz (S'46–A'49–M'54–SM'54–F'66–LF'92) received the B.E.E. degree from Cooper Union, New York, in 1947, the M.E.E. degree from the Polytechnic Institute of New York, Brooklyn, in 1949, and the Ph.D. degree in applied physics from Harvard University, Cambridge, MA, in 1951.

From 1947 to 1952 he was a Project Engineer with the Sperry Gyroscope Company, working in the fields of statistical communication theory, radar detection, and radar system design. From 1952 to 1974 he was Professor of Electrical Engineering at the Polytechnic Institute of New York, Brooklyn, where he served Head of the Electrical Engineering Department from 1961 to 1965. During the year 1965–1966 he was a National Science Foundation Science Faculty Fellow at the Laboratoire de Physique, Ecole Normale Superieure, Paris, France. During the academic year 1973–1974 he was a Visiting Professor at Columbia University, New York, where he became Professor of Electrical Engineering and Computer Science in September 1974. He is currently with Columbia University, New York, as Charles Batchelor Professor Emeritus of Electrical Engineering, and where he is also associated with the Center for Telecommunications Research. For the 1980 calendar year he was on leave as a Visiting Scientist with IBM Research. During 1986 he was on leave as a Resident Consultant with NYNEX Science and Technology. During 1994 he spent half-time at IBM Research, working in the field of wireless communication systems. During the first half of 1995 he was first a Visiting Professor at University College London, under an EPSRC Visiting Fellowship, and then a half-time consultant on wireless networks at AT&T Bell Laboratories. From January to March 1997 he served as Visiting Professor at the University of California, San Diego, with the Center for Wireless Communications. He served from 1985 to 1988 as Director of the Columbia University Center for Telecommunications Research, one of six national engineering research centers established in 1985 under major grants of the National Science Foundation. He is author or coauthor of nine books and more than 150 technical publications on communication theory and systems, signal processing, and computer communication networks. He is on the editorial boards of *Networks*, *Telecommunication Systems*, the Japanese journal *IEICE Transactions on Communications*, the *Journal on Wireless Networks*, *Mobile Computing and Communications Review*, and *Mobile Networks and Applications*.

Dr. Schwartz is a member of the National Academy of Engineering, a Fellow of the American Association for the Advancement of Science (AAAS), Fellow of the International Engineering Consortium, and past Chairman of Commission C of the US National Committee/URSI. He is a former Director of the IEEE, former Chairman of the IEEE Information Theory Group, and past President of the Communications Society. He received a Distinguished Visitor Award from the Australian–American Education Foundation in 1975 and the IEEE Education Medal in 1983. He received the Great Teacher Medal in 1983 from Columbia University, New York. In 1984, the IEEE centennial year, he was cited as one of the ten all-time outstanding electrical engineering educators. In 1986 he was the recipient of the Cooper Union Gano Dunn Award, given annually for outstanding achievement in science and technology. In 1989 he received the IEEE Region I Award for outstanding engineering management and leadership. In 1994 he received the IEEE Communications Society Edwin H. Armstrong Achievement Award for outstanding contributions to communications technology. In 1995 he received the Mayor's Award for Excellence in Technology in New York City.