Smoothing, Statistical Multiplexing and Call Admission Control for Stored Video[*]

Zhi-Li Zhang, James Kurose, James Salehi and Don Towsley
Department of Computer Science
University of Massachusetts
Amherst, MA 01003

UMASS CMPSCI Technical Report UM-CS-96-29

Abstract

VBR compressed video is known to exhibit significant, multiple-time-scale rate variability. A number of researchers have considered transmitting stored video from a server to a client using smoothing algorithms to reduce this rate variability. These algorithms exploit client buffering capabilities and determine a "smooth" rate transmission schedule, while ensuring that a client buffer neither overflows nor underflows.

In this paper, we investigate how video smoothing impacts the statistical multiplexing gains available with such traffic and show that a significant amount of statistical multiplexing gain can be still be achieved. We then examine the implication of these results on network resource management and call admission control when transmitting smooth stored video using variable-bit-rate (VBR) service and *statistical Quality-of-Service (QoS) guarantees*. Specifically, we present a call admission control scheme based on a Chernoff bound method that uses a simple, novel traffic model requiring only a few parameters. This scheme provides an easy and flexible mechanism for supporting multiple VBR service classes with different QoS requirements. We evaluate the efficacy of the call admission control scheme over a set of MPEG video traces.

1 Introduction

Support of Quality-of-Service (QoS) guarantees for real-time transport of stored video over high-speed networks is crucial for the success of many distributed digital multimedia applications, including video on-demand server systems, digital libraries, distance learning, and interactive virtual environments. Video, which is typically stored and transmitted in compressed format, can exhibit significant rate variability, often spanning multiple time scales and in some cases demonstrating *self-similar* behavior [9]. The highly bursty nature of VBR compressed video makes network call admission control and resource management a particularly difficult and complicated task. Hence techniques for reducing the burstiness (rate variability) of video are of significant interest.

A number of researchers have considered using video smoothing algorithms to reduce the variability in transmitting stored video from a server to a client across a high-speed network [23, 24, 8, 20, 19, 27]. These algorithms exploit client buffering capabilities to determine a "smooth" rate transmission schedule, while ensuring that a client

[*] This work was supported by NSF under grant CCR-9119922 and NCR 9508274. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors can be contacted at {zhzhang,kurose,salehi,towsley}@cs.umass.edu.

buffer neither overflows nor underflows. Such video smoothing techniques can achieve significant reduction in rate variability. For example, over a set of MPEG video traces, the smoothing technique in [27] is shown to reduce the peak and standard deviation of the transmitted bit rate by approximately 70%-85%, when smoothed into a 1 MB client buffer.

In this paper, we study several aspects of the problem of supporting stored video with variable-bit-rate (VBR) service and *statistical* QoS guarantees. First, we investigate the extent to which video smoothing reduces the amount of potential statistical multiplexing gain. Statistical multiplexing is an important feature that distinguishes packet-switched networks from their circuit-switched counterparts. VBR network service allows the network to exploit statistical multiplexing gain, since bandwidth is shared dynamically among all traffic streams within a service class. This is in contrast to constant-bit-rate (CBR) service, which provides the abstraction of a fixed-bandwidth pipe to each network user. CBR service is a natural choice for supporting *hard, deterministic* guarantees, but may result in low network utilization since the network must allocate sufficient bandwidth to accommodate the user's peak traffic rate. Because of the significant peak rate reduction, video smoothing can clearly improve the network utilization under *CBR* service [24, 20, 27]. In this paper, we explore the possibility of improving network utilization by exploiting statistical multiplexing gain with *VBR* service.

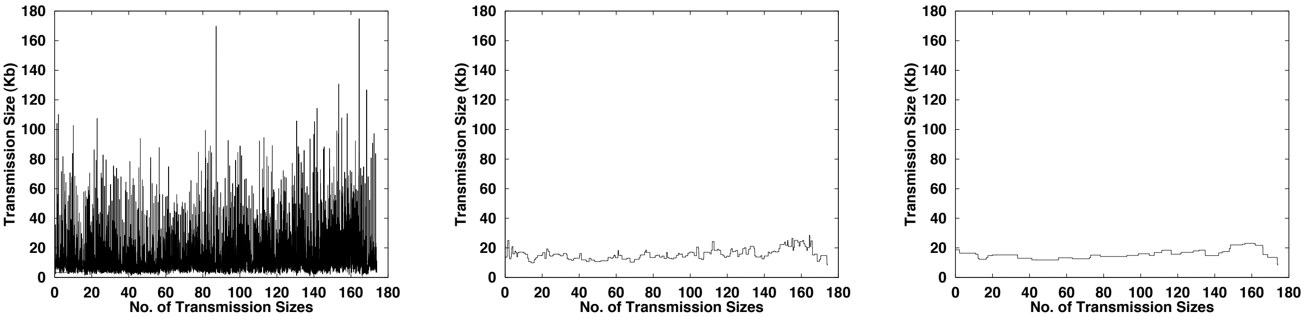
At first glance, there might appear to be only minimal statistical multiplexing gain available with smoothed VBR video traffic, since video smoothing can achieve tremendous reduction in rate variability. However, we find that long-term, slow-time rate variability is still apparent in most smoothed video streams, particularly when client buffers are relatively small. As a consequence, statistical multiplexing gain can still be exploited even after smoothing, thus offering the possibility of reducing the bandwidth required to support a call at a given QoS level, thereby improving network utilization.

In order for VBR service to be a viable alternative to CBR service for real-time video transport, however, it must employ relatively simple, robust resource management and control mechanisms so that the complexity and cost will not offset the utilization gain. A major contribution of this paper is thus to present a call admission control scheme based on a Chernoff bound method [3, 1, 22, 2, 6, 11, 10] that uses a simple, novel traffic model requiring only a few parameters. The Chernoff bound method is shown to provide an effective and robust technique for estimating the potential statistical multiplexing gain and predicting the aggregate bandwidth needed to satisfy a given level of QoS. The traffic model consists of only five parameters that can be easily gathered from a video trace. Our proposed call admission control scheme, coupled with this traffic model, provides an easy and flexible mechanism to support multiple levels of VBR service classes with different QoS requirements.

The remainder of the paper is organized as follows. In Section 2, we study the impact of video smoothing on the statistical characteristics of video traces. In Section 3, the impact of smoothing on statistical multiplexing gain is investigated. We look at call admission control issues for VBR service with statistical QoS guarantees in Section 4. Related work is discussed in Section 5 and the paper is concluded in Section 6.

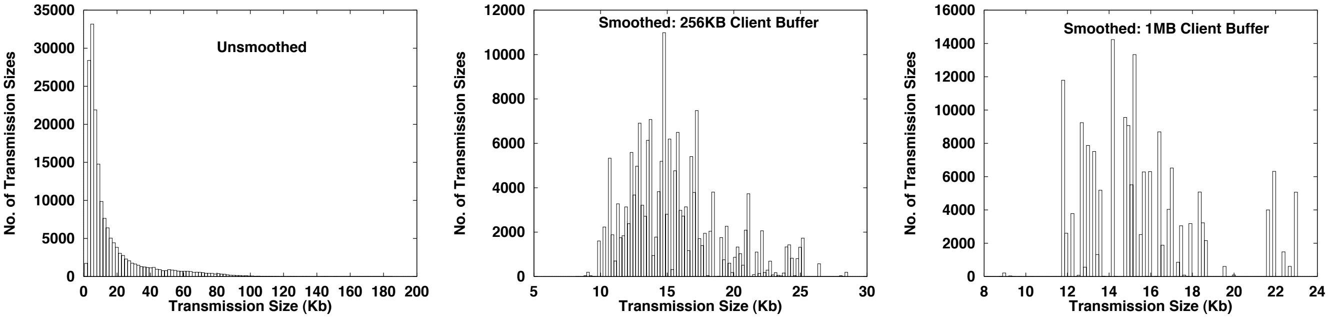
2 Video Smoothing and its Impact on Statistical Characteristics of Smoothed Video

Many multimedia applications transmit stored video streams from a server to a client across a high-speed network. For each stream, the server retrieves data from its video storage system and transfers it onto the high-speed network according to a *transmission schedule*. The client decodes and periodically displays the data it receives from the server. Data arriving ahead of its playback time is stored in a client buffer. In order to ensure continuous playback at the client, the server must transmit the video stream in a manner that ensures the client buffer neither underflows nor overflows.



(a) Unsmoothed (b) Smoothed: 256KB (c) Smoothed: 1MB

Figure 1: Optimal smoothing of a 2-hour MPEG-1 encoding of *Star Wars* latency.

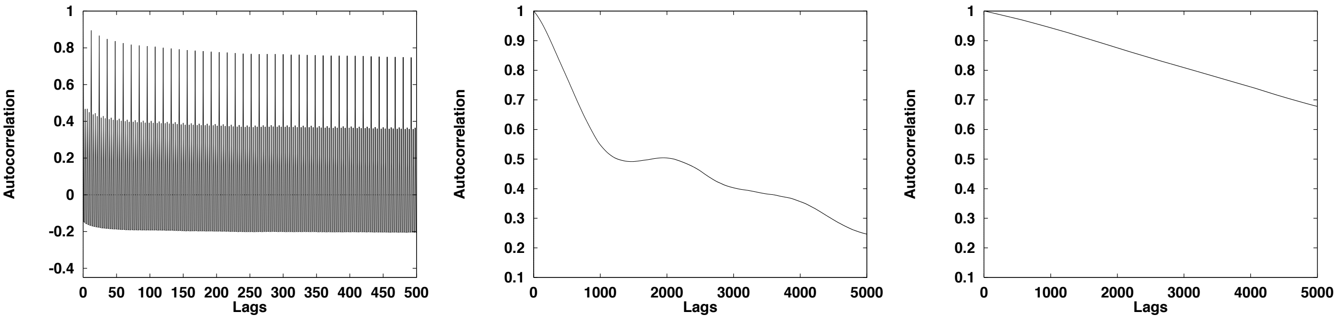


(a) Unsmoothed (b) Smoothed: 256KB (c) Smoothed: 1MB

Figure 2: Impact of the Optimal Smoothing on the Marginal Distributions

Various video smoothing algorithms have been developed [23, 24, 8, 20, 19, 27] that exploit client buffering capabilities to reduce the rate variability existing in VBR-compressed video, while ensuring that a client buffer neither overflows nor underflows. The issue of minimizing buffer requirements for video streams transmitted in a CBR or piece-wise CBR manner is studied in [20, 19]. The authors in [8] examine the issue of minimizing the number of *rate changes* in a server transmission schedule. In [23, 24], video smoothing using client decoder buffer together with a startup delay is studied in a on-line video conferencing setting, and the shortest Euclidean distance algorithm of [16] is used to produce smoothed server transmission schedules under the assumption that frame sizes of the video conference trace are known *a priori*. In [27], a smoothing algorithm is developed that achieves maximal reduction in rate variability for stored video, producing the "smoothest" possible server transmission schedules. The intuitive notion of "smoothness" is formalized using the concept of *majorization* [18], and the optimality of the smoothing algorithm is formally established. Among other things, the optimal smoothing algorithm in [27] produces a transmission schedule that has minimal peak rate and variance for a given client buffer size. Because it reduces rate variability, we will use this algorithm as the smoothing technique in this paper.

Figure 1 visually demonstrates the effect of video smoothing by plotting the transmission sizes (i.e., number of bits per frame time unit, at 24 frames/s), over a two-hour MPEG-1 encoding of *Star Wars* [9]. Both the unsmoothed transmission schedule (a) as well as the smoothed schedules for client buffer sizes of 256 KB (b) and 1 MB (c) are shown. Figure 2 shows the corresponding histograms of the schedules, plotted with 100 bins (note the different scales in the axes). These figures indicate that smoothing significantly reduces the range of transmission sizes – from 0-200 Kb per frame time unit in the unsmoothed schedule, down to 5-30 Kb per frame time unit with a 256 KB client, and 6-24 Kb per frame time unit in the case of 1 MB client buffer. This is a strong indication that rate



(a) Unsmoothed (b) Smoothed: 256KB (c) Smoothed: 1MB

Figure 3: Impact of the Optimal Smoothing on Autocorrelation Structures

variability has been significantly reduced. Note that the transmission schedule in the 1 MB case contains a relatively small number of long, constant rate segments. Furthermore, note that the histogram of a smoothed schedule looks very different from that of the unsmoothed schedule. In particular, the tail distribution of these histograms have very different forms: the long, heavy "tail" of the unsmoothed *Star Wars* trace is transformed into disconnected, conspicuously outstanding "spikes" after smoothing into a 1 MB client buffer.

This drastically altered marginal distribution of smoothed video streams has important consequences for traffic modeling. For example, the traffic modeling techniques presented in [9, 15, 25] that characterize the "heavy-tailed" marginal distributions are not applicable to the smoothed video traces[1]. Different techniques are needed for modeling smoothed video traces. In Section 4, we present a simple technique for characterizing the marginal distribution that is applicable for both smoothed and unsmoothed video streams. The technique is developed for the purpose of call admission control.

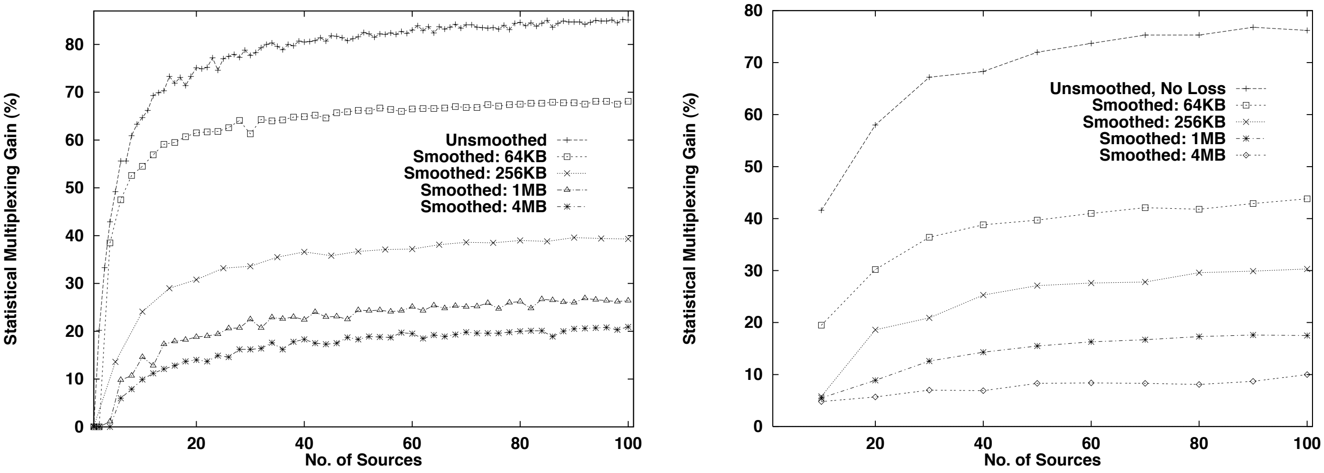
The autocorrelation functions[2] of the unsmoothed and smoothed video traces are shown in Figure 3. Due to the MPEG encoding scheme, the unsmoothed trace demonstrates strong periodic correlation. In Figures 3 (b) and (c) this periodicity has been removed by video smoothing. However, the slowing decaying correlations at large time lags indicate that the traces are still highly correlated. This is because the smoothed video traces consist of many relatively long CBR segments. In the frequency domain, the power spectrums[3] of the video traces (figures of which are not included here due to space limitations) indicate that the variability that still exists is due mostly to slow-time scale variations, while the fast-time scale variability has essentially been removed. This observation can also be visually verified from Figure 1, where we see that the smoothed video streams consist of relatively long CBR segments.

The reduction or removal of fast-time scale rate variability has implications on network resource management, especially buffer allocation within the network. The study in [12, 17] has shown that buffering is only effective in reducing losses due to variability in the high frequency domain, and is not effective for handling variability in low frequency domain. To accommodate low-frequency variability, sufficient bandwidth *must* be allocated in order to maintain the targeted QoS guarantee. This is particularly true in the case of smoothed video streams: because the streams are highly correlated, insufficient bandwidth at one point is likely to lead to consecutive losses over a relatively long period of time, thus greatly affecting the QoS of a client. Consequently, in supporting transport of

[1] In the rest of the paper, we will refer to the smoothed transmission schedule of a video trace as the *smoothed* video trace. It is a sequence of transmission sizes (number of bits per frame time unit) produced by the optimal smoothing algorithm.

[2] The autocorrelation function, $\rho(\tau), \tau = 0, \pm 1, \pm 2, \ldots$, of a stationary discrete random process $\{X_t, t = 0, \pm 1, \pm 2, \ldots\}$ is defined as $\rho(\tau) = \frac{E[(X_t - \mu)(X_{t+\tau} - \mu)]}{\sigma^2}$, where E denotes expectation, μ is the mean of X_t, and σ^2 is the variance of X_t.

[3] Power spectrum of a stationary process is the Fourier transform of its autocorrelation function.



(a) *Star Wars*

(b) 10 *different* video traces

Figure 4: Statistical Multiplexing gain: Unsmoothed and Smoothed Streams, No Loss

smoothed video streams with QoS guarantees, network bandwidth allocation becomes especially critical. At the same time, the amount of buffer space needed within the network can be greatly reduced (*i.e.*, to the amount needed in a network switch for temporarily storing data to be forwarded).

Two advantages are realized with minimal buffer allocation in the network. First, queueing delay jitter introduced by buffering within the network is greatly reduced. Therefore, less client buffer space needs to be set aside to accommodate the network jitter, thus achieving greater reduction in rate variability [27]. From the client's perspective, this also means reduced latency in playback. Second, minimal buffering at the network limits the effect of the autocorrelation structure of the user's traffic on the overall average loss rate. Hence, the difficult task of characterizing the correlation structure of the user traffic is much less important, suggesting that only marginal distribution information (*e.g.*, Figure 2) is needed in traffic specification. For these reasons, we model a network switch as a bufferless multiplexer in the remainder of the paper.

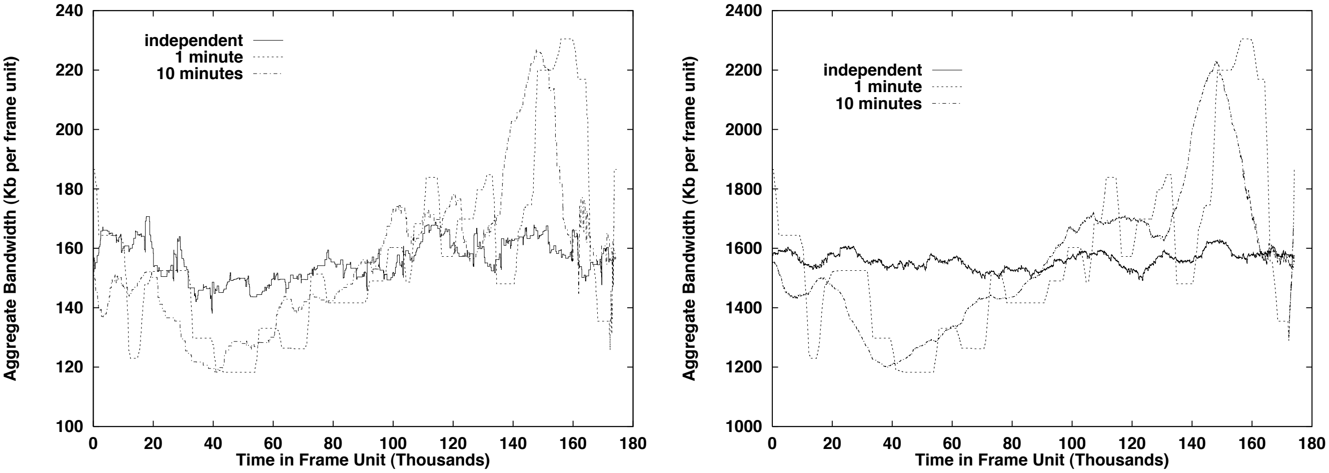
3 Statistical Multiplexing of Smoothed Video Streams

As shown in the previous section, slow-time scale variability still exists in smoothed video streams, particularly with relatively small client buffers. In this section, we empirically determine the amount of statistical multiplexing gain that can be realized when smoothed video streams are aggregated at a network switch or router.

An important assumption underlying most analyses of statistical multiplexing gain is that traffic from different sources are independent of each other. We first evaluate the potential statistical multiplexing gain of smoothed video streams under this independent source assumption, and then investigate the effect of correlated arrivals. Finally, we discuss the implication of this statistical multiplexing gain on network service models and QoS guarantees.

3.1 Independent Arrivals

To investigate the statistical multiplexing gain, we use a simple simulation model. We consider a bufferless multiplexer with n *independent* video streams. For a given QoS requirement (say a loss rate of 10^{-6}), we perform 500 independent runs of a simulation to empirically obtain the minimum bandwidth needed to satisfy the given QoS



(a) 10 instances of *Star Wars* trace

(b) 100 instances of *Star Wars* trace

Figure 5: Aggregate Homogeneous Video Streams under Various Arrival Patterns

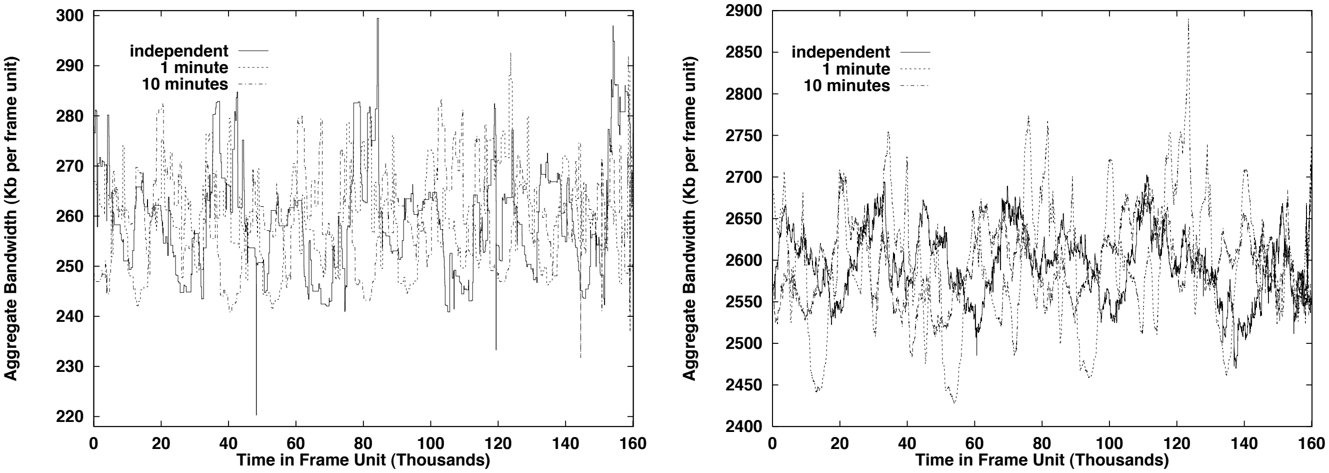
requirement. For each run, we compute the minimum bandwidth required to support the given network load without violating the specified QoS requirement. The maximum value among all runs is used as an indication of bandwidth needed to achieve the target QoS.

In simulating independent arrivals, we assume that the n video streams arriving at the multiplexer are randomly displaced from each other. In other words, for each video stream, the starting frame is equally likely to be any one of the video frames, with appropriate “wrap-around” to ensure that the video streams are of the same length.

To quantify the statistical multiplexing gain, we use the formula $(1 - \frac{r^*}{\hat{r}}) * 100$ as its formal definition, where r^* is the aggregate bandwidth required to satisfy a given QoS requirement (say no loss) for all video streams in the simulation and $\hat{r}$ is the peak rate of the aggregate load (which is the sum of the peak rate of the individual streams). Hence, the statistical multiplexing gain thus defined represents the fractional reduction in aggregate bandwidth requirement needed in the simulation in comparison to peak rate allocation. It thus quantifies the potential utilization improvement that can be realized by VBR service over CBR service with peak rate allocation.

Figure 4 shows the statistical multiplexing gain as a function of number of sources for smoothed video streams with various client buffer sizes, as well as for the unsmoothed video streams. In case (a), all sources are homogeneous, and are generated from the same *Star Wars* trace. In case (b), sources are generated from 10 different video traces. The number of sources from each type of video are increased uniformly. Hence an aggregation of 100 sources consists of 10 sources from each type. The QoS requirement for this example is that no loss occurs at the multiplexer during the entire transmission of the aggregated video streams. The figure indicates that for unsmoothed video streams, a *potential* statistical multiplexing gain of 70%-80% is realizable with VBR service over CBR service with peak rate allocation, while for smoothed streams with various client buffer sizes, a potential statistical multiplexing gain of 10%-60% is realizable. Thus, there are significant statistical multiplexing gains to be exploited by VBR service when individual streams are smoothed, especially when client buffers are relatively small.

In Appendix A, we show that under the independent arrival assumptions, the optimal smoothing algorithm developed in [27] is most likely to yield the *smoothest* aggregate stream in terms of the *peak rate* and *variability*. Hence the statistical multiplexing gains observed when using the optimal smoothing algorithm should, in practice, be lower bounds on the expected statistical multiplexing gains when using other smoothing algorithms.



(a) 1 instance of each type of 10 *different* video traces (b) 10 instances of each type of 10 *different* video traces

Figure 6: Aggregate Heterogeneous Video Streams under Various Arrival Patterns

3.2 Correlated Arrivals

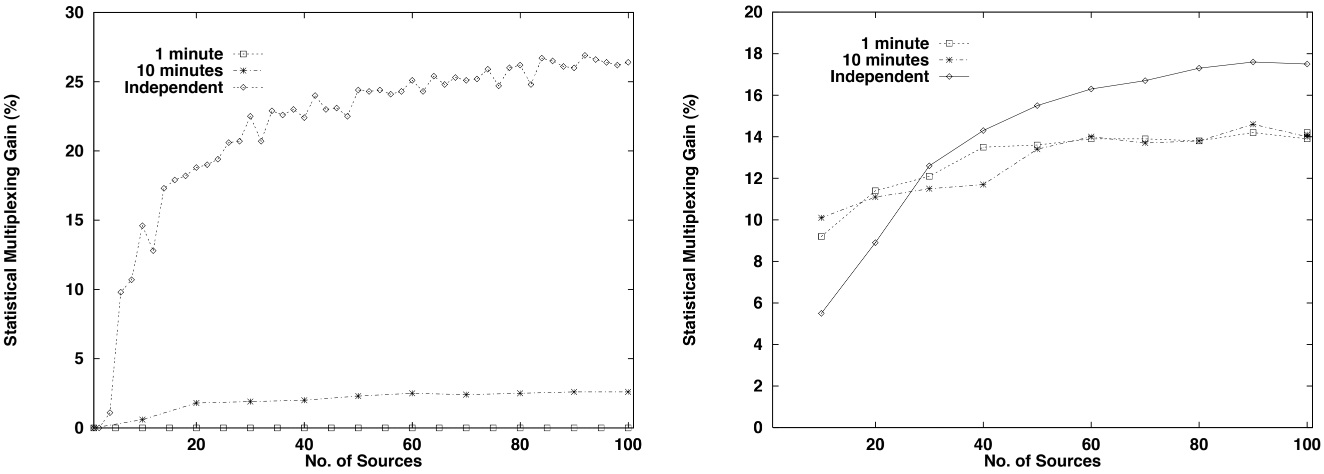
The independent arrival assumption may sometimes be violated in practice. For example, in a video-on-demand system, it is possible that many users may start watching videos within a short time span, thus resulting in correlated arrivals. We next investigate the impact of correlated arrivals on the statistical multiplexing gain.

To investigate this question, we consider scenarios in which all video streams are constrained to begin (*i.e.*, a request for a video playout arrives) within a short time span (*i.e.*, a time span of 1 minute). Within this one minute interval, start times are uniformly, independently and identically distributed. In simulation, this corresponds to randomly choosing the start of a video stream from the first 1 minute segment of the video trace.

Figure 5 illustrates the aggregation of 10 and 100 *Star Wars* sources (smoothed with 1 MB client buffers) under various arrival patterns, where the aggregate instantaneous bandwidth requirement per frame time unit is plotted over the entire duration of the video. The solid line depicts a sample path of the aggregate video stream with independent arrivals, while the two dotted lines depict sample paths of aggregation of video streams when all sources arrive within *1 minute* or *10 minutes* respectively. From the figure, we note that when all sources are homogeneous, the aggregate stream under correlated arrivals is remarkably burstier and has a considerably larger peak rate than under independent arrivals.

Figure 6 illustrates the aggregation of 10 and 100 sources from 10 different video traces (all smoothed with 1 MB client buffers) under the same arrival patterns. In case (a), 10 sources from 10 different video traces are aggregated. In this case, due to the heterogeneous mix of sources, there is little difference in the aggregate streams corresponding to correlated arrivals and independent arrivals. The effect becomes more visible when the number of video sources from the same video traces increases, as shown in case (b), where a total of 100 sources, 10 from each video trace, are aggregated. The maximum aggregate bandwidth requirement in the 1 minute correlated arrival case is considerably larger that that in the independent arrival case (compare the peak of the fine dotted line and that of the solid line). However, the difference between the two cases is less visible in comparison with the homogeneous case consisting only of *Star Wars* streams.

The impact of correlated arrivals on statistical multiplexing gain is shown in Figure 7 where video streams are smoothed into a 1 MB client buffer. Clearly, correlated arrivals have an enormous impact on aggregation of homogeneous sources, leaving almost no statistical multiplexing gain to be exploited. On the other hand, there is a



(a) *Star Wars* (b) 10 *different* video traces

Figure 7: Statistical Multiplexing gain under Correlated Arrivals: Smoothed Video Streams, No Loss

much less severe impact when heterogeneous streams are aggregated and the same number of sources are uniformly dispersed among all types of video streams. In this case, the heterogeneity of the video streams helps alleviate the adverse impact of correlated arrivals on the statistical multiplexing gain.

3.3 Statistical Multiplexing and its Implications on Network Service Models and QoS Guarantees

As shown in section 3.1, VBR service can significantly improve network utilization by exploiting the potential statistical multiplexing gains available with inherently bursty network traffic. However, the potential statistical multiplexing gain can be diminished by correlated arrivals. This observation illustrates an important dimension of network service models — the robustness of network services with QoS guarantees. For a network service model that aims to provide VBR service with *statistical* QoS guarantees by explicitly exploiting statistical multiplexing gain, the term *statistical* takes on two meanings: one at the call level, another at the service level. At the call level, *statistical* QoS guarantees means that QoS fluctuations may occur so long as they remain within the tolerance level specified by the user (*e.g.*, a cell loss rate of at most 10^{-6}), during the call. This is opposed to *deterministic* QoS guarantees, where the QoS (*e.g.*, no cell loss) is hard guaranteed throughout the duration of the call. At the service level, *statistical* guarantees permits the network to fail in providing the promised QoS, for example, in the *rare* event that the users produce correlated traffic. This is in contrast to *guaranteed services*, where as long as the user complies with its traffic specification, the network promises to deliver its QoS guarantee upon which it has agreed with the user. In order to ensure user compliance, traffic specification for guaranteed services must be enforceable and traffic policing and reshaping may be needed within the network.

From the network's perspective, in order to provide for the diverse needs of users, a range of service classes with different levels of service robustness should be provided. By doing so, the network can exploit, to various degrees, potential statistical multiplexing gains and thus increase network utilization while still maintaining the targeted QoS service level. In the next section, we propose a call admission control mechanism that has the flexibility of providing a range of QoS service levels with varying robustness. The tradeoff between the robustness of a network service with QoS guarantees and the realization of statistical multiplexing gain needs to be investigated further and is beyond the effort of this paper.

4 Call Admission Control for Smoothed Video

In the previous section, we demonstrated the potential statistical multiplexing gains available for both smoothed and unsmoothed video streams, and argued for the need to provide a range of QoS guarantee service classes with varying degrees of service robustness. In order to effectively realize the potential statistical multiplexing gains, relatively simple, robust call admission control mechanisms should be employed so that the complexity and cost will not offset the utilization gain. In this section we present a Chernoff-bound-based call admission control scheme and study methods for characterizing the sources' marginal distribution. We propose a simple traffic model with only five parameters. We show that the proposed call admission control scheme, combined with the simple traffic model, provides an easy, effective and flexible mechanism to support multiple levels of VBR service classes with different QoS requirements.

4.1 Chernoff-bound-based Call Admission Control

Consider a bufferless multiplexer where the channel capacity is c. Suppose there are I types of sources, and there are J_i sources of type i, $1 \leq i \leq I$. At any time $t \geq 0$, the amount of traffic arriving from source j of type i is $a_{ij}(t)$. For each type i, we assume that $a_{ij}(t)$ has a stationary distribution given by a K_i-state discrete[4] random variable a_i which takes the values $r_1^{(i)} \leq r_2^{(i)} \leq \ldots \leq r_{K_i}^{(i)}$. In particular, $P\{a_i = r_k^{(i)}\} = p_k^{(i)}$. In other words, with probability $p_k^{(i)}$, the source a_i is in state k, and while in this state, generates $r_k^{(i)}$ amount of traffic. Hence the total amount of traffic at a random time is $a = \sum_{i=1}^{I} \sum_{j=1}^{J_i} a_{ij}$. Given that a_{ij} are all independent, the loss probability at the multiplexer can be estimated by the following well-known *Chernoff Bound* [3, 6] approximation:

$$Pr\{a \geq c\} = Pr\{\sum_{i=1}^{I} \sum_{j=1}^{J_i} a_{ij} \geq c\} \approx e^{-\Lambda^*(c)} \tag{1}$$

where

$$\Lambda^*(\mu) = \sup_{\theta \geq 0} \{\theta\mu - \Lambda(\theta)\} \text{ and } \Lambda(\theta) = \sum_{i=1}^{I} J_i \log M_i(\theta) \tag{2}$$

and $M_i(\theta) = \sum_{k=1}^{K_i} p_k^{(i)} e^{\theta r_k^{(i)}}$ is the moment generating function of a_i.

As $c \to \infty$ with $J_i/c = O(1)$, $1 \leq i \leq I$, the Chernoff Bound (1) can be further refined [22, 2, 1, 6, 10] by adding a prefactor:

$$Pr\{a \geq c\} \approx \frac{1}{\theta^* \sqrt{2\pi\Lambda''(\theta^*)}} e^{-\Lambda^*(c)} \tag{3}$$

where θ^* is the solution to $\Lambda'(\theta) = c$. $\Lambda'(\theta)$ and $\Lambda''(\theta)$ are the first and second derivatives of $\Lambda(\theta)$.

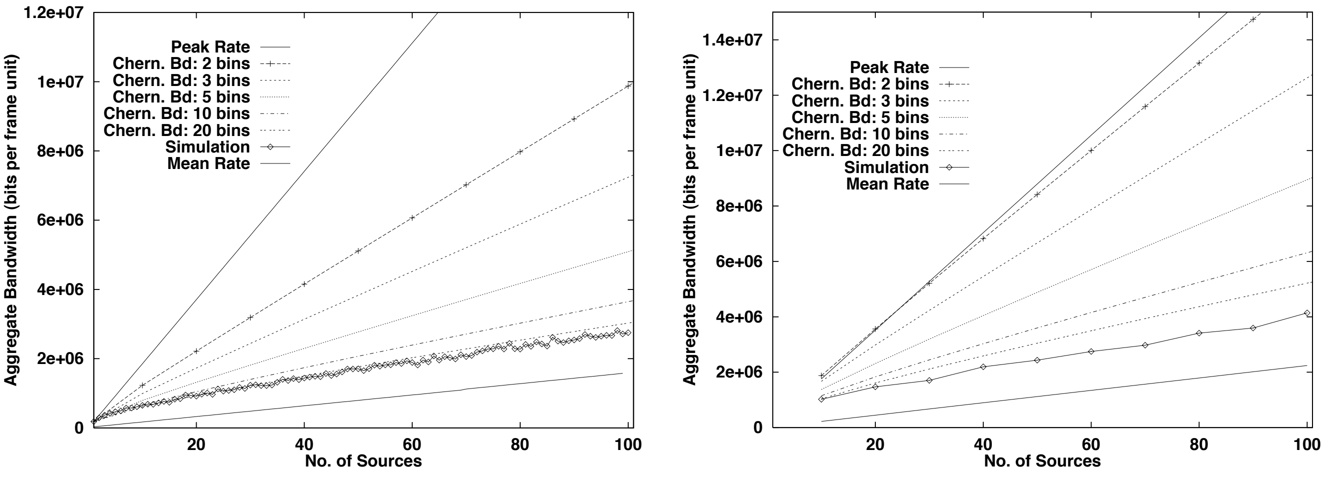
The Chernoff bound can be used to estimate the aggregate bandwidth c^* that is needed to satisfy a given loss probability bound λ at the multiplexer, $Pr\{a \geq c\} \leq \lambda$. The estimated bandwidth c^* is given by the following expression

$$c^* = \sum_{i=1}^{I} J_i \frac{M_i'(\theta^*)}{M_i(\theta^*)} \tag{4}$$

where θ^* is the solution to the following equation

$$\log \lambda = \Lambda(\theta) - \theta\Lambda'(\theta) - \log\theta - \frac{1}{2}\log\Lambda''(\theta) - \log(2\pi). \tag{5}$$

[4] For the sake of simplicity and practicality, we consider only *discrete* random variables.



(a) *Star Wars*, loss rate 10^{-6}

(b) 10 *different* videos

Figure 8: Chernoff Bound Estimation with Histogram: Unsmoothed Streams, Loss Rate 10^{-6}

As the peak rate of the aggregate stream is $\hat{r} = \sum_{i=1}^{I} J_i r_{K_i}^{(i)}$, the statistical multiplexing gain estimated using the Chernoff bound method is $(1 - c^*/\hat{r}) * 100$.

A generic call admission control algorithm based on the Chernoff bound operates as follows. Suppose a new call of source type l arrives. It is accepted if the new aggregate bandwidth estimate c^*, computed using (4) with J_l replaced by $J_l + 1$, is less than c, the channel capacity of the multiplexer.

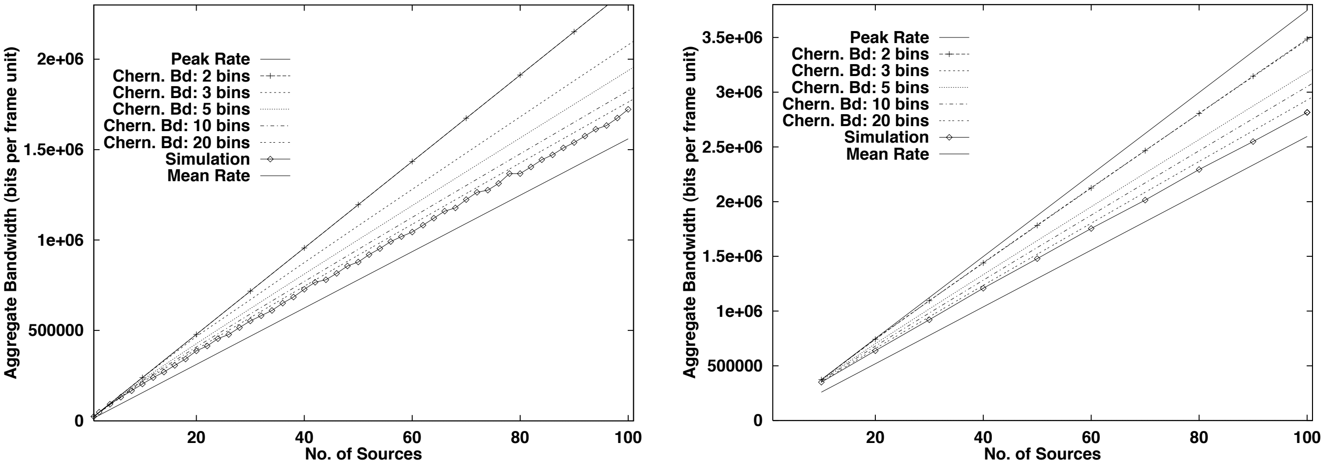
The cost of the call admission algorithm lies mainly in the computation of the marginal moment generating function $M_i(\theta)$ for each source and the solution to the nonlinear equation (5). The latter can generally be solved very fast using the standard Newton-Bisection method. The major cost is associated with the computation of $M_i(\theta)$ and its first and second derivatives used in (4) and (5). The marginal moment generating function is computed from source marginal distribution information $\{(p_k^{(i)}, r^{(i)}), 1 \leq k \leq K_i\}, 1 \leq i \leq I$, provided by the user and maintained by the network.

Clearly, using as few parameters as possible to capture the marginal distribution will not only reduce the computational cost of network call admission control but also the network cost for maintaining relevant state. Therefore, characterizing the marginal distribution of a smoothed or unsmoothed video trace in a manner that permits it to provide sufficient information for the network to exploit statistical multiplexing gains, while at the same time minimizing the associated network cost, is a key question. This will be the focus of the remainder of the paper. This question is particularly challenging, as we have shown that video smoothing drastically alters the marginal distribution of video traces.

4.2 Characterization of Marginal Distribution using Histograms

The histogram method is a standard method for providing a discrete representation of a source marginal distribution. In this section, we evaluate the Chernoff-bound-based call admission control algorithm using the histogram method.

The marginal distribution of a video trace can be characterized using a K-bin histogram as follows. Let $\hat{r}$ be the peak rate of the given trace. We divide the range $(0, \hat{r}]$ into K equal intervals of width $w = \frac{\hat{r}}{K}$ (*i.e.*, bins for histogram). The empirical marginal distribution is then collected by counting number of transmission sizes that fall into each of the K bins. In other words, the marginal distribution is described by a K-state random variable V



(a) *Star Wars*

(b) 10 *different* videos

Figure 9: Chernoff Bound Estimation with Histogram: 512 KB Smoothed Streams, Loss Rate 10^{-6}

with a distribution specified by a set of K (p_k, r_k) pairs. For $1 \leq k \leq K$, the probability that V is in state k is $p_k = \frac{|\{i:(k-1)*w<v_i\leq k*w\}|}{N}$, where $|\cdot|$ denotes the cardinality of a set, v_i denotes the transmission size at frame time i and N is the length of the video, and $r_k = k * w$ is the amount of traffic generated in this state[5].

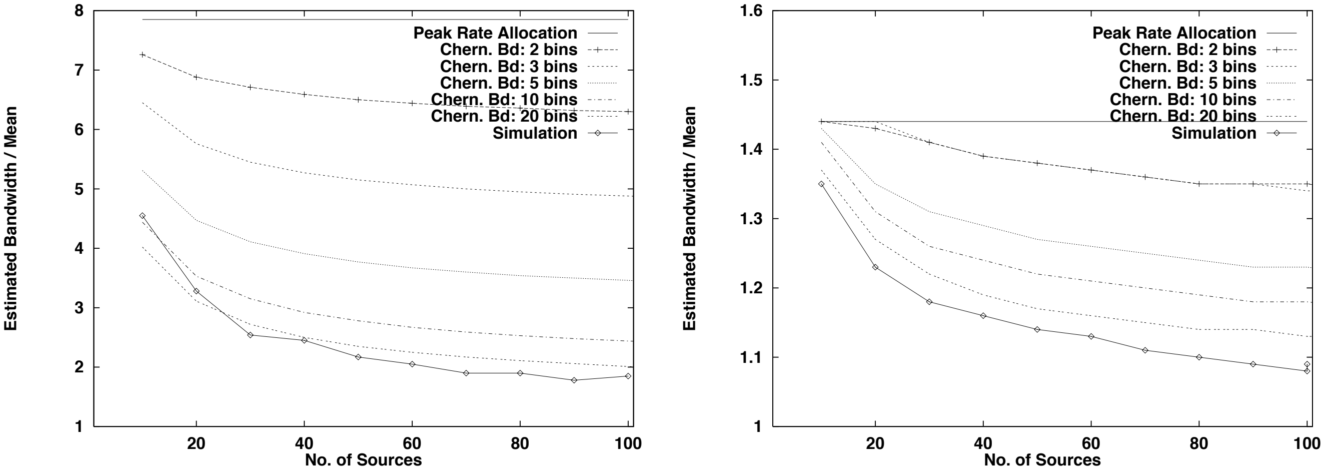
We evaluate the performance of the Chernoff-bound-based call admission control algorithm using histogram characterization of source marginal distributions as follows. For a given loss rate, we compare the bandwidth estimated by the Chernoff bound method using (4) with simulation. The simulation set-up is the same as in Section 3. We perform 500 independent runs, and for each run, compute the minimal bandwidth required to satisfy the given QoS level. The maximum over all 500 runs is taken as the aggregate bandwidth that is needed to satisfy the given QoS level in most cases [6].

The results are shown in Figure 8 for the unsmoothed video streams, and in Figure 9 for the smoothed video streams (with 512 KB client buffers). In both figures, sources in case (a) are homogeneous (generated from the *Star Wars* trace), whereas sources in case (b) are generated from 10 different video traces with an equal number of sources of each type. In all cases, we see that as the number of bins used for describing the marginal distributions increases, the bandwidth requirements estimated by the Chernoff bound method approach the simulation results. This is because with more bins, the marginal distributions of the video traces are more accurately characterized.

Using the same number of bins, the marginal distribution of smoothed video streams can be captured more accurately than that of the unsmoothed video streams, because the smoothed video streams are much less bursty. For example, using $K = 10$ for the smoothed *Star Wars* streams (assuming 512 KB client buffers), the Chernoff bound method predicts that a bandwidth of 183 Kb per frame time unit is needed to transmit 100 smoothed *Star Wars* sources with a loss rate of 10^{-6} as opposed to 239 Kb by peak rate allocation, resulting in a statistical multiplexing gain of 23.6% over the peak rate allocation. The simulation result indicates that a bandwidth of 172 Kb per frame time unit is required, which gives a statistical multiplexing gain of 28.9%. Hence, the Chernoff bound method produces an

[5] Choosing r_k this way results in a histogram that generally has larger mean than the original video trace but the same peak rate. r_k can also be chosen as the mean of all transmission sizes in bin k. This results in a histogram that has the same mean as the original one, but generally with a smaller peak rate.

[6] Another set of independent runs are performed to test the robustness of the aggregate bandwidth thus obtained. For stringent loss rates such as 10^{-5} or 10^{-6} (the latter loss rate essentially yields a lossless transmission), we see almost zero service failure rate. In other words, the maximum bandwidth obtained from the first set of 500 runs is typically sufficient to satisfy the given QoS for the second set of 500 runs.



(a) 10 *different* videos: Unsmoothed

(b) 10 *different* videos: 512KB Smoothed

Figure 10: Comparison of Chernoff Bound Estimation for the Unsmoothed and Smoothed Streams: Loss Rate 10^{-6}

estimate which is only $\frac{183-172}{172} = 6\%$ more than obtained by the simulation. In comparison, using $K = 10$ for the unsmoothed *Star Wars* streams, the Chernoff bound method predicts that a bandwidth of 3647 Kb per frame time unit is needed to transmit 100 unsmoothed *Star Wars* sources with a loss rate of 10^{-6} as opposed to 18526 Kb by peak rate allocation, resulting in a statistical multiplexing gain of 71.3% over the peak rate allocation. The simulation result indicates that a bandwidth of 2750 Kb per frame time unit is required, which gives a statistical multiplexing gain of 75.2%. However, in this case, the Chernoff bound method produces an estimate which is about$\frac{3647-2750}{2750} = 32\%$ more than obtained by the simulation. Therefore, for a given number of bins, the Chernoff bound method produces more accurate bandwidth estimates for smoothed video streams than for unsmoothed video streams, when compared to the simulation results.

In Figure 10, the ratios of the aggregate bandwidth estimated by the Chernoff bound to the aggregate mean rate are shown for unsmoothed and smoothed video streams (with 512 KB client buffers), along with the peak rate allocation and the simulation result. Since the mean bit rate for both smoothed video streams and unsmoothed video streams is the same, comparison of (a) and (b) (note the different scales in y-axes) demonstrates the tremendous impact of video smoothing on bandwidth reduction. Video smoothing greatly reduces the bandwidth required to support a given level of QoS under VBR service, thus improving network utilization. This improvement of network utilization can be effectively realized by using a call admission control algorithm that estimates bandwidth requirement using the Chernoff bound method.

A K-bin histogram requires the specification of $K+1$ parameters by a source: the peak rate $\hat{r}$, and the probabilities of the K bins, $p_1, \ldots, p_K$. By appropriate choice of K, the network can define different levels of service classes with varying robustness of QoS guarantees. For example, by choosing $K = 3$, the network is making rather conservative assumption about the user behaviors, in terms of allocating more bandwidth to provide the requested service to the users. By choosing $K = 5$, or $K = 10$, or larger, the network makes more optimistic and aggressive assumptions about the independent user behaviors (see Figures 8 and 9). Therefore, greater statistical multiplexing gains can be realized, but with the risk of increasing the likelihood of service failure. Larger values of K also result in more overhead and complexity for the network to maintain state and perform call admission control, diminishing the benefits resulting from higher network utilization.

4.3 Parsimonious Bounding Models for Marginal Distribution Characterization

In this section we take a very different approach to the problem of characterizing the sources' marginal distributions for the purpose of call admission control. Consider the following problem: given a user traffic specification described by a set of parameters such as the mean and the peak rate of a source, how should the network construct a marginal distribution that matches the given user parameters? Clearly there are many possible such distributions. Traffic models that make *a priori* assumptions about the user marginal distribution, *e.g.*, that it can be captured by a Gamma or Lognormal or Pareto distribution, have limited applicability. This is particularly true in the light of the impact of video smoothing on the marginal distribution of video traces. Since the network does not have knowledge about the user's marginal distribution beyond the specified user parameters, what assumption should the network make in order to satisfy its QoS? We take a bounding approach in answering this question, and assume that the network should make the *most conservative* assumption so as to account for the "worst-case" marginal distribution that a user may have. This leads to the construction of a marginal distribution such that the bandwidth estimated using the Chernoff bound method yields an *upper* bound on the bandwidth required by any user with any marginal distribution that matches the given set of user-specified parameters.

To address this the problem, we turn to the theory of stochastic ordering. Given two random variables X and Y with respective distributions F and G, we say X is smaller than Y under *increasing convex ordering* (denoted $X \leq_{icx} Y$ or $F \leq_{icx} G$), or informally, X is *stochastically less variable* than Y, if $E[h(X)] \leq E[h(Y)]$ for all increasing, convex functions h. It can be shown (see, *e.g.*, p.271 of [26]) that if X and Y are nonnegative such that $E[X] = E[Y]$, then $X \leq_{icx} Y$ if and only if $E[h(X)] \leq E[h(Y)]$ for all convex h. This ordering is called *variability ordering* in [26]. Intuitively, $X \leq_{icx} Y$ means that X is less variable than Y in the sense that Y gives more weight to the extreme values. In particular, we have that $Var(X) \leq Var(Y)$ and $||X||_\infty \leq ||Y||_\infty$ where $|| \cdot ||$ is the essential supremum of a random variable, defined as $||X||_\infty = \inf\{x : Pr\{X > x\} = 0\}$[7].

With this notion of stochastic variability, the following theorem provides a basis for constructing a worst-case distribution. Informally, the theorem states that among all random variables that have the same user-specified parameters, the random variable that has the worst-case distribution is the one that is *stochastically most variable*.

Theorem 1 *Consider a bufferless multiplexer with channel capacity c. For $1 \leq i \leq I, 1 \leq j \leq J_i$, let a_{ij} denote a random variable with the stationary marginal distribution of source j of type i, and let $\hat{a}_{ij}$ be a corresponding random variable representing the marginal distribution chosen by the network which matches the user specified parameters. In particular, we assume that $E[a_{ij}] = E[\hat{a}_{ij}]$,* i.e., *the mean of the marginal distribution specified by the user is matched by the random variable chosen by the network. Define $a = \sum_{i=1}^{I} \sum_{j=1}^{J_i} a_{ij}$, and $\hat{a} = \sum_{i=1}^{I} \sum_{j=1}^{J_i} \hat{a}_{ij}$. Then, a sufficient condition for the network to provide an upper bound on the loss probability a user may experience,* i.e., *$Pr\{a \geq c\} \leq Pr\{\hat{a} \geq c\}$, as estimated by the Chernoff bound[8] (1), is that $a_{ij} \leq_{icx} \hat{a}_{ij}$ for all i and j.*

Proof: From (1), it suffices to show that $e^{-\Lambda^*(c)} \leq e^{-\hat{\Lambda}^*(c)}$, or $\hat{\Lambda}^*(c) \leq \Lambda^*(c)$. From (2), this is equivalent to

$$\max_{\theta \geq 0}\{\theta c - \hat{\Lambda}(\theta)\} \leq \max_{\theta \geq 0}\{\theta c - \Lambda(\theta)\}. \tag{6}$$

Clearly, (6) holds if $\Lambda(\theta) \leq \hat{\Lambda}(\theta)$ for all $\theta \geq 0$.

Recall that $\Lambda(\theta) = \sum_{i=1}^{I} \sum_{j=1}^{J_i} \log M_{ij}(\theta)$ and $M_{ij}(\theta) = E[e^{\theta a_{ij}}]$. Since $e^{\theta X}$ is a convex function in X and $a_{ij} \leq_{icx} \hat{a}_{ij}$, we have that $\Lambda(\theta) \leq \hat{\Lambda}(\theta)$ for all $\theta \geq 0$. ■

[7] Intuitively, the essential supremum of a random variable is the "peak", or maximal value of X. If X denotes a bounded stationary random arrival rate process, then $||X||_\infty$ is the peak rate of the process.

[8] Since the exponential term in (3) is the dominant term when the number of sources are large, we ignore the prefactor term (*i.e.*, we use (1) instead) in this argument.

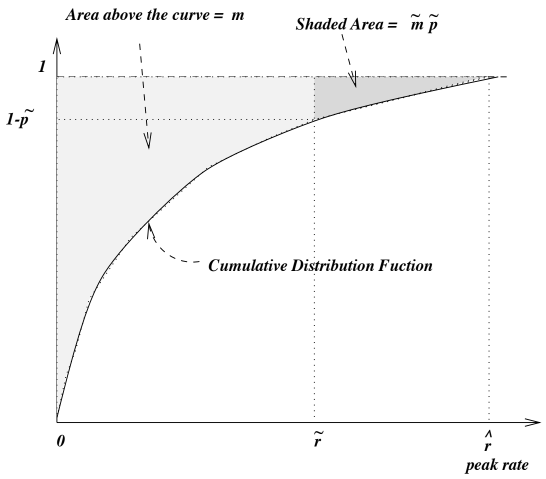


Figure 11: Illustration of the Parameters of the Three-State Model

Remarks

1. Define $L = \max\{a - c, 0\}$ and $\hat{L} = \max\{\hat{a} - c, 0\}$. In other words, L is a random variable representing the amount of loss a user may experience at a given time, and$\hat{L}$ the amount of loss estimated by the network. Then the fact that $a_{ij} \leq_{icx} \hat{a}_{ij}$ for all i and j implies that $a = \sum_{i=1}^{I} \sum_{j}^{J_i} a_{ij} \leq_{icx} \hat{a} = \sum_{i=1}^{I} \sum_{j}^{J_i} \hat{a}_{ij}$. Since $\max\{x, 0\}$ is an increasing convex function in x, we have that $E[L] \leq_{cv} E[\hat{L}]$. Therefore, the average loss experienced by a user is always upper-bounded by that estimated by the network.

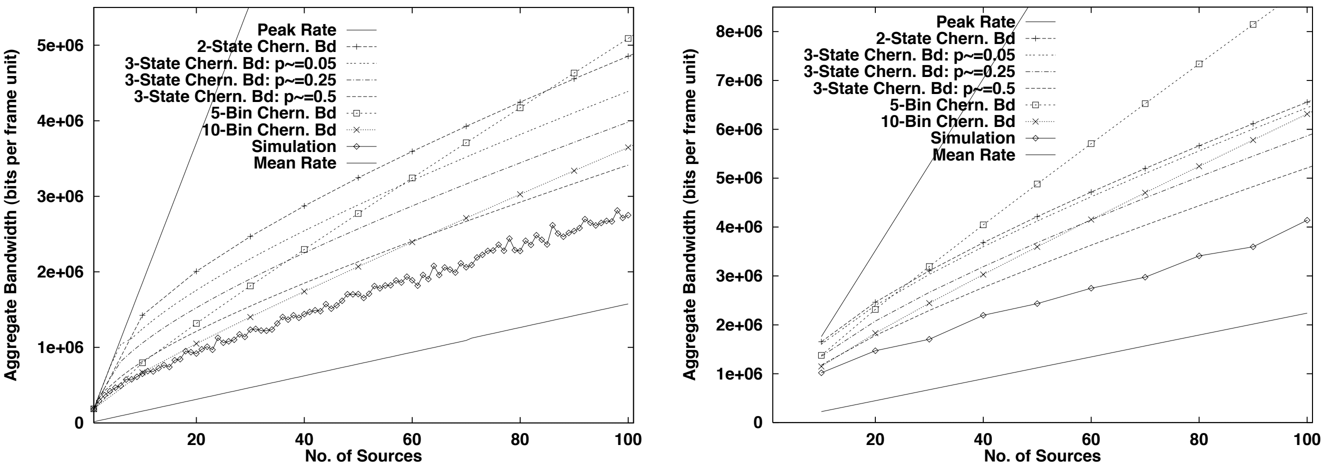
2. Theorem 1 can be strengthened in the following manner. Given that $a \leq_{cx} \hat{a}$, it can be shown that there exists a c_0 such that for $c \geq c_0$, $Pr\{a > c\} \leq Pr\{\hat{a} > c\}$. Hence for $c \to \infty$, $Pr\{a > c\} \leq Pr\{\hat{a} > c\}$. Note that this statement does not require that the loss probability be estimated using the Chernoff bound, as in Theorem 1. On the other hand, as $c \to \infty$, the Chernoff bound provides an asymptotically very tight approximation to the loss probability. Hence the two results are consistent.

4.3.1 Simple Parsimonious Models

Based on Theorem 1, we proceed to construct two simple bounding models which requires only a small number of parameters (*i.e., parsimonious* models). Moreover, these parameters are easy to compute from a video trace.

Perhaps the simplest way to characterize the marginal distribution of a video is to use a model with only two parameters: the peak rate, $\hat{r}$, and the mean rate, m. Among all random variables with the same mean and peak rate, the most *stochastically variable* one, denoted$\hat{X}$, takes two values: $\hat{X} = 0$ with probability $1 - \frac{m}{\hat{r}}$ and $\hat{X} = \hat{r}$ with probability $\frac{m}{\hat{r}}$. $\hat{X}$ has the marginal distribution of a two-state on-off model: it assumes two extreme behaviors of a source, either transmitting at peak rate with probability $m/\hat{r}$, or not transmitting. Thus intuitively,$\hat{X}$ has the "burstiest" behavior. This fact can be easily established formally using the theory of stochastic ordering, the proof of which is relegated to Appendix B.

As we shall see, the two-state model based only on the mean and the peak rate of a source generally does not provide sufficient information about the marginal distribution of the source, therefore resulting in rather conservative bandwidth estimate by the Chernoff bound method. In the following, we thus present a simple "three-state" model to characterize the marginal distribution of a video: in addition to the two parameters representing the mean m and



(a) *Star Wars*

(b) 10 *different* videos

Figure 12: Comparison of Marginal Distribution Models: Unsmoothed Streams, Loss Rate 10^{-6}

the peak $\hat{r}$ of the marginal distribution, we introduce three more parameters to characterize the "tail" of the marginal distribution. Let X be the random variable that has the empirical marginal distribution of a video trace. The three new parameters, $\tilde{r}, \tilde{p}$ and $\tilde{m}$, are defined by the following relations.

$$Pr\{X \geq \tilde{r}\} = \tilde{p} \text{ and } E[X|X \geq \tilde{r}] = \tilde{m}. \tag{7}$$

Intuitively, $\tilde{r}$ defines the rate at which the tail starts, $\tilde{p}$ is the probability that a transmission unit comes from the tail, and $\tilde{m}$ specifies how "heavy" the tail is (while $\hat{r}$ is the "tip" of the tail, and m the center of the mass). The relationship of these parameters is represented visually in Figure 11. The three parameters can be easily computed from a video trace.

Given these parameters, the discrete random variable with the worst-case distribution, $\hat{X}$, is defined as follows. For $0 < \tilde{p} < 1$,

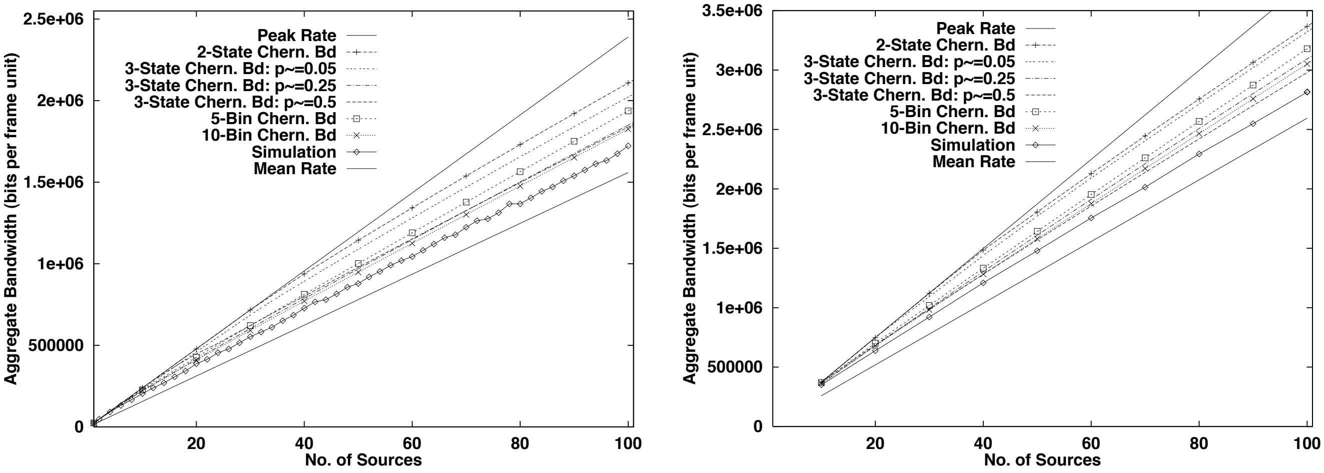
$$\hat{X} = \begin{cases} 0 & \text{with probability } (1 - \frac{\tilde{m}'}{\tilde{r}-1})\tilde{q}; \\ \tilde{r}-1 & \text{with probability } \frac{\tilde{m}'}{\tilde{r}-1}\tilde{q}; \\ \tilde{r} & \text{with probability } (1 - \frac{\tilde{m}-\tilde{r}}{\hat{r}-\tilde{r}})\tilde{p}; \\ \hat{r} & \text{with probability } \frac{\tilde{m}-\tilde{r}}{\hat{r}-\tilde{r}}\tilde{p} \end{cases} \tag{8}$$

where $\tilde{q} = 1 - \tilde{p} = Pr\{X < \tilde{r}\}$ and $\tilde{m}' = E[X|X < \tilde{r}]$. As $m = E[X] = E[X|X < \tilde{r}]Pr\{X < \tilde{r}\} + E[X|X \geq \tilde{r}]Pr\{X \geq \tilde{r}\} = \tilde{m}'\tilde{q} + \tilde{m}\tilde{p}$. Hence $\tilde{m}' = \frac{m - \tilde{m}\tilde{p}}{\tilde{q}}$. We refer to $\hat{X}$ as a "three-state variable" since $\tilde{r} - 1$ and $\tilde{r}$ can be essentially treated as a single state of $\hat{X}$ in practice[9].

In the cases $\tilde{p} = 0$ or $\tilde{p} = 1$, the three-state model degenerates into the two-state model described earlier.

It is easy to check that $E[\hat{X}] = m, \|\hat{X}\|_\infty = \hat{r}, Pr\{\hat{X} \geq \tilde{r}\} = \tilde{p}$ and $E[\hat{X}|\hat{X} \geq \tilde{X}] = \tilde{m}$. We can establish that this 3-state model has the *most stochastically variable* marginal distribution among all discrete random variables X such that $E[\hat{X}] = m, \|\hat{X}\|_\infty = \hat{r}, Pr\{\hat{X} \geq \tilde{r}\} = \tilde{p}$ and $E[\hat{X}|\hat{X} \geq \tilde{X}] = \tilde{m}$. The proof can be found in Appendix B.

[9] In practice, $\hat{r}$ is generally very large. Hence the difference between $\tilde{r} - 1$ and $\tilde{r}$ is negligible. The separation of the two in the definition of $\hat{X}$ is purely due to a technical reason.



(a) *Star Wars*

(b) 10 *different* videos

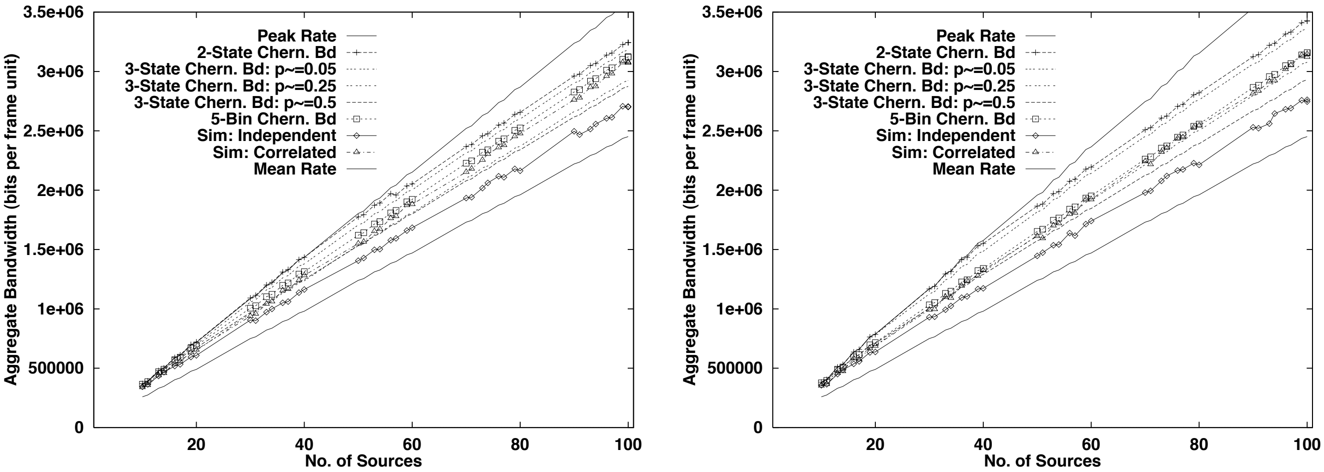
Figure 13: Comparison of Marginal Distribution Models: Smoothed Streams, Loss Rate 10^{-6}

4.3.2 Evaluation

We now examine the performances of the two-state model and the three-state model as the parameter $\tilde{p}$ is varied. Figure 12 shows the performance for unsmoothed video traces, and in Figure 13, smoothed video streams with 512 KB client buffers. For comparison, the performances of the histogram-based method with 5 and 10 bins are also shown in the figures. For $\tilde{p} = 0.5$, the bandwidth estimated by the Chernoff bound method is close to the bandwidth seen by the simulation. As $\tilde{p}$ varies from 0.5 to 0.05 in both figures, the bandwidth estimated using the three-state model approaches the bandwidth estimated using the two-state model. Similar results are obtained by varying $\tilde{r}$ from m to $\hat{r}$ instead of varying $\tilde{p}$. Due to space limitation, these results are not shown here.

In contrast to the histogram based method, the three-state model can provide comparable, if not better, bandwidth estimates with an appropriate choice of $\tilde{p}$. This is achieved without requiring additional parameters in contrast to the histogram-based method. Therefore, without any extra overhead, the three-state model is able to provide bandwidth estimates that range from fairly optimistic (say, by choosing $\tilde{p} = 0.5$ or so) to rather conservative (say, $\tilde{p} = 0.01$ or smaller). This property of the three-state model can be employed by the network to define different levels of service classes. For example, the network can define three different levels of services by choosing $\tilde{p} = 0.5, \tilde{p} = 0.25$ and $\tilde{p} = 0.05$. The user can choose the appropriate service class depending on the level of service robustness required. Since the parameters needed for the traffic specification are fixed and identical for all service classes, the Chernoff-bound-based call admission algorithm has the same implementation.

Two additional examples are shown in Figure 14, where a more diverse mix of video streams is considered. In Example 1, eight of the ten video traces are smoothed using 512 KB client buffers, whereas one trace (*Star Wars*) is smoothed using a 1 MB client buffer, and another trace (*Wizard of Oz*) is smoothed using a 256 KB client buffer. In Example 2, each pair of the ten video traces are smoothed using client buffers of sizes 256 KB, 512 KB, 1 MB, 2 MB and 4 MB respectively . The number of sources of each video type in both cases are not evenly distributed. For eight of the video traces (other than *Star Wars* and *Wizard of Oz*), the number of sources of each type increases gradually from 1 to 5, while the number of *Star Wars* sources increases from 1 to 40 and the number of *Wizard of Oz* sources increases from 1 to 20. To illustrate the need to provide different service levels to account for possible correlated user behaviors, we also consider a scenario of correlated arrivals. In this scenario (the curve labeled "Sim: Correlation" in the figures), the *Star Wars* sources all arrive within a period of 10 minutes, and the *Wizard of Oz*



(a) Example 1

(b) Example 2

Figure 14: Comparison of Marginal Distribution Models: Mixed Smoothed Streams, Loss Rate 10^{-6}

sources within a period of 1 minute.

In this example, the correlated arrivals significantly increase the actual aggregate bandwidth needed to satisfy the desired QoS service level of loss rate of 10^{-6}. Using $\tilde{p} = 0.25$ and $\tilde{p} = 0.5$ for bandwidth estimation in the Chernoff bound method underestimates the bandwidth requirement under such correlated arrivals, thus leading to service failures. The histogram method with 5 bins provides a bandwidth estimation that is barely sufficient. On the other hand, the bandwidth estimated using $\tilde{p} = 0.05$ or by the two-state model is sufficient to accommodate the correlated arrivals with the targeted QoS service level, while still realizing 10%-15% statistical multiplexing gain.

Clearly there is a tradeoff between the robustness of a network service and the amount of statistical multiplexing gain realized. The three-state model we propose here provides a mechanism to balance these two concerns. Appropriate choice of the parameters used in the three-state model plays a critical role in determining the robustness of the QoS guarantees provided by the network. In addition to call admission control, other provisions may be made by either the network or by users to ensure the QoS guarantees can be successfully met. For example, in a video on-demand system, batching [5] of video requests for *hot videos* that arrive within a short period of time, or playback of hot videos at fixed intervals, can be used to alleviate the impact of correlated arrivals.

5 Related Work

There is a vast volume of literature on issues related to statistical multiplexing and call admission control. We will discuss some of the recent work that is most relevant to our work.

The Chernoff bound is a well-known method that has been applied to call admission control with statistical QoS [11, 6, 10, 30]. In [6], a combination of effective bandwidths and the Chernoff bound (called the *Chernoff-Dominant Eigenvalue* method) is proposed for call admission control at a network multiplexer with shared buffers. The method is evaluated using video-conferencing traces. A DAR(1) model is employed to specify the source traffic. However, due to its high burstiness, a DAR(1) model is not appropriate for MPEG compressed video trace. A histogram-based call admission control scheme is proposed in [28], and the loss probability of the aggregate traffic at a network switch is computed using convolution, incurring formidable computational costs when the number of sources is large. In [24], the issue of statistical multiplexing gain is briefly studied using a simple two-parameter model and a call admission control scheme that uses the binomial distribution to estimate loss probability. When the number of sources is large, the computation of the binomial distribution becomes very cumbersome. In this case, the Chernoff bound provides a very good estimate. In [7], a new approach to determining the admissibility of VBR traffic in buffered multiplexers is developed where sources are subject to leaky-bucket regulation. The effectiveness of statistical multiplexing gain for such VBR traffic is then studied.

Recently, several new network services have been proposed which rely on the implicit exploitation of statistical multiplexing gain by adding a renegotiation feature to CBR service [10], and to VBR service with deterministic QoS guarantees [29]. In [10], the entire rate change profile of a renegotiated CBR (RCBR) stream is characterized by a Markovian model and the Chernoff bound method is used for call admission control to limit the probability of service failure. From the call admission control perspective, we can treat an RCBR stream as a VBR stream. When a very small service failure probability is desired, our experience shows that the Chernoff-bound-based call admission control algorithm usually provides a bandwidth estimation that is sufficiently conservative that no renegotiation is actually needed on a per-stream basis to provide the target service level. Hence, VBR service may be likewise employed for such video streams without requiring any explicit renegotiation.

In [4] predictive service is proposed for the future Internet. Predictive service is most appropriate for applications that require QoS guarantees but can tolerate QoS fluctuations. Measurement-based call admission control is proposed and evaluated for predictive service in [13]. Such an approach is an important alternative to the analytic-model-based call admission control proposed in our paper. We believe reduction of variability in video traffic can help the network obtain more stable measurements. However, there are still many issues that remain to be resolved. A key question in measurement-based call admission control is what performance metric to measure, at what time scale to monitor traffic, and how much past history should be taken into consideration. These questions are important in the context of real-time transport of video, due to the slow-time-scale variability and generally long duration of video connections.

Our work, with appropriate modification, can also be applied to predictive service. For example, instead of asking users to provide the parameters used in our simple traffic model, the network can gather, by on-line measurement, the mean m and the tail distribution information $\tilde{m}$ and $\tilde{p}$ with an appropriately chosen $\tilde{r}$. The measured values can be used by the network to explicitly take advantage of statistical multiplexing gain.

Several methods have been used in characterizing the “heavy-tailed” marginal distribution of unsmoothed video traces. For example, in [9] a hybrid model combining Gamma and Pareto distributions is proposed for characterizing the marginal distribution of the JPEG-encoded *Starwars* trace. In particular, the Pareto distribution is used to model the long heavy tail. In [15, 25], the marginal distributions of I, P, B frames are characterized separately using the lognormal distribution. As we have seen, these methods are not applicable to the characterization of the marginal distribution of *smoothed* video streams.

6 Conclusion

In this paper, we have studied the problem of real-time transport of stored video using variable-bit-rate (VBR) service with *statistical* QoS guarantees. In particular, we have investigated the impact of video smoothing on statistical multiplexing gain and its implication in network resource management and call admission control. We started by investigating the issue of statistical multiplexing gain when streams are smoothed and showed how statistical multiplexing gain can be exploited to improve network utilization. We then looked at the issues of call admission control to support VBR service with statistical QoS guarantees. We presented a Chernoff-bound-based call admission control algorithm method that provides an effective mechanism for realizing potential statistical multiplexing gain. We also proposed a simple three-state, five-parameter model for traffic specification. The combined scheme provides a promising, effective and flexible mechanism to support different levels of predictive service with statistical QoS guarantees. We evaluated the efficacy of the scheme over a set of MPEG traces.

In summary, our work supports the contention that by explicitly exploiting multiplexing gain, VBR service with statistical QoS guarantees can provide a viable alternative to CBR service with deterministic QoS guarantees in supporting real-time transport for stored video.

Our work is only an initial study of the problem of real-time transport of stored video; there are still many aspects of the problem that must be investigated. In terms of call admission control, our scheme needs to be further validated in a more complex and dynamic environment. Extending the scheme to incorporate certain measurement-based features is another interesting topic of future research.

Acknowledgements

We would like to thank the researchers who generously shared their MPEG traces. In particular, the contributions of Ed Knightly [14], Marwan Krunz [15], Mark Garrett [9] and Oliver Rose [25] are gratefully acknowledged. We would also like to thank Jayanta Dey and Francesco Lo Presti for many insightful discussions and helpful comments.

A Appendix

In this appendix, we are interested in answering the following question:

> Given n (not necessarily all different) video streams, for each video stream i, let S_i^* be the transmission schedule produced by the optimal smoothing algorithm [27], denoted by $\mathcal{A}^*$, and S_i be any *feasible*[10] transmission schedule produced by an arbitrary smoothing algorithm, denoted by $\mathcal{A}$. Which algorithm is *more likely* to produce a *smoother aggregated* stream with a *lower peak rate* under the independent arrival assumption when the n streams are aggregated at a multiplexer?

This question can be addressed using the stochastic variability ordering introduced in Section 4.3. Recall that given two random variables X and Y with respective distributions F and G, we say X is smaller than Y under *increasing convex ordering* (denoted $X \le_{icx} Y$ or $F \le_{icx} G$), or informally, X is *stochastically less variable* than Y, if $E[h(X)] \le E[h(Y)]$ for all increasing, convex functions h. One important property of increasing convex ordering is the following

Proposition 2 *If $X_1, \ldots, X_n$ are independent and $Y_1, \ldots, Y_n$ are independent, and $X_i \le_{icx} Y_i$, $i = 1, \ldots, n$, then*

$$g(X_1, \ldots, X_n) \le_{icx} g(Y_1, \ldots, Y_n)$$

for all increasing convex functions g.

If X_i and Y_i, $i = 1, \ldots, n$, are all nonnegative, then $X_i \le_{icx} Y_i$ implies that $\sum_{i=1}^n X_i \le_{icx} \sum_{i=1}^n Y_i$ since $g(x_1, \ldots, x_n) = \sum_{i=1}^n x_i$ is an increasing convex function in each x_i.

To apply the increasing convex ordering to the question posed above, we look at the marginal distribution of a smoothed video stream, or equivalently, the corresponding transmission schedule. For each i, $1 \le i \le n$, let $\{v_i^*(t), t = 1, 2, \ldots, N_i\}$ be the optimally smoothed video stream produced by S_i^*, where N_i is the length of the video stream i. For simplicity, we assume that the video stream is stationary. Then its stationary marginal distribution F_i^* can be computed empirically as follows:

$$F_i^*(x) = \frac{|\{t : a = v_i^*(t) \le x\}|}{N_i}.$$

[10] By a *feasible* transmission schedule, we mean a transmission schedule according to which the server never overflows nor underflows the client buffer.

where $| \cdot |$ denotes the cardinality of a set.

Similarly, let $\{v_i(t), t = 1, 2, \ldots, N_i\}$ be the smoothed video stream produced by an arbitrary feasible schedule S. We define its marginal distribution $F_i(x)$ in exactly the same manner.

Let v_i^* and v_i be two random variables with the distributions F_i^* and F_i respectively. We claim that $v_i^* \leq_{icx} v_i$. In [27], it is established that S_i^* is majorized[11] by $S_i, i = 1, \ldots, n,$. Hence we have $E[h(v_i^*)] = \sum_{t=1}^{N_i} \frac{h(v_i^*(t))}{N_i} \leq \sum_{t=1}^{N_i} \frac{h(v_i^*(t))}{N_i} = E[h(v_i)]$, for any convex function h. This, together with Proposition 2, yields the following result.

Theorem 3 *For $i = 1, \ldots, n$, let v_i^* and v_i denote two random variable with the marginal distributions F_i^* and F_i respectively. Then $v_i^* \leq_{icx} v_i$. Consequently, if $v_i^*, i = 1, 2, \ldots, n$ are independent, and $v_i, i = 1, 2, \ldots, n$ are independent, then $\sum_{i=1}^n v_i^* \leq_{icx} \sum_{i=1}^n v_i$.*

The above theorem gives a precise mathematical formulation of the question posed in the beginning of this appendix. It states that if we statistically multiplex n independent video streams produced by $\mathcal{A}^*$ and by $\mathcal{A}$, then at any random point t in time, $\sum_{i=1}^n v_i^*(t) \leq_{icx} \sum_{i=1}^n v_i(t)$. Thus the aggregate stream under $\mathcal{A}^*$ is less variable than the aggregate stream under an arbitrary smoothing algorithm $\mathcal{A}$. In particular, the aggregate stream under $\mathcal{A}^*$ has smaller variance and lower peak rate.

A consequence of Theorem 3 is that if n video streams are fed to a *bufferless* statistical multiplexer with a fixed capacity c, then the average loss suffered by video streams smoothed using the optimal smoothing algorithm $\mathcal{A}^*$ is smaller than that suffered by video streams smoothed using an arbitrary smoothing algorithm $\mathcal{A}$. This follows easily from Theorem 3: let L^* (resp. L) be the random variable representing the amount of the loss suffered by the streams smoothed by the optimal smoothing algorithm $\mathcal{A}^*$ (resp. an arbitrary smoothing algorithm $\mathcal{A}$), then $E[L^*] = E[\max\{\sum_{i=1}^n v_i^* - c, 0\}] \leq E[\max\{\sum_{i=1}^n v_i - c, 0\}] = E[L]$, as $\max\{x, 0\}$ is an increasing convex function in x.

B Appendix

In this appendix, we establish that the random variables constructed by the two-state model and the three state-model have the "worst-case" distribution among all the distributions that match the given user parameters.

Theorem 4
(1) If X is an arbitrary nonnegative random variable such that $E(X) = m$ and $\|X\|_\infty = \hat{r}$, and $\hat{X}$ is defined by $Pr\{\hat{X} = 0\} = 1 - m/\hat{r}$ and $Pr\{\hat{X} = \hat{r}\} = m/\hat{r}$, then $X \leq_{icx} \hat{X}$.
(2) If X is an arbitrary nonnegative discrete random variable such that $E[X] = m$,$\|X\|_\infty = \hat{r}$, $Pr\{X \geq \tilde{r}\} = \tilde{p}$ and $E[X | X \geq \tilde{r}] = \tilde{m}$, and $\hat{X}$ is defined as in (8), then $X \leq_{icx} \hat{X}$.

Before we prove the theorem, we first state an important property of the increasing convex ordering, and then establish a useful lemma using this fact.

Lemma 5 *Let X and Y be two nonnegative random variables with the cumulative distributions F and G respectively. Then $X \leq_{icx} Y$ if and only if for any $a \geq 0$,*

$$\int_a^\infty \bar{F}(x)dx \leq \int_a^\infty \bar{G}(x)dx. \tag{9}$$

where $\bar{F}(x) = 1 - F(x)$ and $\bar{G}(x) = 1 - G(x)$.

[11] See [27] for the definition of majorization and its application to video smoothing.

For a proof, see Proposition 8.5.1 of [26].

Lemma 6 *Let Y_i and Z_i, $i = 1, 2$, be two pairs of nonnegative random variables such that $Y_1 \leq_{icx} Y_2$ and $Z_1 \leq_{icx} Z_2$. Define two new random variables $X_i, i = 1, 2$, as follows:*

$$X_i = \begin{cases} Y_i, & \textit{with probability } p, \\ Z_i, & \textit{with probability } 1-p \end{cases}$$

where $0 \leq p \leq 1$. Then $X_1 \leq_{icx} X_2$.

Proof: For $i = 1, 2$, let F_i, G_i and H_i be the cumulative distributions of Y_i, Z_i and X_i respectively. By the definition of X_i, it is clear that for any $a \geq 0$, $H_i(a) = pF_i(a) + (1-p)G_i(a)$. Then from Lemma 5, it is easy to see that $Y_1 \leq_{icx} Y_2$ and $Z_1 \leq_{icx} Z_2$ implies that $X_1 \leq_{icx} X_2$. ■

Proof of Theorem 4:
(1) Let F and G denote the cumulative distributions of X and$\hat{X}$. Note that $G(x) = \frac{m}{\hat{r}}$ for $0 \leq x < \hat{r}$ and $G(x) = 1$ when $x \geq \hat{r}$. From Lemma 5, it suffices to show that for any $a \geq 0$,

$$\int_a^\infty \bar{F}(x)dx \leq \int_a^\infty \bar{G}(x)dx. \tag{10}$$

Define $\alpha = \inf\{a : F(a) \geq 1 - \frac{m}{\hat{r}}\}$. For any $a \geq \alpha$, if $\hat{r} > x \geq a$, then $F(x) \geq 1 - \frac{m}{\hat{r}} = G(x)$, and for $x \geq \hat{r}$, $F(x) = G(x) = 1$. Hence for any $x \geq a, \bar{F}(x) \leq \bar{G}(x)$. Hence

$$\int_a^\infty \bar{F}(x)dx = \int_a^{\hat{r}} \bar{F}(x)dx \leq \int_a^{\hat{r}} \bar{G}(x)dx = \int_a^\infty \bar{G}(x)dx.$$

For any $0 \leq a < \alpha$, if $0 \leq x \leq a$, then $F(x) < 1 - p = G(x)$. Hence

$$\int_0^a F(x)dx \leq \int_0^a G(x)dx.$$

Therefore,

$$\begin{aligned} \int_a^\infty \bar{F}(x)dx &= \int_0^\infty \bar{F}(x) - \int_0^a \bar{F}(x)dx \\ &= m - \int_0^a (1 - F(x))dx = m - a + \int_0^a F(x)dx \\ &\leq m - a + \int_0^a G(x)dx = \int_a^\infty \bar{G}(x)dx \end{aligned}$$

where in the above we have used the fact that$\int_0^\infty \bar{F}(x) = \int_0^\infty \bar{G}(x) = m$.

(2) Let Y be a discrete random variable with the distribution $Pr\{Y = x\} = Pr\{X = x | X \geq \tilde{r}\}$. Then $E[Y] = \tilde{m}$ and $\|Y\|_\infty = \hat{r}$. Let $\hat{Y}$ be a random variable with the distribution $Pr\{\hat{Y} = \tilde{r}\} = 1 - \frac{\tilde{m}-\tilde{r}}{\hat{r}-\tilde{m}}$ and $Pr\{\hat{Y} = \hat{r}\} = \frac{\tilde{m}-\tilde{r}}{\hat{r}-\tilde{r}}$. Then $E[\hat{Y}] = \tilde{m}$ and $\|\hat{Y}\|_\infty = \hat{r}$. From (1), we see that $Y - \tilde{r} \leq_{icx} \hat{Y} - \tilde{r}$, thus $Y \leq_{icx} \hat{Y}$. Similarly, let Z be a discrete random variable with the distribution $Pr\{Z = x\} = Pr\{X = x | X < \tilde{r}\}$. Then $E[Z] = E[X | X < \tilde{r}] = \tilde{m}'$ and $\|Z\|_\infty < \tilde{r}$. Let $\hat{Z}$ be a random variable with the distribution $Pr\{\hat{Z} = 0\} = 1 - \frac{\tilde{m}'}{\tilde{r}-1}$

and $Pr\{\hat{Z} = \tilde{r} - 1\} = \frac{\tilde{m}'}{\tilde{r}-1}$. Then $E[\hat{Z}] = \tilde{m}'$ and $\|\hat{Y}\|_\infty = \tilde{r} - 1$. Using the same argument as in (1), we can prove that $Z \leq_{icx} \hat{Z}$.

As, for any $x \geq 0$, $Pr\{X = x\} = Pr\{X = x|X \geq \tilde{r}\}Pr\{X \geq \tilde{r}\} + Pr\{X = x|X < \tilde{r}\}Pr\{X < \tilde{r}\} = Pr\{Y = x\}\tilde{p} + Pr\{Z = x\}(1 - \tilde{p})$, and $Pr\{\hat{X} = x\} = Pr\{\hat{Y} = x\}\tilde{p} + Pr\{\hat{Z} = x\}(1 - \tilde{p})$, from Lemma 6, we have that $X \leq_{icx} \hat{X}$. ■

Remarks:

1. In [21], a result to the same effect of Theorem 4 (1) is proved using a different approach.

2. We can extend the three-state model to a K-state model by specifying the following parameters in addition to the mean rate m and the peak rate $\hat{r}$: $Pr\{\tilde{r}_{k-1} \leq X < \tilde{r}_k\} = \tilde{p}_k$, $E[X|\tilde{r}_{k-1} \leq X < \tilde{r}_k] = \tilde{m}_k, 1 \leq k < K$, where $\tilde{r}_0 = 0$ and $\tilde{r}_{K-1} = \hat{r}$. Based on an extension of Lemma 6, the “worst-case” distribution for this K-state model can be constructed likewise.

References

[1] R. R. Bahadur and R. Rao. On deviations of the sample mean. *Ann. Math. Statis.*, 31:1015–1027, 1960.

[2] N. R. Chaganty and J. Sethuraman. Strong large deviation and local limit theorems. *Ann. Probab.*, 21(3):1671–1690, 1993.

[3] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statis.*, 23:493–507, 1952.

[4] D. D. Clark, S. Shenker, and L. Zhang. Supporting real-time applications in an integrated services packet network:architecture and mechanism. In *Proc. ACM SIGCOMM*, August 1992.

[5] A. Dan, D. Sitaram, and P. Shahabuddin. Scheduling policies for an on-demand video server with batching. In *Second ACM International Conference on Multimedia (ACM Multimedia)*, pages 15–24, San Francisco, CA, October 1994.

[6] A. Elwalid, D. Heyman, T. V. Lakshman, D. Mitra, and A. Weiss. Fundamental bounds and approximations for atm multiplexers with applications to video teleconferencing. *IEEE/ACM Transactions on Networking*, pages 1004–1016, August 1995.

[7] A. Elwalid, D. Mitra, and R. H. Wentworth. A new approach for allocating buffers and bandiwdth to heterogeneous regulated traffic in an atm node. *IEEE/ACM Transactions on Networking*, pages 1115–1127, August 1995.

[8] W.-c. Feng and S. Sechrest. Smoothing and buffering for delivery of prerecorded compressed video. In *IS&T/SPIE Multimedia Computing and Networking*, pages 234–232, San Jose, CA, February 1995.

[9] M. Garrett and W. Willinger. Analysis, modeling and generation of self-similar VBR video traffic. In *Proc. ACM SIGCOMM*, pages 269–280, London, England UK, August 1994. ACM.

[10] M. Grossglauser, S. Keshav, and D. Tse. RCBR: A simple and efficient service for multiple time-scale traffic. In *Proc. ACM SIGCOMM*, pages 219–230, Boston, MA, August 1995.

[11] J. Y. Hui. *Switching and Traffic Theory for Integrated Broadband Networks*. Boston: Kluwer, 1990.

[12] C.-L. Hwang and S.-Q. Li. On input state space reduction and buffer noneffective region. In *Proc. IEEE INFOCOM*, pages 1018–1028, March 1994.

[13] S. Jamin, P. Danzig, S. Shenker, and L. Zhang. A measurement-based call admission control for integrated serives packet networks. In *Proc. ACM SIGCOMM*, pages 2–13, Boston, MA, August 1995.

[14] Edward W. Knightly, Dallas E. Wrege, Jörg Liebeherr, and Hui Zhang. Fundamental limits and tradoffs of providing deterministic guarantees to VBR video traffic. In *Proc. ACM SIGMETRICS*, pages 98–107, Ottawa, Canada, May 1995.

[15] M. Krunz and H. Hughes. A traffic model for MPEG-coded VBR streams. In *Proc. ACM SIGMETRICS*, pages 47–55, Ottawa, Canada, May 1995.

[16] D. T. Lee and F. P. Preparata. Euclidean shortest path in the presence of rectilinear barriers. *Networks*, 14:393–410, 1984.

[17] S.-Q. Li, S. Chong, and C.-L. Hwang. Link capacity alllocation and network control by filtered input rate in high-speed networks. *IEEE/ACM Transactions on Networking*, 3(1):10–25, February 1995.

[18] A. W. Marshall and I. Olkin. *Inequalities: Theory of Majorization and its Applications*. New York, Academic Press, 1979.

[19] J. M. McManus and K. W. Ross. Prerecorded VBR sources in ATM networks: Piecewise-constant-rate transmission and transport. *Manuscript*, September 1995.

[20] J. M. McManus and K. W. Ross. Video on demand over ATM: Constant-rate transmission and transport. In *Proc. IEEE INFOCOM*, San Francisco, CA, March 1996.

[21] D. Mitra and J. A.Morrison. Independent regulated processes to a shared unbuffered resource which maximize the loss probability. *Preprint*.

[22] V. V. Petrov. On the probabilities of large deviations for sums of independent random variables. *Theory of Prob. and its Applications*, X(2):287–298, 1965.

[23] A. R. Reibman and A. W. Berger. On VBR video teleconferencing over ATM networks. In *Proc. IEEE GLOBECOM*, pages 314–319, 1992.

[24] A. R. Reibman and A. W. Berger. Traffic descriptors for VBR video teleconferencing over ATM networks. *IEEE/ACM Transactions on Networking*, 3(3):329–339, June 1995.

[25] O. Rose. Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems. Technical Report 101, University of Würzburg Institute of Computer Science, February 1995. Many MPEG-1 traces are available via FTP from `ftp-info3.informatik.uni-wuerzburg.de` in `pub/MPEG`.

[26] S. M. Ross. *Stochastic Processes*. New York, Wiley, 1983.

[27] 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. In *ACM International Conference on Measurement and Modeling of Computer Systems (ACM SIGMETRICS)*, Philadelphia, PA, May 1996.

[28] P. Skelly, M. Schwartz, and S. Dixit. A histogram-based model for video traffic behavior in an atm multiplexer. *IEEE/ACM Transactions on Networking*, 1(4):446–459, August 1993.

[29] H. Zhang and E. W. Knightly. A new approach to support delay-sensitive VBR video in packet-switched networks. In *Proc. 5^{th} Workshop on Network and Operating Systems Support for Digital Audio and Video*, pages 275–286, Durham, NH, April 1995.

[30] Z.-L. Zhang, D. Towsley, and J. Kurose. Statistical analysis of the generalized processor sharing scheduling discipline. *IEEE Journal of Selected Areas in Communications*, pages 1071–1080, August 1995.