Call Admission for Prerecorded Sources with Packet Loss

Martin Reisslein and Keith W. Ross, *Senior Member, IEEE*

***Abstract*— We develop call admission policies for statistically multiplexing prerecorded sources over a bufferless transmission link. Our model is appropriate for video on demand, as well as other on-demand multimedia applications. In particular, we allow users to specify when the sources begin transmission; we also allow the user to invoke VCR actions such as pause and temporal jumps. We suppose that the quality of service (QoS) requirement allows for a small amount of packet loss.**

We develop a stochastic model which captures the random phases of the sources. We then apply large deviation theory to our model to develop global admission rules. The accuracy of the large deviation approximation is verified with simulation experiments employing importance sampling techniques. We also propose a refined admission rule which combines the global test and a myopic test. Numerical results are presented for the *Star Wars* trace; we find that the statistical multiplexing gain is potentially high and often insensitive to the QoS parameter. Finally, we develop efficient schemes for the real-time implementation of our global test. In particular, we demonstrate that the Taylor series expansion of the logarithmic moment generating function of the frame size distribution allows for fast and accurate admission decisions.

***Index Terms*— Call admission, packetized video, prerecorded sources, video on demand.**

I. Introduction

In recent years, there has been an explosion of research in packetized variable bit-rate (VBR) video, with the great majority of the work addressing *live video* such as videoconference and the broadcast of a sporting event. While this research on live video certainly merits the attention it has received, much (if not the majority) of the video carried on high-speed packet-switched networks will emanate from *prerecorded sources*. These sources include full-length movies, music video clips, and educational material. From the perspective of the transport network, prerecorded VBR video sources are fundamentally different from live video sources: for live video, the exact dynamics of the VBR traffic are unknown; for prerecorded sources, the amount of traffic in each frame is known before the frames are transmitted into the network.

In this paper, we develop call-admission policies for statistically multiplexing prerecorded sources over a single transmission link. Our model is appropriate for video on demand (VoD) as well as other multimedia and on-demand applications. However, to fix ideas, we will assume that all sources are video sources. We suppose that the transmission link is bufferless, so that packet loss occurs whenever the sum of the traffic rates, over the videos in progress, is greater than the link rate. We permit the users to begin the videos at random independent times; we also permit the users to pause and force temporal jumps (rewind and fast forward). We refer to pause and temporal jumps as VCR features.

We shall consider two quality of service (QoS) requirements. To define these QoS requirements, let ϵ be a fixed positive number, where a typical value of ϵ is 10^{-6}. The first QoS requirement is that the expected fraction of time during which cell loss occurs must be less than ϵ. The second QoS requirement is that the expected fraction of bits lost must be less than ϵ.

An *ideal admission policy* will accept a new video connection if and only if the QoS requirement will continue to hold with the additional connection. For live video, one is typically forced to employ a conservative admission policy since one never knows with certainty the type of traffic the live videos will generate. In this paper, we show how to construct an ideal admission policy for prerecorded video which can be efficiently implemented in real time.

Recently, several research groups have proposed schemes for the *lossless* multiplexing of prerecorded traffic. McManus and Ross [15]–[17], Kesidis and Hung [9], and Salehi *et al*. [21] have proposed schemes which use receiver memory and preplay delay. The principal feature of these schemes is that they produce high link utilizations, thereby reducing the network transport cost. But these schemes also require a more expensive receiver (due to the additional receiver memory) and are not easily amenable to temporal jumps. Knightly *et al*. [11], Knightly and Zhang [12], [13], and Liebeherr [3], [14] also propose schemes for lossless multiplexing for prerecorded sources; their approach is to place a buffer before the transmission link and admit new connections as long as the buffer is guaranteed to never overflow. It is shown in McManus and Ross [16] that unless the link buffer is large (implying a large playback delay), this last set of schemes give low link utilizations when the traffic is highly variable (such as in action movies).

Elwalid *et al*. [2] study admission control policies for statistically multiplexing VBR sources over a single buffered link; they assume the traffic is leaky-bucket controlled and organized in classes. The class structure is not appropriate for VoD systems supporting a large number of videos; moreover, leaky buckets provide a loose bound on video traffic, and result in pessimistic admission decisions.

Manuscript received April 30, 1996; revised September 30, 1996. This work was supported in part by NSF Grant NCR93-04601.

The authors are with the Department of Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104 USA.

Publisher Item Identifier S 0733-8716(97)04187-5.

When a video experiences a VCR action, its phase alignment changes with respect to the other videos being transmitted over the link. In Section II, we develop a novel stochastic model for prerecorded traffic with random phases, a model which captures the random occurrences of VCR actions in the videos. In Section III, we apply the central limit theorem and large deviation theory to our stochastic model to develop connection admission rules. These rules admit new connections only if it is highly likely that the QoS requirements will be satisfied for the duration of the video, even if the videos experience VCR actions. We refer to these admission rules as *global rules* since they account for the long-term expected behavior of the video traffic. In Section IV, we present the result of our numerical experiments with an MPEG encoding of *Star Wars*. We first develop an importance sampling heuristic for estimating cell loss with Monte Carlo simulation. We then show that one of our approximations is extremely accurate, and therefore leads to an ideal admission policy. We also find the statistical multiplexing gain to be high, especially when each video is smoothed over each of its group of pictures (GOP's). A brute force implementation of our ideal admission policy can be computationally prohibitive. In Section V, we describe two modifications which significantly reduce the amount of on-line computation that is needed for admission control. In Section VI, we introduce a refinement of our admission control procedure which takes both a global and myopic view of the traffic offered to the link. This policy admits a new connection only if: 1) there will be negligible cell loss in (say) the minute following call admission, assuming that no VCR actions occur; and 2) the QoS requirements are likely to be met over a long period of time (say, an hour). We find that an additional myopic test does not significantly reduce link utilization. We summarize our findings in Section VII.

II. Modeling Prerecorded Traffic with Random Phase Shifts

As mentioned in Section I, we assume that each video can experience VCR actions. We model these interactive features by associating with each video in progress an independent and random phase shift. We begin with some notation.

Consider J video streams multiplexed over a bufferless link with transmission rate C bits/s; see Fig. 1. For simplicity, assume that each video has N frames and has a frame rate of F frames/s. Let $x_n(j)$ be the number of bits in the nth encoded frame $(1 \le n \le N)$ of the jth video. Because we suppose that all videos are prerecorded, the sequence $\{x_n(j), 1 \le n \le N\}$ for the jth video is a known sequence of integers. We shall find it convenient to extend the definition of this sequence for n ranging from $-\infty$ to $+\infty$ as follows:

$$(\cdots, x_{-2}(j), x_{-1}(j), x_0(j), x_1(j), x_2(j), \cdots)$$

where

$$x_{n+N}(j) = x_n(j), \qquad \text{for} \quad n = \cdots, -2, -1, 0, 1, 2, \cdots.$$

Thus, the infinite sequence is created by repeating the video trace over and over again.

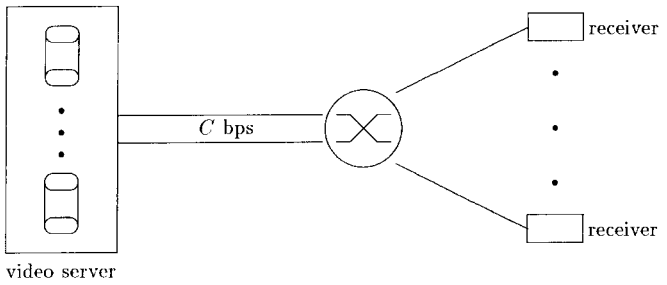

Fig. 1. Prerecorded videos multiplexed over a link of capacity C bits/s. Throughout, we assume that the switch and the connections to the receivers introduce no delay.

To model the random start times and the VCR actions, we assign to the jth video, $j = 1, \cdots, J$, the random phase θ_j. We suppose that the θ_j's, $j = 1, \cdots, J$, are independent, and that each θ_j is uniformly distributed over $\{0, \cdots, N-1\}$. (We choose a uniform distribution to fix ideas; however, the theory can be developed with an arbitrary distribution.) Our model supposes that the amount of traffic generated by the jth video at frame time n is

$$X_n(j, \theta_j) := x_{n+\theta_j}(j).$$

Therefore, for a given phase profile $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_J)$, the total amount of traffic generated by the J videos at frame time n is

$$X_n(\boldsymbol{\theta}) = \sum_{j=1}^{J} X_n(j, \theta_j) = \sum_{j=1}^{J} x_{n+\theta_j}(j).$$

This completes our formal definition of the traffic model. Henceforth, we write $X_n(j)$ and X_n for $X_n(j, \theta_j)$ and $X_n(\boldsymbol{\theta})$, respectively.

Having defined the traffic model, we now highlight some of its implications. First, note that for each fixed $n, X_n(1), \cdots, X_n(J)$ are independent random variables. Second, note that the probability mass function for $X_n(j)$ can be calculated directly from the known trace $\{x_1(j), x_2(j), \cdots, x_N(j)\}$ as follows:

$$P(X_n(j) = l) = \pi_j(l) \tag{1}$$

where we define

$$\pi_j(l) := \frac{1}{N} \sum_{n=1}^{N} 1(x_n(j) = l).$$

Observe, in particular, that the distribution of $X_n(j)$ does not depend on n. We tacitly assume here that VCR actions do not change the frame size distribution, that is, we assume that viewers do not invoke VCR actions to see more high-action scenes.

A. Quality of Service Measures

Over the period of time during which a video is in progress, the video will see a variety of different phase profiles $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_J)$ with respect to the other videos in progress. For each phase profile, there will be a fraction of frame periods during which the traffic rate exceeds the link rate and cell loss occurs. By taking the expectation over all possible phase

profiles, we obtain the expected fraction of frame periods during which cell loss occurs. We are therefore motivated to define $P_{\text{loss}}^{\text{time}}$ by

$$P_{\text{loss}}^{\text{time}} := E\left[\lim_{M\to\infty}\frac{1}{M}\sum_{m=1}^{M}1\left(X_m > \frac{C}{F}\right)\right].$$

Note that for all $M \geq 1$

$$P_{\text{loss}}^{\text{time}} := E\left[\frac{1}{M}\sum_{m=1}^{M}1\left(X_m > \frac{C}{F}\right)\right]$$

and that for all n

$$P_{\text{loss}}^{\text{time}} = P\left(X_n > \frac{C}{F}\right). \tag{2}$$

In the subsequent sections, we will find it convenient to work with the expression (2) for $P_{\text{loss}}^{\text{time}}$.

Because X_n is the sum of J independent random variables, the probability in (2) can be calculated by convolution. However, convolution often leads to numerical problems, and its computational complexity may not be suitable for real-time admission control. For these reasons, we shall explore approximations and bounds for $P(X_n > (C/F))$ in the next section.

We shall also study the QoS measure $P_{\text{loss}}^{\text{info}}$, where

$$P_{\text{loss}}^{\text{info}} := \frac{E\left[\left(X - \frac{C}{F}\right)^+\right]}{E[X]}. \tag{3}$$

From the fact

$$\sum_{n=1}^{N}X_n = \sum_{n=1}^{N}\sum_{j=1}^{J}x_n(j)$$

it follows that

$$P_{\text{loss}}^{\text{info}} = E\left[\frac{\sum_{n=1}^{N}\left(X_n - \frac{C}{F}\right)^+}{\sum_{n=1}^{N}X_n}\right]$$

$$= E\left[\lim_{M\to\infty}\frac{\sum_{m=1}^{M}\left(X_m - \frac{C}{F}\right)^+}{\sum_{m=1}^{M}X_m}\right].$$

These two last expressions evoke an average—over all possible phase profiles—of the fraction of information (bits) lost.

To simplify notation, we henceforth write X for X_n and write $X(j)$ for $X_n(j)$. Also, let $a := C/F$.

III. Bounding and Approximating QoS Measures

In this section, we develop a central limit and large deviation approximation for $P_{\text{loss}}^{\text{time}} = P(X > a)$ and $P_{\text{loss}}^{\text{info}} = E[(X - a)^+]/E[X]$.

A. Central Limit Approximation

For the jth video, the average number of bits in a frame is

$$m(j) = \frac{1}{N}\sum_{n=1}^{N}x_n(j)$$

and the sample variance is

$$\sigma^2(j) = \frac{1}{N-1}\sum_{n=1}^{N}[x_n(j) - m(j)]^2.$$

By the central limit theorem [1, p. 310], X is approximately a normal random variable with mean and variance

$$m = \sum_{j=1}^{J}m(j) \quad \sigma^2 = \sum_{j=1}^{J}\sigma^2(j).$$

Throughout the remainder of this subsection, we assume that the approximation is exact, that is, we assume that

$$X \triangleq N(m, \sigma^2).$$

With the established traffic model, the expected fraction of time that there is cell loss $P_{\text{loss}}^{\text{time}} = P(X > a)$ may now be easily computed from the tail of the normal distribution. Given a specific QoS requirement $P_{\text{loss}}^{\text{time}} \leq \epsilon$, where ϵ is a small number such as 10^{-7}, the QoS requirement $P_{\text{loss}}^{\text{time}} \leq \epsilon$ is met if and only if

$$\frac{1}{2}\text{erfc}\left(\frac{a-m}{\sqrt{2\sigma^2}}\right) \leq \epsilon \tag{4}$$

where $\text{erfc}(\cdot)$ denotes the well-known complementary error function defined as

$$\text{erfc}(a) = \frac{2}{\sqrt{\pi}}\int_a^\infty e^{-t^2}\,dt.$$

Using this admission rule (4), whenever a new video $J+1$ requests establishment, we update $m \leftarrow m + m(J+1)$ and $\sigma^2 \leftarrow \sigma^2 + \sigma^2(J+1)$. (It is natural to assume that $m(j)$ and $\sigma^2(j)$ have been calculated off line.) We then admit the new video if and only if (4) is met.

Now, suppose that the QoS requirement $P_{\text{loss}}^{\text{info}} \leq \epsilon$ is imposed. Given the normal distribution of the amount of traffic that arrives during a specific frame period, $P_{\text{loss}}^{\text{info}} = E[X - a)^+]/E[X]$ can be calculated by noting that

$$E[(X-a)^+] = \int_a^\infty (x-a)f_X(x)\,dx$$

where

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-[(x-m)^2/2\sigma^2]}$$

and furthermore, $E[X] = m$. Some straightforward calculus shows that the QoS requirement $P_{\text{loss}}^{\text{info}} \leq \epsilon$ is met if and only if

$$\frac{1}{2}\left(1 - \frac{a}{m}\right)\text{erfc}\left(\frac{a-m}{\sqrt{2\sigma^2}}\right) + \frac{1}{\sqrt{2\pi}}\,\frac{\sigma}{m}\exp\left[\frac{-(a-m)^2}{2\sigma^2}\right] \leq \epsilon.$$

B. Large Deviation Bound and Approximation

For small tail probabilities, the central limit approximation can underestimate the loss probability (see numerical results in Section IV). In this subsection, we provide an upper bound for $P_{\text{loss}}^{\text{time}}$ and large deviation approximations for $P_{\text{loss}}^{\text{time}}$ and $P_{\text{loss}}^{\text{info}}$. To this end, for any random variable Y, let

$$\mu_Y(s) := \ln\, E[e^{sY}].$$

Note that $\mu_Y(s)$ is the logarithm of the moment generating function for Y.

For any random variable Y and constant c, the Chernoff bound gives an upper bound for $P(Y > c)$ (see Hui [8], [7], Kelly [10], Mitra *et al*. [2], and Roberts [20, p. 109]). Applying the Chernoff bound to $P(X > a)$, we obtain

$$P_{\text{loss}}^{\text{time}} \le e^{-s^* a + \mu_X(s^*)} \tag{5}$$

where s^* is the unique solution to

$$\mu_X'(s^*) = a. \tag{6}$$

Note that

$$\mu_X(s) = \sum_{j=1}^{J} \mu_{X(j)}(s).$$

Given a specific QoS requirement $P_{\text{loss}}^{\text{time}} \le \epsilon$, the Chernoff bound gives the admission control condition

$$e^{-s^* a + \mu_X(s^*)} \le \epsilon. \tag{7}$$

Using (7), the admission control mechanism operates as follows. The logarithmic generating functions $\mu_{X(j)}(s)$ are first calculated off line for all videos. Now, suppose a new video $J+1$ requests establishment. We update $\mu_X(s) \leftarrow \mu_X(s) + \mu_{X(J+1)}(s)$ and then find the s^* that satisfies (6). Finally, we admit the video if and only if (7) is satisfied. We remark that Grossglauser *et al*. [5] have recently proposed a large deviation approximation for $P_{\text{loss}}^{\text{time}}$. In their scheme, the prerecorded traffic is first smoothed into piecewise constant rates. For each change in the constant rate, they propose a renegotiation of network bandwidth. Their "probability of renegotiation failure" is similar to $P_{\text{loss}}^{\text{time}}$.

The bound given by (5) can be converted to an accurate approximation by applying the theory of large deviations (see Hui [8, p. 202], Hsu and Walrand [6], and Roberts [20, p. 109])

$$P_{\text{loss}}^{\text{time}} = P(X \ge a) \approx \frac{1}{s^* \sqrt{2\pi \mu_X''(s^*)}} e^{-s^* a + \mu_X(s^*)}.$$

The large deviation (LD) approximation gives the following admission control condition:

$$\frac{1}{s^* \sqrt{2\pi \mu_X''(s^*)}} e^{-s^* a + \mu_X(s^*)} \le \epsilon. \tag{8}$$

There is also a large deviation approximation for $P_{\text{loss}}^{\text{info}}$ (see Roberts [20, p. 154]):

$$P_{\text{loss}}^{\text{info}} = \frac{E[(X-a)^+]}{E[X]} \approx \frac{1}{m s^{*2} \sqrt{2\pi \mu_X''(s^*)}} e^{-s^* a + \mu_X(s^*)}.$$

TABLE I
STATISTICS OF *Star Wars* TRACE

Frames			GOPs			Bitrate	
Mean bits	St. Dev. bits	Peak/ Mean	Mean bits	St. Dev. bits	Peak/ Mean	Mean Mbps	Peak Mbps
15,598	18,165	11.9	187,178	72,869	5.05	0.37	4.45

Thus, the large deviation approximation gives the following admission control criterion:

$$\frac{1}{m s^{*2} \sqrt{2\pi \mu_X''(s^*)}} e^{-s^* a + \mu_X(s^*)} \le \epsilon. \tag{9}$$

IV. NUMERICAL RESULTS FOR THE *Star Wars* TRACE

In order to evaluate the admission control conditions introduced in the previous section, we conducted some numerical experiments with the MPEG-I *Star Wars* bandwidth trace, available via anonymous FTP from Bellcore [4]. The trace gives the number of bits in each video frame. In our experiments, all of the J videos use the *Star Wars* trace, but each video has its own random phase. The movie was compressed with the group of pictures (GOP) pattern *IBBPBBPBBPBB* (12 frames) at a frame rate $F = 24$ frames/s. The trace has a total number of $N = 174\,136$ frames, which corresponds to a run time of approximately 2 h. Some salient statistical properties of the *Star Wars* trace are given in Table I. We consider a single bufferless ATM node with transmission rate $C = 155$ Mbit/s. (At the end of this section, we also consider $C = 45$ Mbits/s.) We furthermore assume that all 48 bytes of the ATM cell payload are used to transport the video frames.

In order to validate the approximation described in Section III, we ran simulation experiments using the *Star Wars* trace. In the simulation algorithms described below, we focus on the $P_{\text{loss}}^{\text{time}}$ criterion; the algorithms for $P_{\text{loss}}^{\text{info}}$ are similar. Based on our assumption of uniformly distributed phases (see Section II), there are two different simulation approaches possible.

One approach works as follows. For a fixed number of connections J, draw the phases $\theta_j, j = 1, \cdots, J$, from a discrete uniform distribution over $[0, N-1]$ (denote this distribution by DU[0, $N-1$]) and check whether loss occurs for this phase profile, that is, check whether $\Sigma_{j=1}^{J} x_{\theta_j}(j) > a$. Repeat this procedure many times in order to obtain an estimate for the fraction of phase profiles that have loss, that is, for $P_{\text{loss}}^{\text{time}}$.

An alternative approach proceeds as follows. Draw the start phases $\theta_j \sim \text{DU}[0, N-1], j = 1, \cdots, J$, and then simulate the transmission of the entire *Star Wars* trace and count the number of frames with loss (see Fig. 2 for the details of this algorithm). In this algorithm, we wrap the trace around when the index extends beyond the end of the trace, that is, for $n + \theta_j > N$, we replace $n + \theta_j$ by $n + \theta_j - N$.

We used the second approach in our simulation experiments since it gives tighter confidence intervals than the first approach for the same amount of CPU time. The first approach is computationally more expensive because it requires a new set of random phases for each simulated frame time, while the second approach allows us to use the same set of random phases for N frame times. Note that there is a statistical

```
1.  Fix J;
2.  o = 0; p = 0;
3.  For l = 1 to L do
4.    Draw θ_j ~ DU[0, N − 1], j = 1, ..., J;
5.    q = 0;
6.    For n = 1 to N do
7.      X_n = Σ_{j=1}^{J} x_{n+θ_j}(j);
8       q = q + 1(X_n > C/F);
9.    o = o + q;
10.   p = p + q^2;
11. P̂_loss^time = o/(NL); (sample mean)
12. Ŝ^2 = 1/(N^2(L−1)) (p − o^2/L); (sample variance)
```

Fig. 2. Simulation algorithm for $P_{\text{loss}}^{\text{time}}$.

TABLE II
SIMULATION RESULTS FOR J UNSMOOTHED *Star Wars* CONNECTIONS WITHOUT AND WITH IMPORTANCE SAMPLING

	no IS		IS	
J	$\hat{P}_{\text{loss}}^{\text{time}}$	90% CI	$\hat{P}_{\text{loss}}^{\text{time}}$	90% CI
268	$4.59\ 10^{-8}$	$[0, 4.54\ 10^{-7}]$	$1.80\ 10^{-7}$	$[1.37\ 10^{-7}, 2.23\ 10^{-7}]$
276	$8.43\ 10^{-7}$	$[0, 4.61\ 10^{-6}]$	$1.35\ 10^{-6}$	$[9.83\ 10^{-7}, 1.72\ 10^{-6}]$
284	$8.50\ 10^{-6}$	$[0, 5.12\ 10^{-5}]$	$5.88\ 10^{-6}$	$[3.76\ 10^{-6}, 7.99\ 10^{-6}]$
292	$8.28\ 10^{-5}$	$[0, 4.46\ 10^{-4}]$	$7.27\ 10^{-5}$	$[1.40\ 10^{-5}, 1.31\ 10^{-4}]$

difference between the two approaches. In the first approach, every simulated frame period constitutes an independent trial; in the second approach, on the other hand, the simulated transmission of the entire trace is an independent trial since the frame sizes are correlated within the video trace. This fact has to be accounted for when computing the sample variance (see Fig. 2, line 12).

In Table II (column no IS), we give the sample mean and 90% confidence intervals for $P_{\text{loss}}^{\text{time}}$ for the unsmoothed *Star Wars* trace. In this experiment, we ran $L = 500$ replications for a total of $500 \times 174\,136 \approx 87 \times 10^6$ simulated frame periods. We note that the half lengths of the confidence intervals are larger than the sample means; this implies that the lower end of the confidence interval for $P_{\text{loss}}^{\text{time}}$ is zero, as subtracting the half length from the sample mean would result in negative probabilities. This phenomenon is a consequence of the simulation of rare loss events; tightening the confidence intervals further without employing variance reduction techniques, such as importance sampling (discussed below), would require immense computational resources. Each of the four simulation results in the no IS column in Table II took about two days on a SPARCstation 2.

Table III shows the results obtained after smoothing the *Star Wars* trace over each group of pictures (GOP's). In this experiment we ran $L = 2000$ replications for a total of $2000 \times 14\,511 \approx 29 \times 10^6$ simulated GOP's. We observe that the simulation of 29×10^6 GOP's gives confidence intervals that are significantly tighter than those obtained for the unsmoothed trace, even though the simulation for the unsmoothed trace required three times the computational effort. This seems to indicate that the GOP smoothed trace is more amenable to simulation experiments.

Importance Sampling: In order to obtain better estimates for $P_{\text{loss}}^{\text{time}}$, particularly for the unsmoothed trace, and to reduce the simulation time, we apply importance sampling (IS) techniques to our problem. The basic idea of IS is to draw the random phases $\theta_j, j = 1, \cdots, J$, from a distribution $g_j(m_j) = P(\theta_j = m_j)$ that leads to more frequent losses than the discrete unform distribution used in the experiments described above. Let $f_j(m_j) = 1/N$ denote this discrete uniform distribution. Since we use the *Star Wars* trace for all J video streams, we use the same distribution for all videos, and write henceforth $g(m_j)$ and $f(m_j)$ for $g_j(m_j)$ and $f_j(m_j)$. The fraction of frame periods for which there is cell loss is given by

$$P(X > a) = E[1(X > a)] = \sum_{m_1=1}^{N} \cdots \sum_{m_J=1}^{N} h(m_1, \cdots, m_J) \prod_{j=1}^{J} g(m_j)$$

where

$$h(m_1, \cdots, m_J) = 1(x_{m_1} + \cdots + x_{m_J} > a) \left(\prod_{j=1}^{J} \frac{f(m_j)}{g(m_j)} \right).$$

The Monte Carlo estimate for $P_{\text{loss}}^{\text{time}}$ is given by

$$\hat{P}_{\text{loss}}^{\text{time}} = \frac{1}{L} \sum_{l=1}^{L} h(m_1^{(l)}, \cdots, m_J^{(l)})$$

where the samples $m_j^{(l)}, j = 1, \cdots, J$, are all drawn from $g(\cdot)$ for $l = 1, \cdots, L$. With IS for each l, we use a new set of random phases (the θ_j's). Recall that without IS, we use the same set of phases for any entire simulated trace.

The challenge of the IS approach lies in finding the distribution $g(\cdot)$ that gives small variances for $P_{\text{loss}}^{\text{time}}$. For the GOP-smoothed trace, we obtained the IS results in Table III for $L = 100\,000$ with

$$g(m) = \frac{y_m}{\sum_{n=1}^{N} x_n}.$$

Here, y_m is the size of the mth GOP, that is, $y_m = \sum_{n=(m-1)G+1}^{mG} x_n$ and G denotes the number of frames per GOP. Each simulation result took about 1 h, roughly 20 times faster than the corresponding simulation without IS. Note this choice of $g(\cdot)$ favors larger frames [as compared with $f(\cdot)$]. This will cause $h(m_1, \cdots, m_J)$ to be strictly positive more frequently; but when $h(m_1, \cdots, m_J)$ is strictly positive, it will not be excessively large since its denominator is likely large. Thus, the sampling function $g(\cdot)$ causes the output stream of $h(\cdot)$ to be less variable, thereby giving a tighter confidence interval.

TABLE III
SIMULATION RESULTS FOR J GOP-SMOOTHED *Star Wars* CONNECTIONS WITHOUT AND WITH IMPORTANCE SAMPLING

	no IS		IS	
J	$\hat{P}_{\text{loss}}^{\text{time}}$	90% CI	$\hat{P}_{\text{loss}}^{\text{time}}$	90% CI
338	$1.10\ 10^{-6}$	$[3.90\ 10^{-7}, 1.82\ 10^{-6}]$	$5.37\ 10^{-7}$	$[4.89\ 10^{-7}, 5.86\ 10^{-7}]$
342	$7.17\ 10^{-6}$	$[5.29\ 10^{-6}, 9.05\ 10^{-6}]$	$6.37\ 10^{-6}$	$[5.54\ 10^{-6}, 7.19\ 10^{-6}]$
346	$6.15\ 10^{-5}$	$[5.77\ 10^{-5}, 6.52\ 10^{-5}]$	$5.04\ 10^{-6}$	$[4.27\ 10^{-5}, 5.81\ 10^{-5}]$

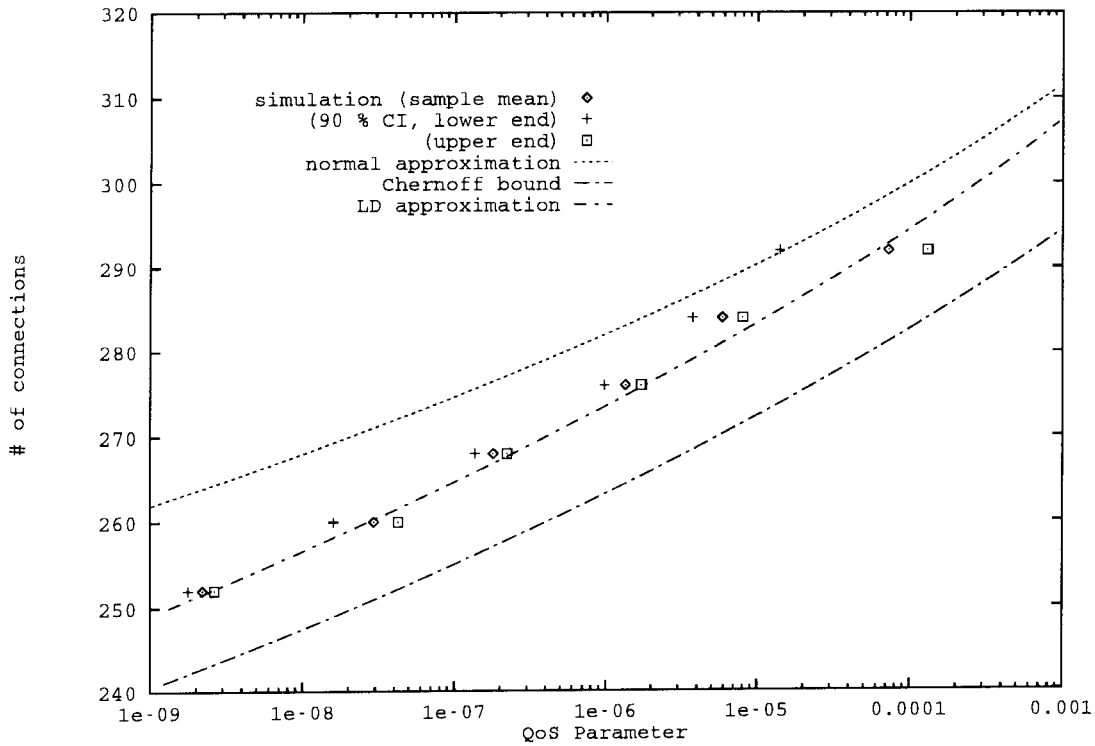

Fig. 3. Comparison of approximations and simulation (unsmoothed, $P_{\text{loss}}^{\text{time}}$ criterion).

We found that the discrete empirical distribution

$$g(m) = \frac{\sqrt{x_m}}{\sum_{n=1}^{N} \sqrt{x_n}}$$

works well for the unsmoothed *Star Wars* trace. The square root in $g(\cdot)$ dampens the tendency to draw only large frames. Given the large ratio of peak-to-mean frame size (see Fig. 1), the distribution $g(m) = x_m / \Sigma_{n=1}^{N} x_n$ would misrepresent the trace. A small number of excessively large frames would dominate the simulation, while the large number of small frames would be ignored. The chosen distribution, however, still encourages large values for the x_n's and leads to more frequent losses; the $g(\cdot)$ in the denominator of the expression for $h(\cdot)$ compensates for this effect. The IS results displayed in Table II were obtained for $L = 2 \times 10^6$; note that IS has significantly reduced the width of the confidence intervals. Each of the four simulation results took approximately 2 h on a SPARCstation 2.

Approximations and Bounds: In Fig. 3, we compare the results from IS, the normal approximation, the LD approximation, and the Chernoff bound. The figure shows the number of *Star Wars* connections allowed for by the condition $P_{\text{loss}}^{\text{time}} \leq \epsilon$ as the QoS parameter ϵ varies. The calculations are based on the unsmoothed *Star Wars* trace, that is, each frame is transmitted in its assigned interval of length $1/F$. The diamonds represent the sample mean $\hat{P}_{\text{loss}}^{\text{time}}$ resulting from IS. We ran $L = 2 \times 10^6$ replications for each $J = 252, 260, \cdots, 292$. For a fixed number of connections J, the crosses and boxes indicate the lower and upper ends of the 90% confidence interval for $P_{\text{loss}}^{\text{time}}$. The curve labeled "normal approximation" is calculated using the admission control condition (4). Given that the simulation represents the actual loss probabilities, the normal approximation is an optimistic admission policy. The curves labeled "Chernoff bound" and "LD approximation" are computed using conditions (7) and (8), respectively. The Chernoff bound appears to be a conservative admission control policy as it lies about ten connections below the boxes representing the upper end of the 90% confidence interval for $P_{\text{loss}}^{\text{time}}$. The LD approximation appears to be the appropriate tool for admission control for prerecorded VBR video as it almost coincides with the sample means from the simulation. We will henceforth focus on the LD approximation in our analysis.

In Fig. 4, we investigate the effect of smoothing the trace over a GOP. The curves were obtained by using the $P_{\text{loss}}^{\text{time}}$ criterion (8). The "unsmoothed curves" already appeared in Fig. 3, and are replicated here for comparison purposes. The GOP curves were calculated by first smoothing the *Star Wars* trace over each GOP (12 frames). The figure reveals that smoothing over the GOP increases the admission region substantially, resulting in higher network utilization for a given QoS requirement. For a QoS parameter $\epsilon = 10^{-6}$, for example, smoothing over the GOP increases the number of admissible connections by approximately 24%. This corresponds to an increase in the average link utilization (defined as the number of admissible connections $\times m/a$) from 73% for the unsmoothed trace to 90% for the GOP smoothed trace. We also observe from Fig. 4 that the number of allowed connections is nearly constant with respect to the QoS requirement for GOP smoothing. Peak rate admission allows for 31 unsmoothed *Star Wars* connections, giving an average link utilization of 8.3%. After smoothing the *Star Wars* trace over the GOP, a peak

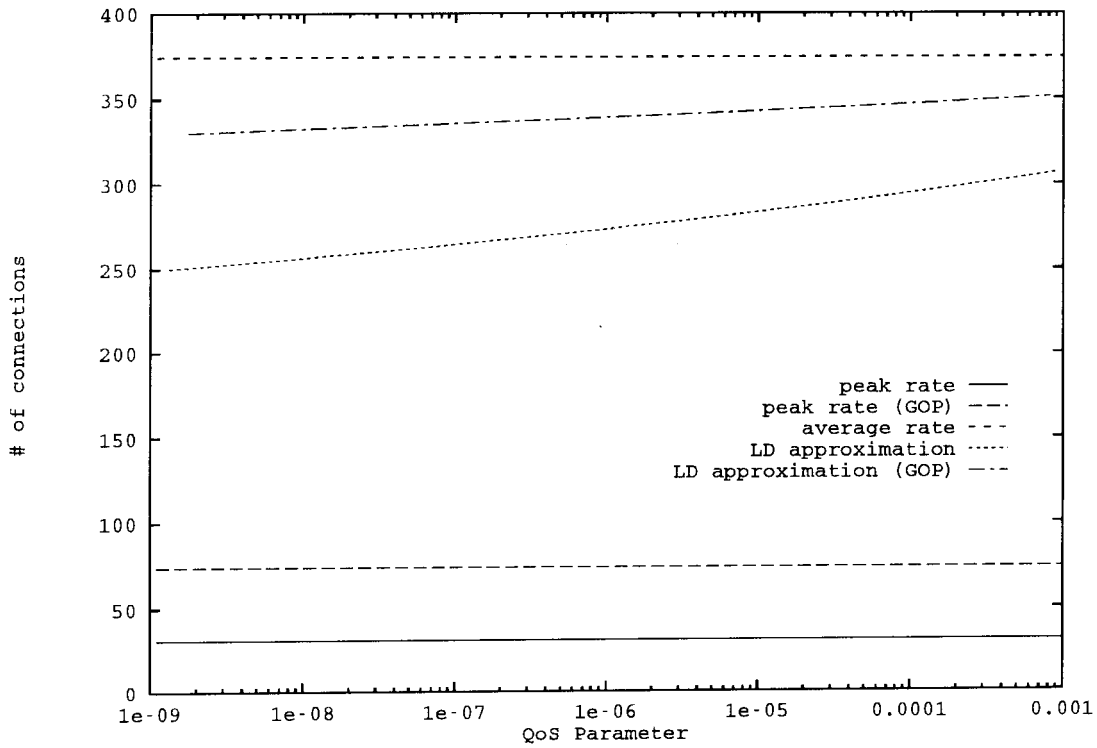

Fig. 4. Effect of GOP smoothing on the number of admissible connections as a function of $P_{\text{loss}}^{\text{time}}$.

rate admission policy would allow for 74 connections, which results in an average link utilization of 20%. Average-rate admission permits 374 connections $(= C/Fm)$.

In Fig. 5, we compare the criteria $P_{\text{loss}}^{\text{time}}$ and $P_{\text{loss}}^{\text{info}}$. We notice that the $P_{\text{loss}}^{\text{info}}$ criterion allows for slightly more connections. For $\epsilon = 10^{-7}$, for example, the number of admissible unsmoothed *Star Wars* connections is increased by approximately 6%, while the number of GOP-smoothed connections is 2.4% higher. This observation can be intuitively explained by noting that only the fraction $(X-a)/a$ of cells is lost during periods of overload. While the criterion $P_{\text{loss}}^{\text{time}}$ is based on the long run fraction of time there is cell loss, the criterion $P_{\text{loss}}^{\text{info}}$ accounts for the fact that not all cells are lost during overload.

A. Numerical Experiments for 45 Mbits/s Link

In order to investigate the effect of network bandwidth on our results, in particular on the accuracy of the LD approximation, we chose to replicate the experiments described in the previous section for an ATM link with bandwidth $C = 45$ Mbit/s. As in the previous experiments, we use the MPEG 1 encoded *Star Wars* trace, and assume that all 48 bytes of the ATM cell payload are used to transport the frames.

In Fig. 6, we depict the results from IS simulation, normal approximation, Chernoff bound, and LD approximation for the unsmoothed *Star Wars* trace for $C = 45$ Mbits/s. As in the experiment for $C = 155$ Mbits/s (see Fig. 3), we set the parameter L of the IS simulation to 2×10^6 Notice that the IS heuristic introduced in the previous section gives even tight confidence intervals for $C = 45$ Mbits/s. The simulation also verifies that the LD approximation remains accurate when fewer connections are multiplexed.

Fig. 7 shows the effect of smoothing the trace over each GOP. We also plot the number of admissible connections when the *Star Wars* trace is smoothed over three GOP's. Note that the additional smoothing hardly increases the admission region. We also see, as in the case of $C = 155$ Mbits/s (see Fig. 4), that smoothing over one GOP increases the number of admissible connections substantially. For the QoS parameter $\epsilon = 10^{-6}$, for instance, the number of admissible connections is increased by about 54%.

Note that the link utilizations are not as high as for the 155 Mbits/s link. The QoS requirement $P_{\text{loss}}^{\text{time}} \leq 10^{-6}$ permits 57 unsmoothed *Star Wars* connections, which corresponds to a link utilization of approximately 53% (we had 73% for $C = 155$ Mbits/s). For the same QoS requirement, 88 GOP smoothed connections are allowed, resulting in a link utilization of about 81% (90% for $C = 155$ Mbits/s). As in the case of $C = 155$ Mbits/s, we see that the number of admissible GOP-smoothed connections is almost insensitive to the QoS parameter.

V. Efficient Calculation of Admission Control Decisions

In this section, we address the real-time implementation of the proposed admission control tests. The criterion derived for the normal approximation (4) can be readily tested in real time (assuming that $m(j)$'s and $\sigma(j)$'s have been calculated off line for all videos).

The conditions resulting from the Chernoff bound and LD approximation, however, require more computational effort. Recall from Section II-C that these conditions are based on $\mu_X(s)$, the logarithmic generating function of the total amount

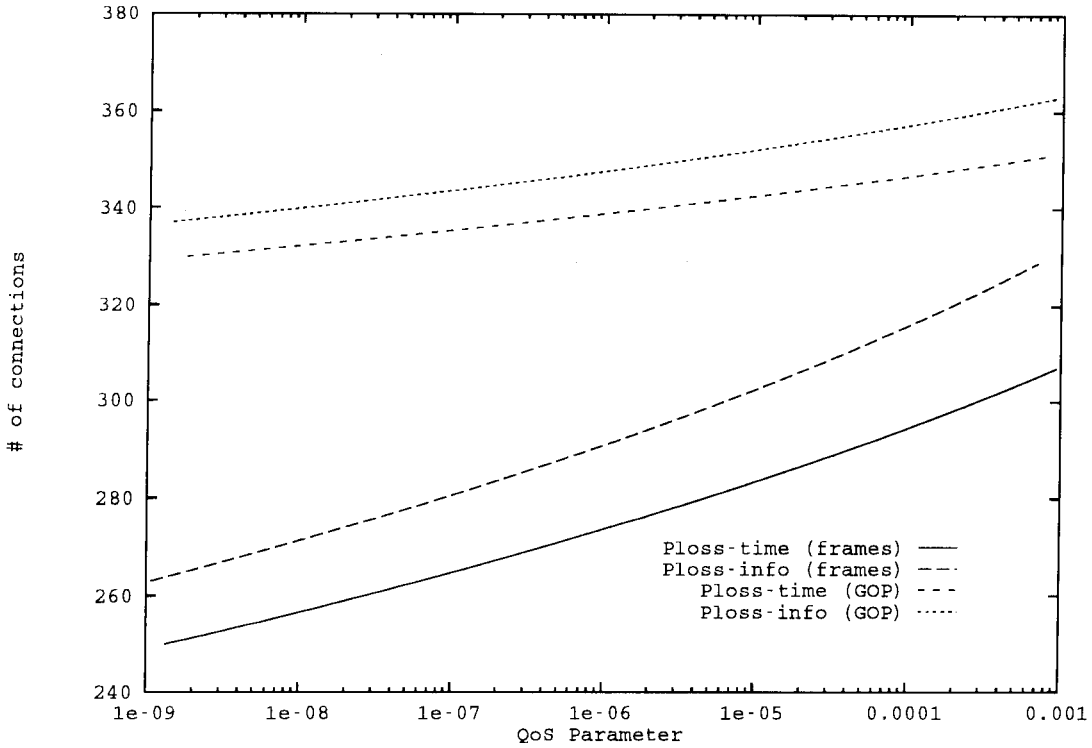

Fig. 5. Comparison of $P_{\text{loss}}^{\text{time}}$ and $P_{\text{loss}}^{\text{info}}$ criteria.

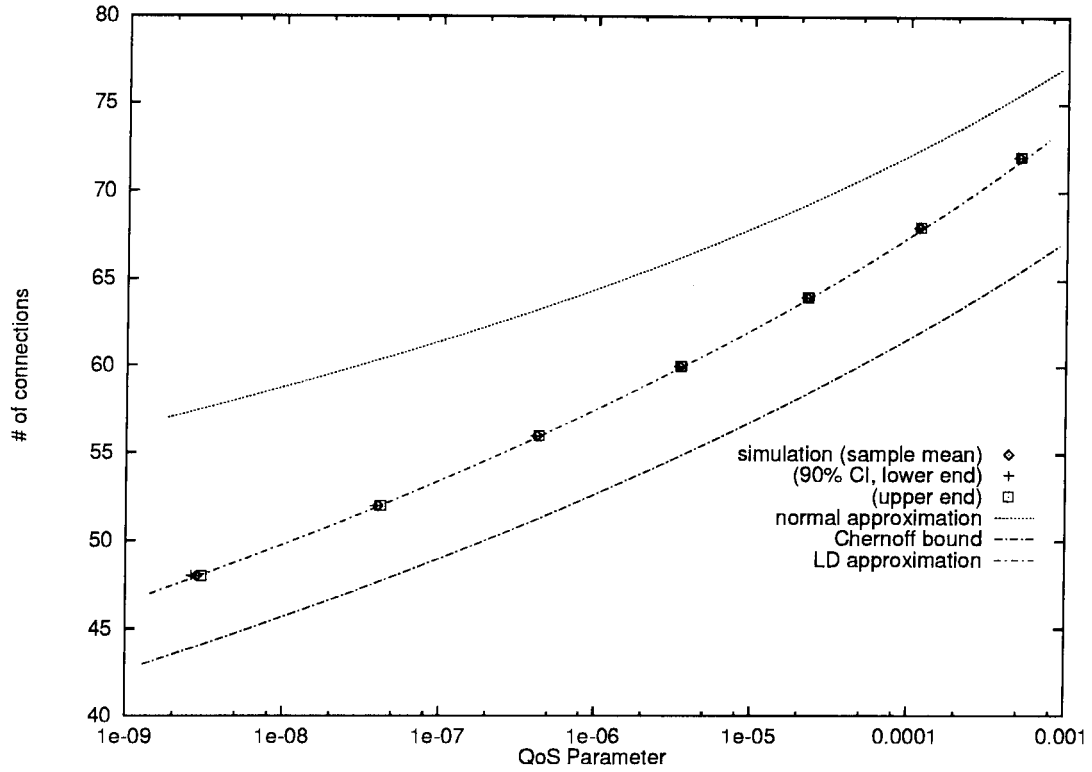

Fig. 6. Comparison of approximations and simulation for $C = 45$ Mbits/s as a function of $P_{\text{loss}}^{\text{time}}$.

of traffic arriving from all active video streams in one frame period. With (1), this function can be explicitly written as

$$\mu_X(s) = \sum_{j=1}^{J} \ln\left(\sum_{l=x_{\min}}^{x_{\max}} \pi_j(l) e^{sl}\right) \tag{10}$$

where $x_{\min}$ and $x_{\max}$ denote the smallest and largest frame size, respectively. We assume for simplicity that $x_{\min}$ and $x_{\max}$ are identical for all videos. We furthermore assume that the histogram of the frame sizes $\pi_j(l)$ has been computed off line for all videos.

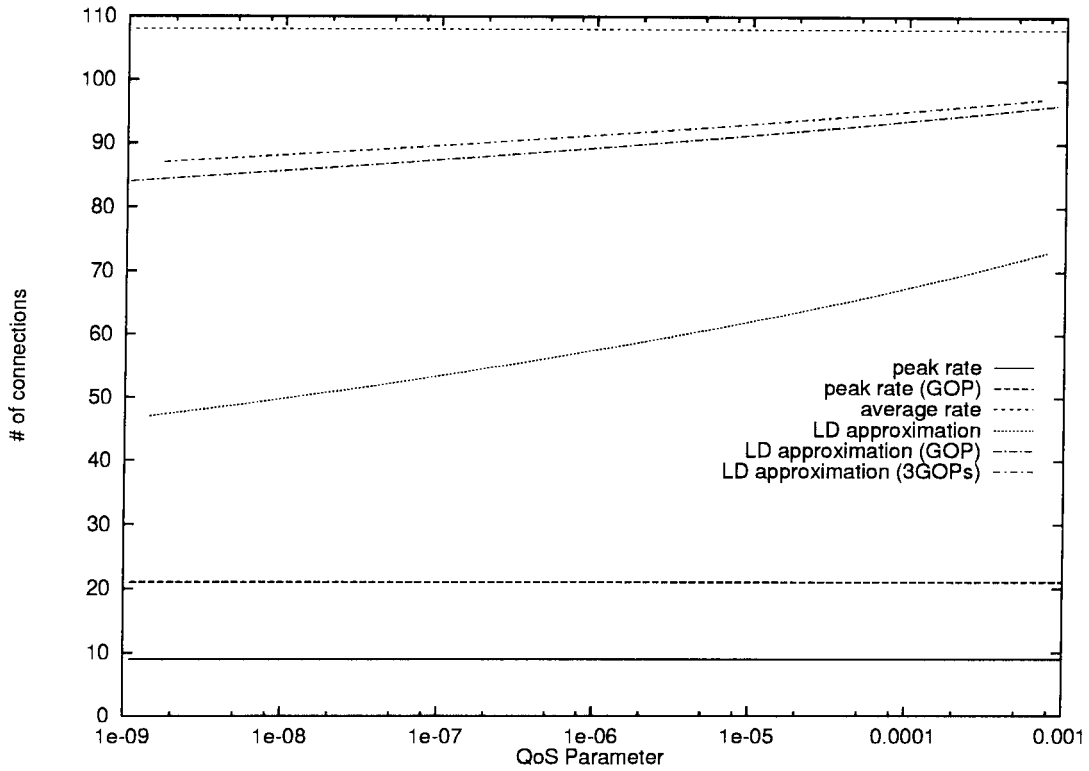

Fig. 7. Effect of smoothing over one GOP or three GOP's on the number of admissible connections as a function of $P_{\text{loss}}^{\text{time}}$.

Now, suppose that a new video $J+1$ requests connection establishment. The admission control will proceed as follows. First, find the s^* that satisfies $\mu'_X(s^*)+\mu'_{X(J+1)}(s^*) = a$. This can best be done with Newton's method [18] starting from the s^* of the last admission test. The new connection is accepted if and only if (9) is satisfied. This procedure can require an excessive amount of CPU time for real-time implementation. For this reason, we now investigate computational procedures for accelerating the run time. (We note that such procedures are not needed for the scheme of Grossglauser *et al.* [5] because their scheme employs a relatively small number of rates.)

Considerable speed-up of the described computation can be achieved by reducing the resolution of the histogram $\pi_j(l)$. So far, our calculations are based on $\pi_j(l)$ computed according to (1) for $l = x_{\min}, \cdots, x_{\max}$, where the step size for l is 1 bit. In order to reduce the resolution, we may compute the histogram as

$$\pi_j(l) = \frac{1}{N}\sum_{n=1}^{N} 1(l - \text{binsize} < x_n(j) \le l) \quad \text{for } l = x_{\min}, x_{\min} + \text{binsize}, \cdots, x_{\max}.$$

Collecting the frames into bins in this fashion ensures that the probability of cell loss is an increasing function of the bin size, and hence leads to conservative admission decisions. Fig. 8 shows the number of admissible *Star Wars* connections computed with varying resolutions (binsizes) for the *Star Wars* histogram. Table IV gives the typical CPU times required for an admission test. All computations are performed on a Sun SPARCstation 2. The graph shows that increasing the binsize to one ATM cell (384 bits) reduces the number of admissible connections by approximately 1.1% while reducing the CPU time by a factor of about 370.

TABLE IV
CPU TIMES FOR ADMISSION TESTS BASED ON HISTOGRAM WITH VARYING RESOLUTION

binsize	1 bit	1 cell	10 cells	20 cells
CPU time	80 min	13 sec	4.5 sec	0.75 sec

We mention that a quick approximate test can be performed in the following fashion. Store s^* and $\mu_X(s^*)$ of the previous admission test, compute only $\mu_{X(J+1)}(s^*)$ on line without changing s^* (takes approximately 1.4 s for a histogram with resolution of 1 bit, 3.7 ms for a resolution of one cell), and check if (9) holds. If the new connection is accepted, update s^* and $\mu_X(s^*)$ while the new video is being transmitted. Note that updates must also be done when a connection terminates. This procedure is motivated by the fact that s^* typically varies only slightly when adding a new connection. Furthermore, using a suboptimal s^* leads to conservative acceptance decisions since the expression in the exponent of the Chernoff bound (5), which dominates the LD approximation, is strictly convex [2, p. 1119]. We refer to this procedure as the *on-line procedure*.

The admission control algorithms discussed so far compute the logarithmic generating function $\mu_X(s)$ directly from the histogram of the frame sizes. As an alternative, we apply an idea of Hui [8, p. 206] to our problem at hand: we expand the logarithmic generating function in a Taylor series, and base the acceptance decision on precomputed coefficients representing the videos. To this end, we first note that the moment generating function of the jth video may easily be

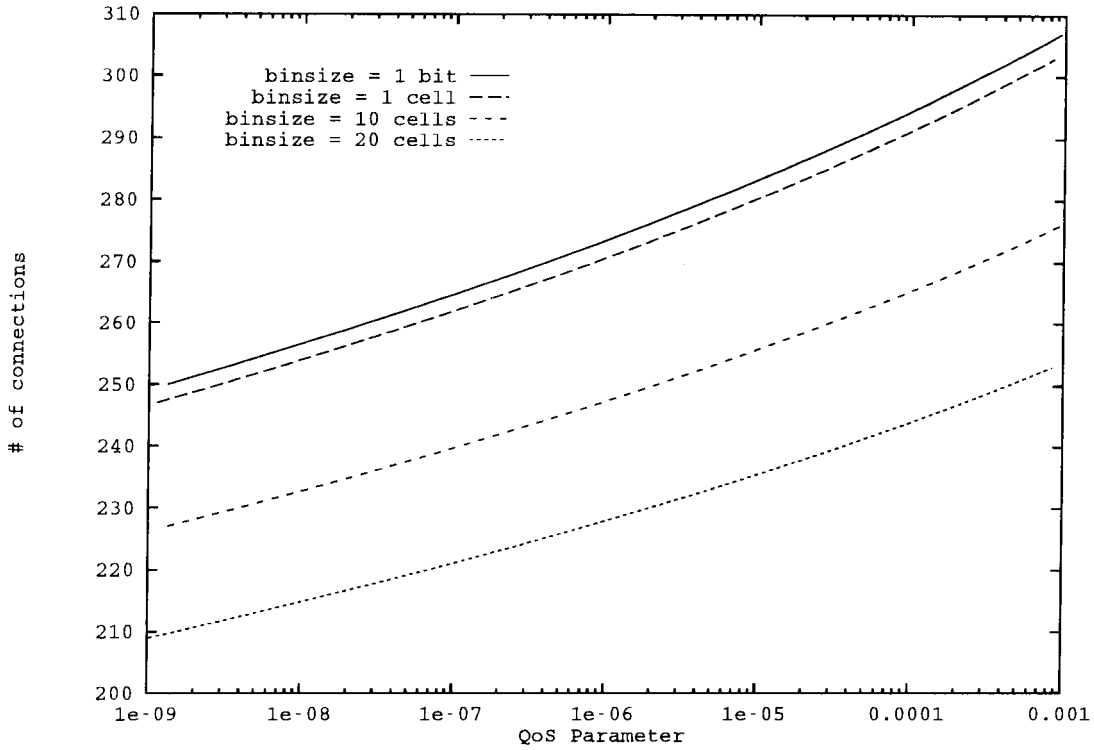

Fig. 8. Number of connections with varying resolution of the histogram as a function of $P_{\text{loss}}^{\text{time}}$.

developed into a Taylor series

$$M_{X(j)}(s) = \sum_{l=x_{\min}}^{x_{\max}} \pi_j(l) e^{sl} = \sum_{l=x_{\min}}^{x_{\max}} \left[\pi_j(l) \sum_{u=0}^{\infty} \frac{(sl)^u}{u!} \right] = \sum_{i=0}^{\infty} a_i(j) s^i$$

where

$$a_i(j) \equiv \frac{1}{i!} \sum_{l=x_{\min}}^{x_{\max}} \pi_j(l) l^i.$$

We note that the frame sizes l (as well as the link capacity a in the conditions of Section III-B) should be scaled to improve the convergence of the admission region obtained from the series approximation to the admission region computed directly from the histogram. We achieved rapid convergence (see Fig. 10) with a scale factor q of 60 cells for the unsmoothed *Star Wars* trace. This factor is somewhat higher than m, the average frame size of the *Star Wars* video, and ensures that $E[X/q] < 1$. Without this scaling, the l^i in the expression for $a_i(j)$ can easily lead to huge values; we conjecture that the common scale factor used for admission control should be larger than the largest $m(j)$ of all videos available on the video server.

In a second step, we may compute the series expansion of the logarithmic moment generating function

$$\mu_{X(j)}(s) = \ln M_{X(j)}(s) = \ln \sum_{i=0}^{\infty} a_i(j) s^i = \sum_{k=1}^{\infty} c_k(j) s^k \tag{11}$$

where the coefficients $c_k(j)$ are given by (see Hui [8, p. 206])

$$c_k(j) = a_k(j) - \frac{1}{k} \sum_{i=1}^{k-1} i a_{k-i}(j) c_i(j).$$

The coefficients for the *Star Wars* video, computed with a scale factor of 60 cells, are depicted in Fig. 9. Our experiments seem to indicate that a relative small number K of coefficients in (11) approximates the logarithmic moment-generating function with sufficient accuracy for the purpose of admission control. Fig. 10 shows the number of admissible *Star Wars* connections computed from the series expansion of μ_X with different K. The figure shows that the admission region obtained from the series approach converges rather quickly to the admission region calculated with the direct approach. For $K \geq 4$, both curves are almost indistinguishable. This seems to indicate that videos can be characterized by a set of coefficients $\{c_k(j), 1 \leq k \leq K\}$. These coefficients can be computed off line and stored with the actual video. Given these video descriptors, the admission control tests of Section III-B can now be conducted very efficiently by noting that

$$\mu_X(s) = \sum_{k=1}^{K} c_k s^k$$

where

$$c_k = \sum_{j=1}^{J} c_k(j), \qquad 1 \leq k \leq K.$$

This method avoids the expensive computation of the sum of exponentials in the direct approach (10), and is furthermore independent of the number of videos already in progress.

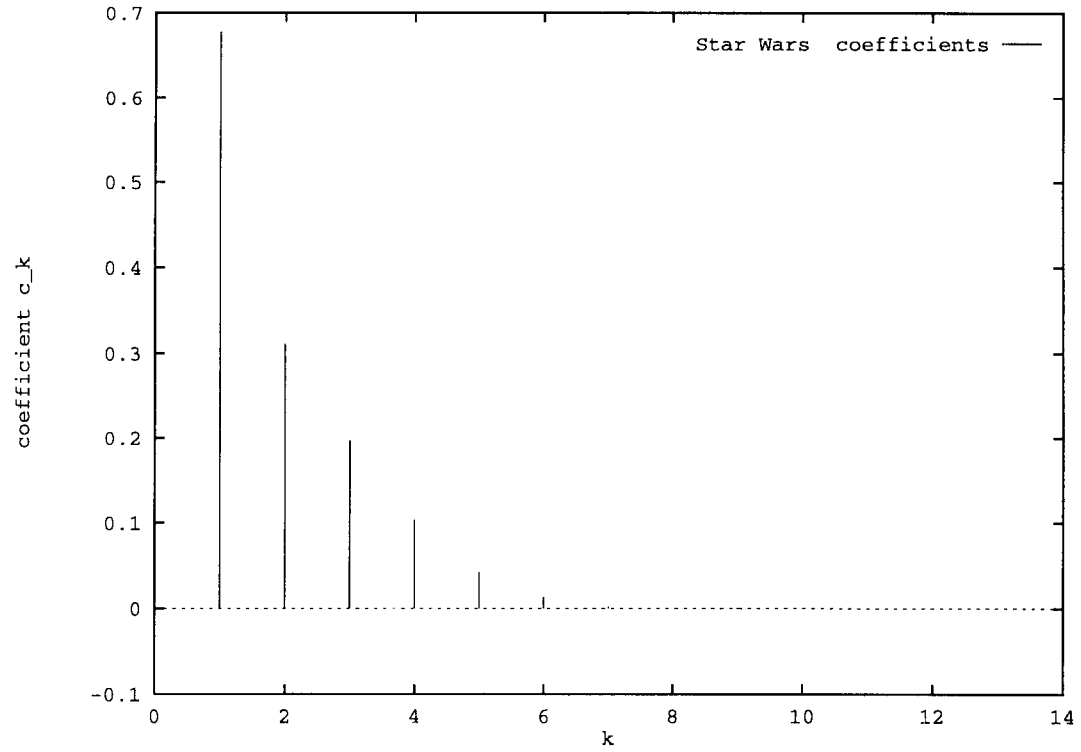

Fig. 9. c_k coefficients for *Star Wars* video.

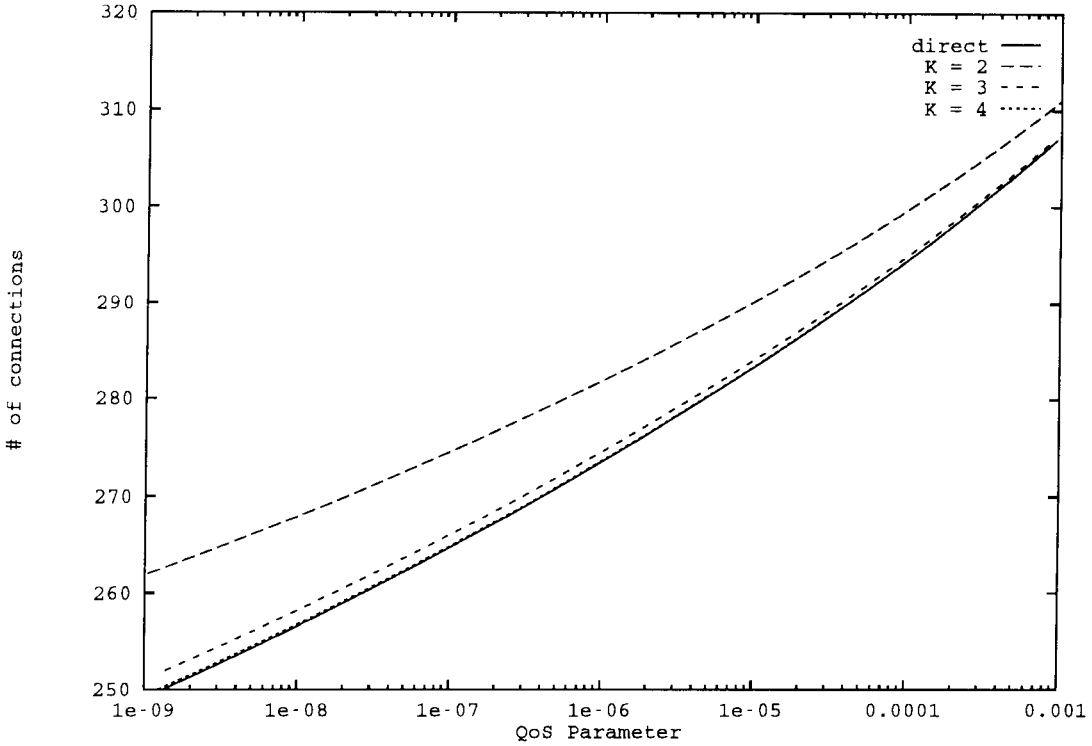

Fig. 10. Effect of the number of coefficients in Taylor series approximation as a function of $P_{\text{loss}}^{\text{time}}$.

If a new video $J+1$ requests connection establishment, we first update $c_k \leftarrow c_k + c_k(J+1), 1 \leq k \leq K$, and find the s^* that satisfies $\mu'_X(s^*) = a$, starting Newton's method with the s^* found in the last admission test. Finally, we admit the new connection if and only if (9) is satisfied. This admission test takes approximately 21 ms for $K = 6$, and is hence viable for real-time admission control.

Owing to the results in this section, we believe that there are two options for real-time admission control employing the LD approximation: 1) the on-line procedure discussed earlier,

and 2) the test based on the precomputed $c_k(j)$ coefficients as we just described. We acknowledge that it would be desirable to support these conclusions with additional experiments using other video traces as well as heterogeneous mixes of traces. Unfortunately, there is currently a lack of publicly available traces that give a correct representation of the bandwidth requirements of MPEG-compressed videos.

VI. A Refined Admission Control Procedure

We begin this section by supposing that VCR control (pause, rewind, fast forward) is no longer permitted. We do suppose, however, that the various videos begin playback at different times. For $n = 1, \cdots, N$, define $X_n(j)$ as in Section II. But now define $X_n(j) = 0$ for all $n > N$.

Suppose that, during frame time l, there are J videos in progress, with the jth video having trace $\{x_1(j), x_2(j), \cdots, x_N(j)\}$. Suppose, during frame time l, frame θ_j of video j is transmitted. Now, consider admitting a new video $J+1$ (with trace $\{x_1(J+1), x_2(J+1), \cdots, x_N(J+1)\}$) which is to begin transmission in frame time $l+1$. With this new video, the amount of offered traffic at frame time $n+l$ is

$$x_{n+\theta_1}(1) + \cdots + x_{n+\theta_J}(J) + x_n(J+1).$$

An admission rule which guarantees no loss for the duration of the new video is

$$\sum_{j=1}^{J+1} x_{\theta_j+n}(j) \le a, \qquad n = 1, \cdots, N \tag{12}$$

where $\theta_{J+1} := 0$. An admission rule which permits loss for a fraction of frame periods is

$$\frac{\sum_{n=1}^{N} 1\left[\sum_{j=1}^{J+1} x_{\theta_j+n}(j) > a\right]}{N} \le \epsilon. \tag{13}$$

Similarly, one can easily define admission rules which permit a fraction of bit loss (analogous to $P_{\text{loss}}^{\text{info}}$).

Now, let us once again suppose that VCR features are permitted. The theory developed in Section II takes a global view on the phase profiles: it essentially assumes that, over the course of a video, there will be many phase profiles. An admission control procedure that takes a more myopic view to call admission is one akin to rule (12), but over fewer frame periods, e.g.,

$$\sum_{j=1}^{J+1} x_{\theta_j+n}(j) \le a, \qquad n = 1, \cdots, M \tag{14}$$

where the θ_j's are the current phases at the call admission time and $M \ll N$.

We can define an appealing admission rule by combining one of the global tests in Section III with the test (14) (or with a test similar to (14), such as a test permitting some cell loss). For a given QoS parameter ϵ, such a combined test admits fewer connections than the isolated global test. However, this combined test guards against the possibility of excessive cell loss due to unusual phase profiles at call admission.

```
1.  Fix J, M;
2.  o = 0; p = 0;
3.  For l = 1 to L do
4.    q = 0; I = P;
5.    Draw θ_i ~ DU[0, N/G − 1], i = 1, ..., I;
6.    While (q < 1) and (I ≤ J) do
7.      I = I + 1;
8.      Draw θ_I ~ DU[0, N/G − 1];
9.      For m = 1 to M do
10.       X_m = Σ_{i=1}^{I} x_{m+θ_i}(i);
11.       q = q + 1(X_m > Ga);
12.   o = o + I − 1;
13.   p = p + (I − 1)^2;
14. Î = o/L; (sample mean)
15. Ŝ^2 = 1/(L−1) (p − o^2/L); (sample variance)
```

Fig. 11. Simulation algorithm for combined admission test.

We conducted simulation experiments with the GOP-smoothed *Star Wars* trace to evaluate the combination of global and myopic admission test. Fig. 11 presents the employed simulation algorithm for the combination of any of the global tests of Section III and the myopic test (14). (An admission rule involving a test that is akin to (14), but allows for some loss, can be simulated in a similar fashion.) We first fix the number of connections J allowed by the global admission test for a specific QoS parameter ϵ. We then simulate the transmission of M consecutive GOP's of I videos, where each video has an independent starting phase that is drawn from a discrete uniform distribution over $[0, N/G - 1]$, where G denotes the number of frames per group of pictures (GOP). Starting from P, the number of connections allowed by a peak rate admission scheme, we increase the number of video streams I until we experience loss during the transmission of the M GOP's or hit the limit given by the global admission test. We thus find the maximum number of connections allowed by the combined admission test. We run L replications to find a confidence interval for the expected maximum number of connections.

Fig. 12 shows the number of *Star Wars* connections admitted by the combination of the LD approximation for $P_{\text{loss}}^{\text{info}}$ (9) and the myopic test (14) as the parameter M varies. The latter criterion ensures that there is no loss during the transmission of M GOP's following the admission of a new video connection, provided none of the videos experiences VCR actions. For this simulation run, we fixed $\epsilon = 10^{-4}$ for the condition (9), this gives $J = 357$ GOP smoothed *Star Wars* connections for the global test (see Fig. 5). We ran $L = 40$ replications for $M = 200, 400, \cdots, 2000$. Note that M GOP's correspond to $M/2$ real-time seconds. *Star Wars* has a total number of 14 511 GOP's. We see from the figure that the confidence intervals for the expected number of connections admitted by the combined test are centered only slightly below the $P_{\text{loss}}^{\text{time}}$ limit of 357 connections. This shows that, even for large values of M, the myopic test (14) has little impact on admission control, implying that the global admission test will typically not lead to large losses.

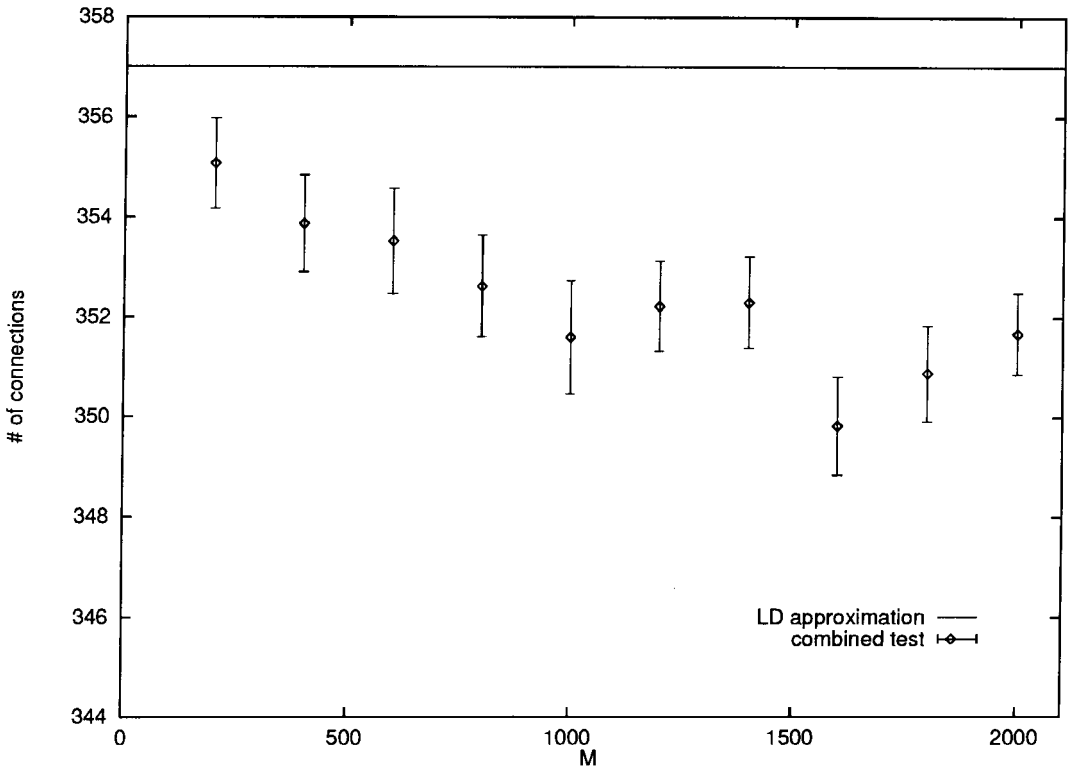

Fig. 12. Combined admission test [LD approximation for $P_{\text{loss}}^{\text{info}}$ and (14)] for varying M.

VII. Conclusion

In this paper, we have developed rules for admitting prerecorded sources to an unbuffered link. We have considered two QoS requirements: the first requires that the fraction of frame periods during which loss occurs be less than a given ϵ; the second requires that the fraction of cells lost be less than ϵ. We have presented several approximations for loss for prerecorded traffic, and for the *Star Wars* trace, we have found that the large deviations approximation is quite accurate. Our numerical results have also shown, for multiple copies of *Star Wars*, that it is possible to get a high degree of statistical multiplexing, particularly when each trace is smoothed over its GOP's. We also observed that the number of allowed connections is often insensitive to the cell-loss requirement ϵ. We then explored efficient on-line calculation of admission control decisions, and indicated two procedures which appear promising. Finally, we refined the global admission test in order to guard against the possibility of excessive cell loss due to unusual phase profiles at call admission.

There are several avenues for further research.

1) Perform more computational tests with mixtures of movies once more traces become available.
2) Investigate the character of human VCR actions, and determine whether they significantly change the frame size distributions $\pi_j(\cdot)$.
3) Determine whether performance significantly improves by including small buffers at the receivers. In this case, it is possible to prefetch frames when spare bandwidth is available [19].

Acknowledgment

The authors would like to thank S. Rajagopal and J. McManus for their comments during the course of this research. They also thank D. Mitra for pointing out [20] to them. They would furthermore like to thank M. Garrett and A. Fernandez of Bellcore for making the *Star Wars* trace available.

References

[1] P. Billingsley, *Probability and Measure*. New York: Wiley, 1979.
[2] A. Elwalid, D. Mitra, and R. H. Wentworth, "A new approach for allocating buffers and bandwidth to heterogeneous regulated traffic in an ATM node," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1115–1127, Aug. 1995.
[3] D. Ferrari, J. Liebeherr, and D. Wrege, "Exact admission control for networks with bounded delay services," Dept. Comput. Sci., Univ. Virginia, Charlottesville, Tech. Rep. CS-94-29, July 1994.
[4] M. W. Garrett, "Contributions toward real-time services on packet networks," Ph.D. dissertation, Columbia Univ., New York, NY, May 1993. ftp address and directory of the used video trace: bellcore.com/pub/vbr.video.trace/.
[5] M. Grossglauser, S. Keshav, and D. Tse. "RCBR: A simple and efficient service for multiple time-scale traffic," in *ACM SIGCOMM*, 1995.
[6] I. Hsu and J. Walrand, "Admission control for ATM networks," in *IMA Workshop Stochastic Networks*, Minneapolis, MN, Mar. 1994.
[7] J. Y. Hui, "Resource allocation for broadband networks," *IEEE J. Select. Areas Commun.*, vol. 6, pp. 1598–1608, Dec. 1988.
[8] ——, *Switching and Traffic Theory for Integrated Broadband Networks*. Boston, MA: Kluwer, 1990.
[9] A. Hung and G. Kesidis, "Resource management of prerecorded VBR sources in ATM networks," Dept. EECE, Univ. Waterloo, Waterloo, Ont., Canada, Tech. Rep. 95-05.
[10] F. P. Kelly, "Effective bandwidth at multiple class queues," *Queueing Syst.*, vol. 9, pp. 5–16, 1991.
[11] E. Knightly, J. Liebeherr, D. Wrege, and H. Zhang, "Fundamental limits and tradeoffs for providing deterministic guarantees to VBR video traffic," in *Proc. IEEE INFOCOM'95*, Boston, MA, Apr. 1995.

[12] E. W. Knightly and H. Zhang, "Providing end-to-end statistical performance guarantees with bounding interval dependent stochastic models," in *Proc. ACM SIGMETRICS'94*, 1994.
[13] ——, "Traffic characterization and switch utilization using a deterministic bounding interval dependent traffic model," in *Proc. IEEE INFOCOM'95*, Boston, MA, Apr. 1995.
[14] J. Liebeherr and D. Wrege, "Video characterization for multimedia networks with a deterministic service," in *Proc. IEEE INFOCOM'96*, San Francisco, CA, Mar. 1996.
[15] J. M. McManus and K. W. Ross, "Prerecorded VBR sources in ATM networks: Piecewise-constant-rate transmission and transport," Dept. Syst. Eng., Univ. Pennsylvania, Philadelphia, Tech. Rep., 1995.
[16] ——, "A comparison of traffic management schemes for prerecorded video with constant quality service, Dept. Syst. Eng., Univ. Pennsylvania, Phildelphia, Tech. Rep., 1996.
[17] ——, "Video on demand over ATM: Constant-rate transmission and transport," *IEEE J. Select. Areas Commun.*, vol. 14, pp. 1087–1098, Aug. 1996.
[18] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C, The Art of Scientific Computing*, 2nd ed. Cambridge, England: Cambridge Univ. Press, 1992.
[19] M. Reisslein and K. W. Ross, "A prefetching design for VBR-encoded video on demand," Dept. Syst. Eng., Univ. Pennsylvania, Phildelphia, Tech. Rep., 1996.
[20] J. W. Roberts, Ed., "Performance evaluation and design of multiservice networks," Commission of the European Communities, Luxemburg, COST 224 Final Rep., 1992.
[21] J. Salehi, Z.-L. Zhang, J. Kurose, and D. Towsley, "Supporting stored video: Reducing rate variability and end-to-end resource requirements through optimal smoothing," Univ. Massachusetts, Amherst, Tech. Rep., 1995.

Martin Reisslein received the Dipl.-Ing.(FH) degree from the Fachhochschule Dieburg, Germany, in 1994, and the M.S.E. degree from the University of Pennsylvania, Philadelphia, in 1996, both in electrical engineering.

He is currently working towards the Ph.D. degree in systems engineering at the University of Pennsylvania, Philadelphia. His research interests include multimedia networking and stochastic modeling.

Keith W. Ross (S'82–M'86–SM'90) received the B.S. degree from Tufts University, Medford, MA, in 1979, the M.S. degree from Columbia University, New York, NY, in 1981, and the Ph.D. degree from the University of Michigan, Ann Arbor, in 1985.

He is an Associate Professor in the Department of Systems Engineering, University of Pennsylvania, Philadelphia, and holds secondary appointments in the Computer Information Science and the Operations and Informations Management (Wharton) Departments. In 1980, he designed satellite radar systems as an employee of AVCO. He has been a visiting scholar at several research and academic institutions in France. His current research interests are in protocols and traffic management in high-speed telecommunication networks, including local-area networks, wide-area data networks, voice networks, and broad-band integrated services, digital networks. He is currently performing research in ATM and video on demand, and teaches courses on Internet technology and commerce, and on broad-band networking.

Dr. Ross is the recipient of numerous grants from the National Science Foundation and AT&T, and was the Program Chairman of the 1995 INFORMS Telecommunications Conference.