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

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

To appear in
IEEE Journal of Selected Areas in Communications
Special Issue on Real-Time Video Services in Multimedia Networks

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 gains can still be achieved. We then examine the implication of these results on network resource management and call admission control when transmitting smoothed stored video using variable-bit-rate (VBR) service with *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-1 coded video traces.

1 Introduction

Support for Quality-of-Service (QoS) guarantees for real-time transport of stored video over high-speed networks is crucial to 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 [7]. The highly bursty nature of VBR-compressed, constant-quality video makes network call admission control and resource management a particularly difficult and complicated task. Hence techniques for reducing the burstiness (rate variability) of such video are of significant interest.

*This work was supported by NSF under grant NCR-9206908 and by ARPA under ESD/AVS contract F-19628-92-C-0089. The authors can be contacted at {zhzhang,kurose,salehi,towsley}@cs.umass.edu.

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 [6, 17, 18, 21, 22, 27]. These algorithms exploit client buffering capabilities to determine a "smooth" rate transmission schedule, while ensuring that the client buffer neither overflows nor underflows. Such techniques can achieve significant reduction in rate variability. For example, over a set of MPEG-1 coded 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. These results demonstrate that video smoothing is a powerful technique that will likely be deployed for real-time transport of VBR-compressed stored video.

The objective of this paper is to study the impact of video smoothing on network resource control and management. Specifically, we investigate the suitability of constant-bit-rate (CBR) and variable-bit-rate (VBR) network service models for real-time transport of smoothed video in an ATM environment, and how such an application can be supported. CBR service, introduced as an emulation of circuit-switched networks, provides the abstraction of a fixed-bandwidth pipe to each network user. In contrast, VBR service exploits the cell-switching nature of the underlying infrastructure and allows statistical multiplexing of traffic streams within a service class, thus enabling dynamic bandwidth sharing among the streams. Under CBR service, network resource control and management are very simple. By requiring users to specify only their peak rate requirement, *hard, deterministic* guarantees can be supported with peak rate allocation. Under this scheme, a new session is admitted into the network if and only if the sum of the peak rates of the on-going sessions and the new session is less than the channel capacity allocated for CBR services at all network switches along the route of the new session. Thus for constant-bit-rate traffic such as uncompressed audio and video streams, CBR service is the natural choice of service. On the other hand, for bursty traffic such as constant-quality VBR-compressed video, CBR service can result in low network utilization as a result of the peak rate allocation. VBR service offers the possibility of improving network utilization by exploiting the potential statistical multiplexing gain offered by the bursty traffic. In order for VBR service to be a viable alternative to CBR service, however, it must employ relatively simple, robust resource control and management mechanisms so that the complexity and cost will not offset the utilization gain.

By applying video smoothing techniques to real-time video transmission, the peak rate and rate variability of the smoothed video stream can be significantly reduced, thus improving the network utilization under *CBR* service [8, 22, 27]. However, a completely constant-bit-rate video stream may require an extremely large client buffer and long start-up latency [18]. With relatively small client buffers (say, in the range of 64 KB to 1 MB), smoothed video streams continue to exhibit long-term, slow-time rate variability. As a consequence, *there is still an opportunity to exploit statistical multiplexing gains*, thus offering the possibility of reducing the bandwidth required to support a video stream at a given QoS level and improving network utilization.

In the first part of this paper, we evaluate the potential statistical multiplexing gains of smoothed video streams under VBR service through a simulation-based empirical study, and establish the advantage of VBR service over CBR service in supporting real-time transport of stored video[1]. We investigate the effect of correlated video streams

[1]Note that in the paper, since we are primarily interested in comparing VBR and CBR network services for real-time video transport of *VBR coded* video streams, *statistical multiplexing gain* is defined as the percentage of reduction in bandwidth required under *VBR service*

on statistical multiplexing gains, and demonstrate the need for the network to support multiple QoS service levels with varying robustness (*see* Section 3.3 for the definition of *robustness*). Throughout the paper, *loss rate* is used as the QoS parameter of network services, although other performance metrics (e.g., delay or delay jitter) could be also used as well.

In the second part of the paper, we present a call admission control scheme with a simple, novel traffic model for VBR service that can effectively realize the potential statistical multiplexing gains and is capable of supporting multiple QoS service levels. The call admission control scheme is based on the well-known Chernoff-bound method [5, 8, 9, 23]. Our contribution lies in the traffic model used in the scheme. We propose a parsimonious bounding model approach that uses only a few generic parameters to characterize the marginal distribution of video streams. Specifically, we introduce a new five-parameter traffic model to capture the marginal distribution (in particular, its tail) of an arbitrary video stream, either smoothed or unsmoothed. The bounding properties of this model are established. The parameters can be easily obtained from the stored video. We show that the Chernoff bound method coupled with this traffic model provides an effective and robust technique for estimating the potential statistical multiplexing gain and predicting the aggregate bandwidth needed to satisfy a given QoS requirement. Moreover, by appropriately setting some of the parameters in the traffic model, the network can easily control the performance of the proposed call admission control scheme, thereby providing a flexible mechanism to support multiple levels of VBR service classes with different QoS requirements.

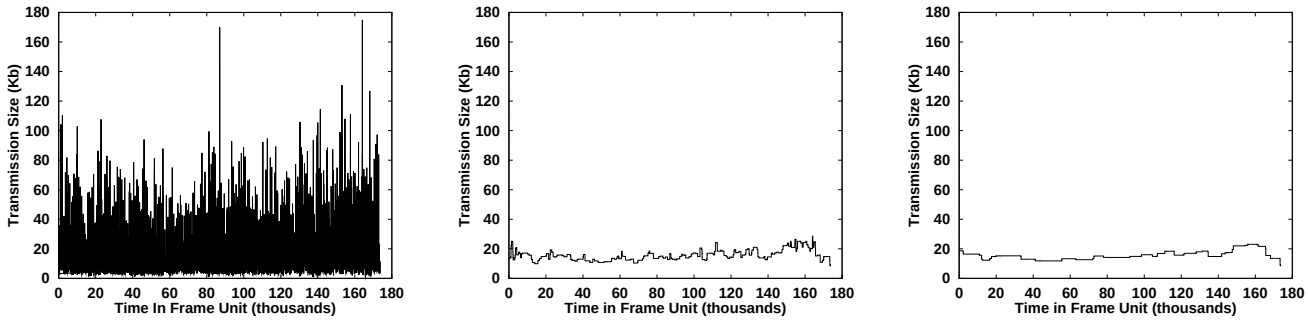
The remainder of the paper is organized as follows. In Section 2, we examine the impact of video smoothing on the statistical characteristics of video traces. In Section 3, the impact of smoothing on statistical multiplexing gains is investigated. We study call admission control issues for VBR service 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 that the client buffer neither underflows nor overflows.

Various video smoothing algorithms have been developed [6, 17, 18, 21, 22, 27] that exploit client buffering capabilities to reduce the rate variability existing in VBR compressed video, while ensuring that the client buffer neither overflows nor underflows. The issue of minimizing buffer requirements for stored video streams transmitted

over that under *CBR service* when transmitting the same set of VBR coded video streams with comparable level of QoS (*see* Section 3.3 for a precise definition). This should not be confused with another definition of statistical multiplexing gain, i.e., the advantage of statistically multiplexed *VBR coded* video over the use of *CBR coded* video. This definition has also been frequently used in the literature, in particular, in the study of video coding techniques, *see*, e.g., [22], where the relative merits of VBR coded video over CBR coded video are studied.



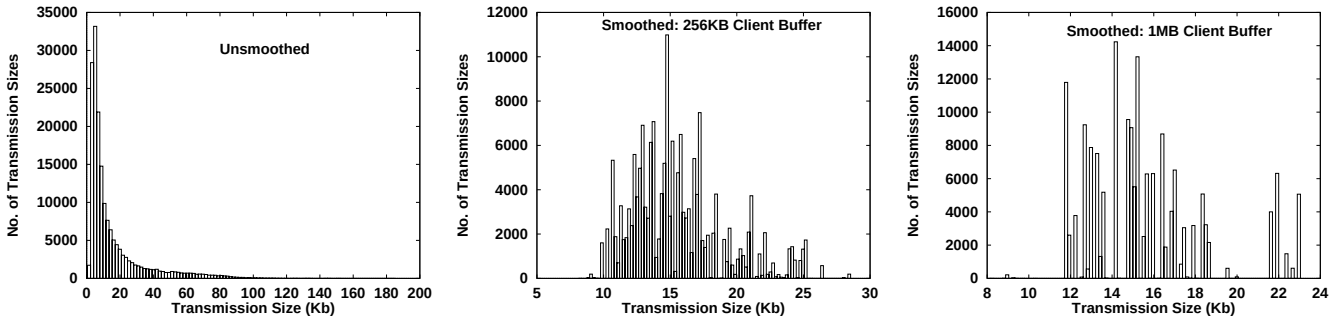
(a) Unsmoothed (b) Smoothed: 256KB (c) Smoothed: 1MB

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

in a CBR or piece-wise CBR manner is studied in [18, 17]. The authors in [6] examine the issue of minimizing the number of *rate changes* in a server transmission schedule. In [21, 22], video smoothing using client decoder buffer together with a startup delay is studied in an on-line video conferencing setting, and the shortest Euclidean distance algorithm of [13] is used to produce smoothed server transmission schedules under the assumption that the frame sizes of the video conference trace are known *a priori*. In [27], a smoothing algorithm is presented that achieves the maximal reduction in rate variability for stored video, producing the "smoothest" possible server transmission schedule. The intuitive notion of "smoothness" is formalized using the concept of *majorization* [16], 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 both minimal peak rate and variance for a given client buffer size. Because it optimally reduces rate variability, we use this algorithm as the video smoothing technique throughout the paper.

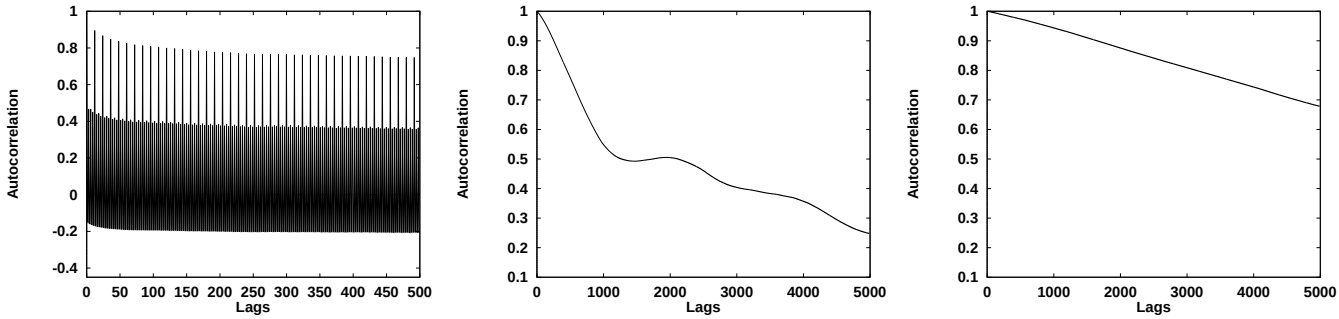
Figure 1 visually demonstrates the effect of video smoothing by plotting the transmission sizes over a two-hour MPEG-1 encoding of *Star Wars* [7], where both the unsmoothed transmission schedule (a) as well as the smoothed transmission schedules for client buffer sizes of 256 KB (b) and 1 MB (c) are shown. The transmission size is defined as the number of bits sent by the server per frame unit of time (approximately 42 ms, given the 24 frames/s frame rate for the *Star Wars* encoding). In the rest of the paper, we will refer to the smoothed transmission schedule of a video trace as the *smoothed trace*. It is a sequence of transmission sizes produced by the optimal smoothing algorithm of [27]. Note that implicit in our study of video smoothing techniques is the assumption that the server transmits ATM cells within a frame (or a transmission size) periodically using the intra-frame *deterministic smoothing* method [28]. Under this assumption, a frame unit of time is the most natural choice of time reference. In the latter part of the paper we will see that the marginal distribution of video transmission rates is the only information required by our call admission control scheme, any smaller time unit, say, half a frame unit, will not change the description of the marginal distribution (only the scale is changed), while any larger time unit may alter this description, resulting in a "coarser" description.

Figure 2 shows the corresponding histograms of the unsmoothed and smoothed video traces of Figure 1, plotted with 100 bins (note the different scales of the axes in Figure 2). These figures indicate that smoothing significantly



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

Figure 2: Impact of the Optimal Smoothing on the Marginal Distributions of *Star Wars*



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

Figure 3: Impact of the Optimal Smoothing on Autocorrelation Structures of *Star Wars*

reduces the range of transmission sizes – from 0-200 Kb per frame unit of time in the unsmoothed schedule, to 5-30 Kb per frame unit of time with a 256 KB client buffer, and to 6-24 Kb per frame unit of time in the case of 1 MB client buffer. Note that the 1 MB client buffer smoothed trace (Figure 1(c)) contains a relatively small number of long, constant rate segments. Furthermore, note that the histogram of a smoothed trace differs significantly in appearance from that of the unsmoothed trace. In particular, the tail distribution of these histograms have very different forms: the long, heavy "tail" of the unsmoothed *Star Wars* trace (Figure 2(a)) is transformed into disconnected, conspicuously outstanding "spikes" after smoothing into a 1 MB client buffer (Figure 2(c)).

These drastically altered marginal distributions of smoothed video streams have important consequences for traffic modeling. For example, the traffic modeling techniques presented in [7, 12, 24] that characterize the "heavy-tailed" marginal distributions are not applicable to the smoothed video traces. Neither is the DAR(1) traffic model in [5] which assumes that the marginal distribution can be approximated by a negative geometrical distribution. Clearly, 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 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 slowly 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 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 [10, 14] 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 the 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: the stringent delay requirement of real-time video transport means that the network buffer allocated for the video streams must be relatively small. Therefore when the streams are highly correlated, insufficient bandwidth at one point in time is likely to lead to consecutive losses over a relatively long period of time, thus greatly affecting the client's QoS. These observations have been confirmed by our experiments with smoothed video streams. Consequently, in supporting the real-time transport of 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 (e.g., to the amount needed in a network switch for temporarily storing data to be forwarded), since buffering is only effective in reducing losses due to fast-time scale rate variability, of which there is little for smoothed video streams. In general, the optimal buffer/bandwidth trade-off depends on the characteristics of source traffic and is an interesting subject worth further study (*see* [15] for results along this line in the context of leaky-bucket regulated sources).

Two advantages are realized with minimal buffer allocation in the network. First, queueing delay jitter within the network is greatly reduced, implying that less client buffer space is needed to accommodate it. From the client's perspective, this also means reduced latency in playback. Second, minimal buffering in the network limits the effect of the autocorrelation structure of the user's traffic on the overall average loss rate [26]. Hence, the difficult task of characterizing the correlation structure of the user traffic is much less important. For these reasons, we will assume that the network employs very little buffering internally for real-time video transport, and in fact, we model a network switch as a bufferless multiplexer in the remainder of the paper. Under such a model, only marginal distribution information (e.g., Figure 2) is needed in traffic specification.

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

Name of Video	Beauty & Beast	CNN News	Jurassic Park	MTV News	Princess Bride	Silence of the Lambs	Soccer	Star Wars	Termi-nator	Wizard of Oz
Mean Rate	40.0	40.0	13.1	24.6	40.0	7.3	27.1	15.6	10.9	41.2
Peak Rate	251.7	246.6	119.6	229.2	243.6	134.2	187.2	185.3	79.6	343.1

Table 1: Statistics of the 10 MPEG-1 Coded Video Traces (in Kb/Frame)

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 gains of smoothed video streams under this independent source assumption, and then investigate the effect of correlated video streams. Finally, we discuss the implication of this statistical multiplexing gain on network service models and QoS guarantees.

3.1 Independent Video Streams

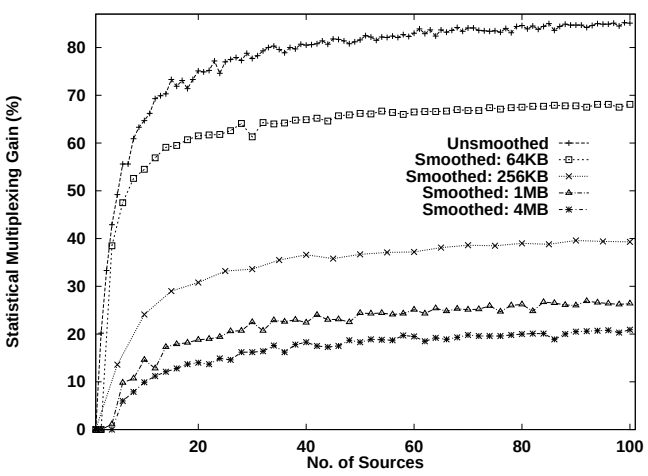
To investigate the statistical multiplexing gain, we use a simple simulation model. We consider a bufferless multiplexer with n independent video streams. The QoS requirement in our study is the loss rate encountered by the video streams at the multiplexer, which is calculated as the ratio of the total amount of loss over the total amount of video transmitted. For a given QoS requirement (say a loss rate of 10^{-6}), we perform 500 independent simulation runs to empirically obtain the minimum bandwidth needed to satisfy the given QoS 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 level QoS[2].

In simulating independent video streams, 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.

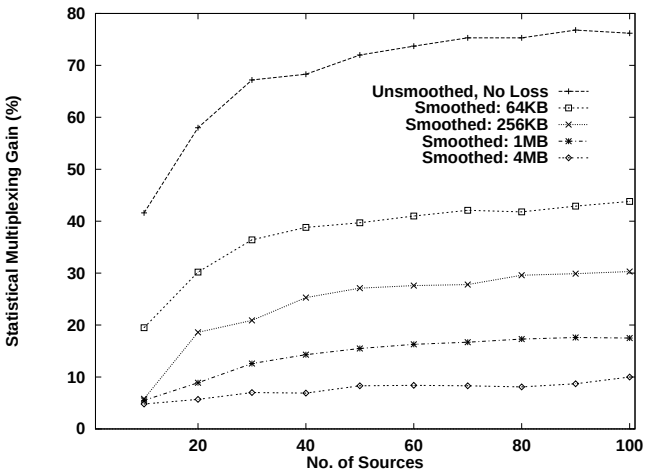
To quantify the statistical multiplexing gain, we use the formula $(1 - r^*/\hat{r}) * 100$ as its formal definition, where r^* is the aggregate bandwidth required to satisfy a given QoS requirement (say, no loss) for all video streams in the simulation and $\hat{r}$ is the peak rate of the aggregate 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 the 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. Although we use *Star Wars* in this (and all other) homogeneous-source experiments, the results hold qualitatively for all of the video traces in our test set. In case (b),

[2]Another set of independent runs are performed to test the robustness of the aggregate bandwidth value. For stringent loss rates such as 10^{-5} or 10^{-6} (the latter loss rate essentially yields a lossless transmission). The maximum bandwidth obtained from the first set of 500 runs is almost always sufficient to satisfy the given QoS in the second set of 500 runs.



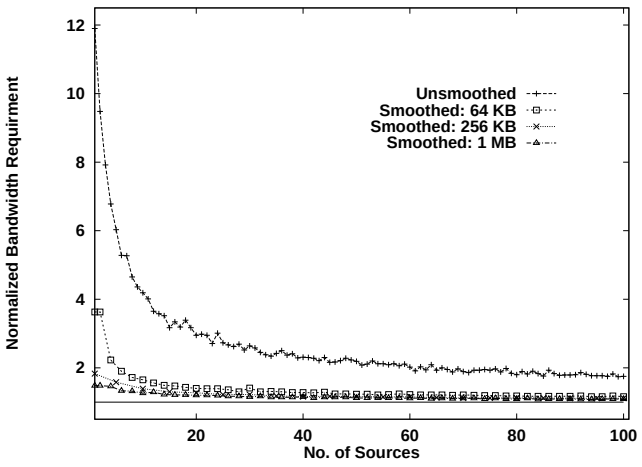
(a) *Star Wars*



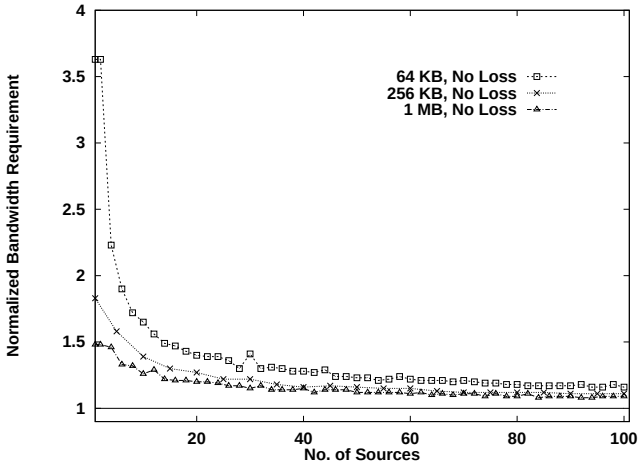
(b) 10 *different* video traces

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

sources are generated from 10 different video traces (their peak/mean rates are listed in Table 1). The number of sources from each type of video are increased uniformly as the number of sources increases. 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, while for smoothed streams with various client buffer sizes, a potential statistical multiplexing gain of 10%-60% is realizable. Thus, *there are still significant statistical multiplexing gains to be exploited by VBR service when individual streams are smoothed, especially when client buffers are relatively small.*



(a) Unsmoothed and Smoothed *Star Wars* Streams



(b) Smoothed *Star Wars* Streams Only

Figure 5: Effect of Statistical Multiplexing on Per-Stream Bandwidth Requirement

Figure 5(a) shows the effect of statistical multiplexing on the per-stream bandwidth requirement (normalized by the mean rate) to achieve lossless transport when all video streams are homogeneous *Star Wars* traces, either unsmoothed or smoothed. To emphasize the potential statistical multiplexing gains after smoothing, the same curves for the smoothed video streams in Figure 5(a) are reproduced in Figure 5(b) alone (with a different y-axis scale). Note that since the mean rate for both smoothed video streams and unsmoothed video streams are the same, the normalized bandwidth required when there is a single source shows the impact of video smoothing on bandwidth reduction. This illustrates that video smoothing can achieve significant network utilization improvement under CBR service. However, the network utilization can be further improved if VBR service is used, as these figures demonstrate that statistical multiplexing gains can significantly reduce the bandwidth required to support a given QoS level. For example, consider an OC-3 link which has a bandwidth of approximately 155 Mb/s. Suppose we have a client buffer of size 256 KB. Given the software MPEG-1 coded *Star Wars* video trace (which has an average rate of roughly 0.73 Mb/s), about 185 smoothed *Star Wars* streams can be supported with no loss under CBR service using peak rate allocation. Under VBR service, our simulation results show that *an additional* 119 *Star Wars* streams can be supported without experiencing any loss. This yields a *potential* 67% utilization gain. However, this potential utilization gain is by no means guaranteed due to the nature of statistical multiplexing. Traffic arrival patterns play a critical role in determining the realizable statistical multiplexing gain.

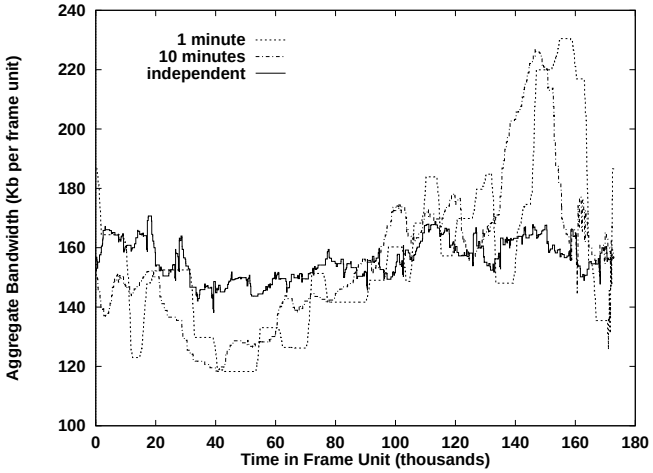
3.2 Correlated Video Streams

The assumption that video start times are independent of each other may sometimes be violated in practice. For example, in a video-on-demand system, many users may start watching videos within a short time span, thus producing correlated video streams. We next investigate the impact of correlated video streams on the statistical multiplexing gain.

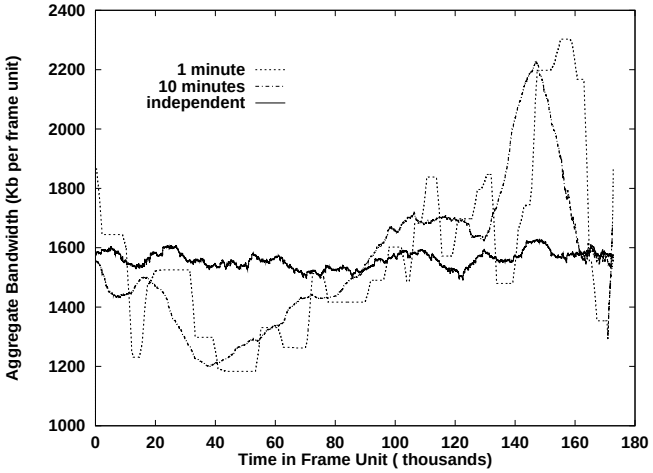
To investigate this question, we consider scenarios in which all video streams are constrained to begin within a short interval of time, say of length Δ minutes. Within this time interval, start times are uniformly, independently and identically distributed. In our simulation, this corresponds to randomly choosing the start of a video stream from the first Δ minutes of the video trace.

Figure 6 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 where each individual source arrives at the multiplexer *independently*, 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 video streams is remarkably burstier and has a considerably larger peak rate than under independent video streams.

Figure 7 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 observable difference in the behavior of the

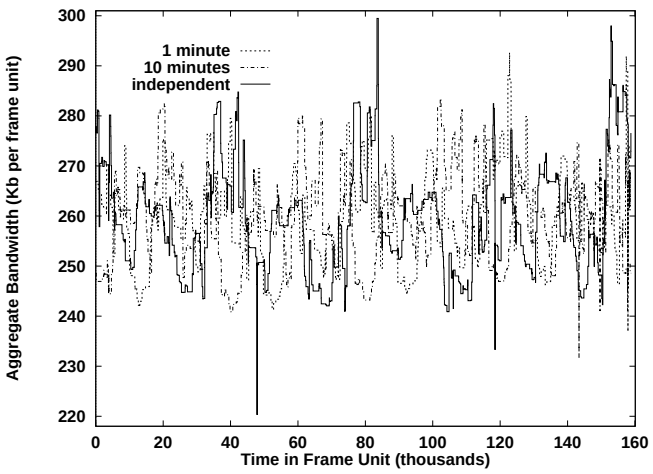


(a) 10 sources of *Star Wars*

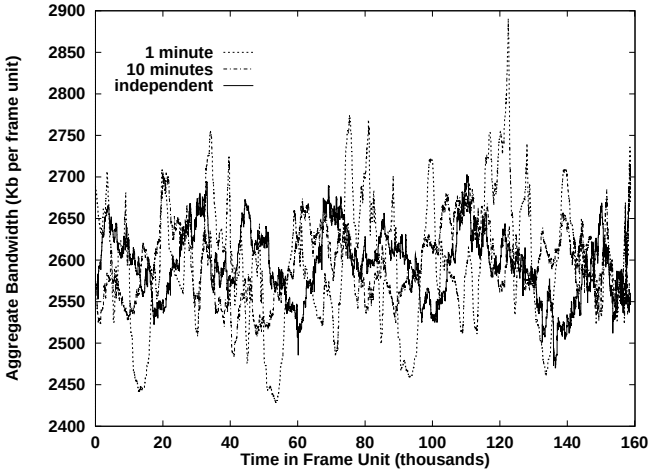


(b) 100 sources of *Star Wars*

Figure 6: Aggregate Smoothed (1MB) Homogeneous Video Streams under Various Arrival Patterns

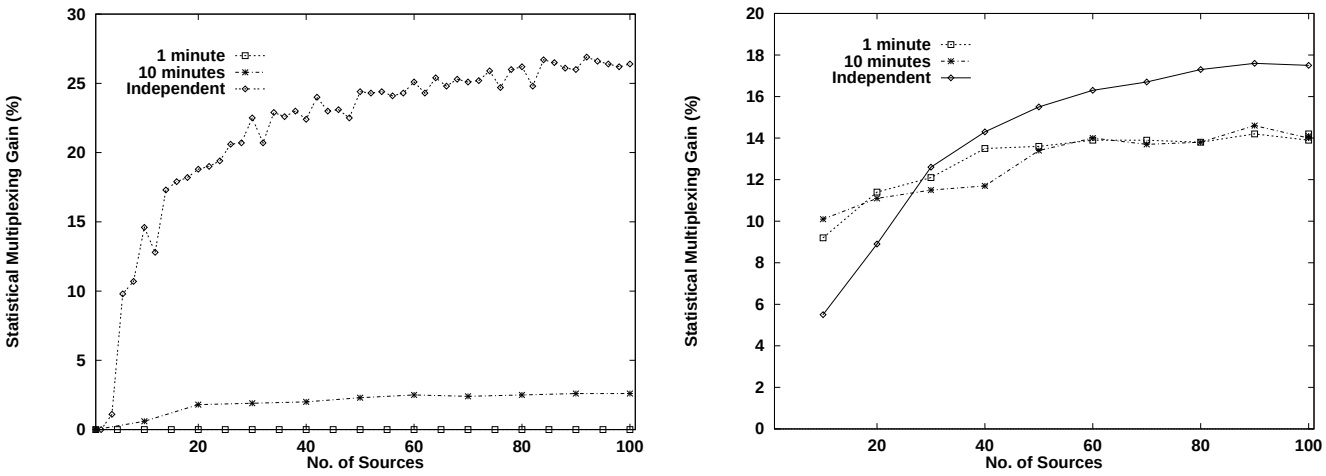


(a) 10 sources: 10 different videos, 1 instance each



(b) 100 sources: 10 different videos, 10 instances each

Figure 7: Aggregate Smoothed (1MB) Heterogeneous Video Streams under Various Arrival Patterns



(a) *Star Wars* (b) 10 *different* videos

Figure 8: Statistical Multiplexing gain under Correlated Video Streams: Smoothed Video Streams, No Loss

correlated and independent video streams. 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 stream case is considerably larger than that in the independent stream 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 video streams on statistical multiplexing gain is shown in Figure 8 where video streams are smoothed into a 1 MB client buffer. Clearly, correlated video streams have an enormous impact on aggregation of homogeneous sources, leaving almost no statistical multiplexing gains to be exploited. On the other hand, there is much less severe impact when heterogeneous streams are aggregated. In this case, the heterogeneity of the video streams helps alleviate the adverse impact of correlation on the statistical multiplexing gain.

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

We have seen that VBR service can significantly improve network utilization by exploiting the potential statistical multiplexing gains available with inherently bursty network traffic. However, we have also seen that the potential statistical multiplexing gain can be diminished by correlated video streams. 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, the other 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 in contrast 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 service* permits the network to fail to provide the promised call-level QoS guarantee (referred to as *service failure* in the rest of the paper), for example, in the *rare* event that the users produce correlated traffic. *Robustness* of a network service is then represented by the likelihood that a *promised* call-level QoS guarantee would fail, i.e., the probability of service failure. This is in contrast to *guaranteed service*, where as long as the user complies with its traffic specification, the network promises to deliver the QoS it has guaranteed to 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 target call-level QoS guarantee. In other words, the amount of bandwidth allocated to a given video stream would differ for services with the same target call-level QoS guarantee but with varying robustness, depending on how much potential statistical multiplexing gains are to be realized. Since the extent of statistical multiplexing gains that can be realized depends on the user behaviors, which are almost impossible to predict and characterize, the robustness of a network service is difficult to quantify mathematically. Despite this difficulty, the robustness of a service may still be empirically verified or tested by the network service provider. Now the fundamental question is: How can we design an effective call admission control scheme that provides a flexible mechanism to support a range of network services with varying robustness? In the next section, we take a systematic approach to address this problem. In particular, we propose a *uniform* call admission control scheme that has the flexibility of providing multiple levels of QoS services with varying robustness.

As an aside, we point out that in addition to providing multiple levels of QoS services with varying robustness through call admission control, other provisions may be made by either the network or by users to ensure the promised call-level QoS guarantees can be successfully met. For example, in a video-on-demand system, batching [4] 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 video streams.

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 first describe a Chernoff-bound-based call admission control algorithm and then study methods for characterizing the sources' marginal distribution. In particular, we present a simple, novel three-state traffic model with only five parameters that can be easily obtained from the stored video. Using this simple traffic model, we devise a uniform call admission control scheme based on the Chernoff bound method, and show that it provides an 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 $t = 0, 1, \ldots$, 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 random variable a_i which takes the values $r_1^{(i)} \leq r_2^{(i)} \leq \ldots \leq r_{K_i}^{(i)}$. In particular, $P\{a_{ij} = r_k^{(i)}\} = p_k^{(i)}$. In other words, with probability $p_k^{(i)}$, a_{ij} is in state k, and while in this state, the source 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, 5] 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 $\Lambda^*(\mu) = \sup_{\theta \geq 0}\{\theta\mu - \Lambda(\theta)\}$, $\Lambda(\theta) = \sum_{i=1}^{I} J_i \log M_i(\theta)$ and $M_i(\theta) = \sum_{k=1}^{K_i} p_k^{(i)} e^{\theta r_k^{(i)}}$ is the moment generating function of a_{ij}.

As $c \to \infty$ with $J_i/c = O(1)$, $1 \leq i \leq I$, the Chernoff Bound (1) can be further refined [20, 2, 1, 5, 8] by adding a prefactor:

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

where θ^* is the solution to $\Lambda'(\theta) = c$. Here $\Lambda'(\theta)$ 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, i.e., $Pr\{a \geq c^*\} \leq \lambda$. From (2), we have that the estimated bandwidth c^* is given by the following expression:

$$c^* = \sum_{i=1}^{I} J_i \frac{M_i'(\theta^*)}{M_i(\theta^*)} \tag{3}$$

where θ^* is the solution to the following equation:

$$\log \lambda = \Lambda(\theta) - \theta\Lambda'(\theta) - \log \theta - \frac{1}{2} \log \Lambda''(\theta) - \frac{1}{2} \log(2\pi). \tag{4}$$

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 (3) with J_l replaced by $J_l + 1$, is less than c, the channel capacity of the multiplexer.

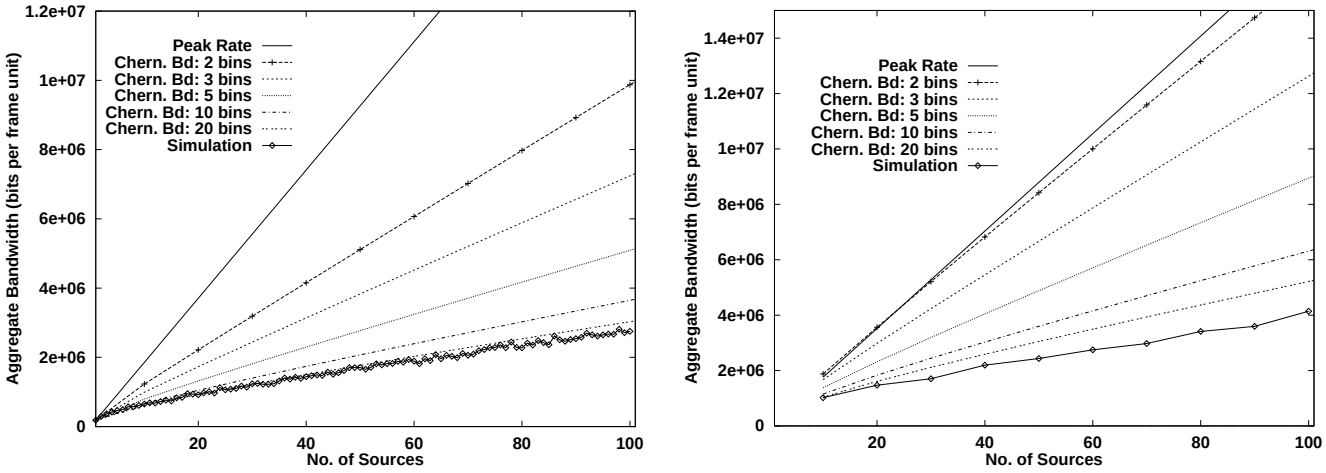
The cost of the call admission algorithm lies mainly in the computation of the marginal moment generating function $M_i(\theta)$ for each source and the solution to the nonlinear equation (4). In our experience, the latter can be solved very fast using the standard Newton-Bisection method. The major cost is associated with the computation of $M_i(\theta)$ and its first and second derivatives used in (3) and (4). The marginal moment generating function is computed from source marginal distribution information $\{(p_k^{(i)}, r_k^{(i)}), 1 \leq k \leq K_i\}$, $1 \leq i \leq I$, provided by the user and maintained by the network. Clearly, the computational cost is a function of the number of states used to describe the source marginal distribution.

In applying the above call admission algorithm to stored video, one important issue is the manner in which the source's marginal distribution is specified. At first glance, this may not seem to be a difficult issue. For stored video, it appears that the server can easily obtain the marginal distribution information from a video trace and simply supply it to the network. However, from the perspective of the network, this approach may not be feasible in practice, since maintaining distribution information for hundreds or thousands of traffic streams can be formidable. Therefore, a key issue is how to characterize 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 amount of information and the processing costs associated with this information. This question is particularly challenging, as we have shown that video smoothing drastically alters the marginal distribution of video traces.

The focus of the remainder of the paper is thus on the marginal distribution characterization. In Section 4.2, we look at a standard method for characterizing the marginal distribution — the histogram method [29]. Under this method, by increasing the number of bins used to describe the histogram, a more accurate bandwidth estimation can be obtained using the Chernoff-bound method. However, this better performance is achieved by incurring greater network overhead. Furthermore, to provide multiple QoS guarantee services with varying robustness, a variable number of user specifiable parameters needs to be supported in the call admission control and traffic specification schemes, adding more complexity to the network control and management mechanism. To overcome these problems, in Section 4.3, a different approach for characterizing the marginal distribution is proposed. Instead of requiring the user to specify the marginal distribution directly, as in the case of the histogram method, the network only requests the user to specify a few generic parameters of the distribution (such as the peak and mean rates) and constructs a distribution that matches these parameters. We present a simple, novel five-parameter traffic model and show that by appropriately choosing parameter values, the five-parameter traffic model demonstrates comparable (if not superior) performance to the histogram method. The five-parameter model presents a uniform traffic specification for stored video to the call admission control scheme without incurring extra overhead and complexity, thereby providing a flexible mechanism to support multiple levels of QoS services with varying robustness.

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.



(a) *Star Wars* (b) 10 *different* videos

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

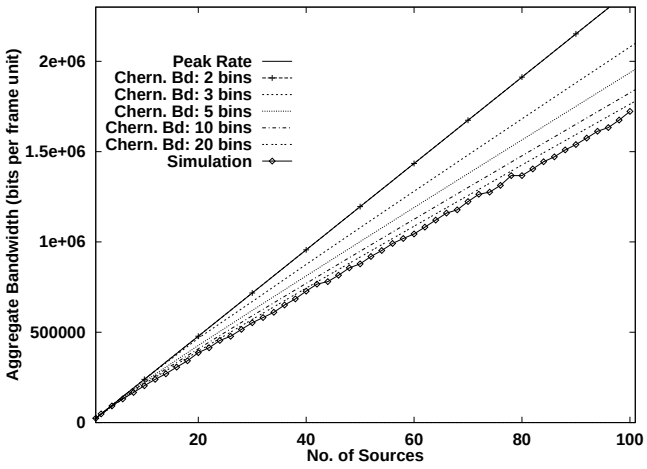
The marginal distribution of a video trace can be characterized using a K-bin histogram as follows. Let $\hat{r}$ be the peak rate of the given trace. We divide the range $(0, \hat{r}]$ into K equal intervals of width $w = \frac{\hat{r}}{K}$ (i.e., bins for histogram). The empirical marginal distribution is then collected by counting the 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 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, N is the length of the video, and $r_k = k * w$ is the amount of traffic generated in this state[3].

We evaluate the performance of the Chernoff-bound-based call admission control algorithm using histogram characterizations of source marginal distributions as follows. For a given loss rate, we compare the bandwidth estimated by the Chernoff bound method using equation (3) with that obtained from simulation. The simulation set-up is the same as in Section 3.

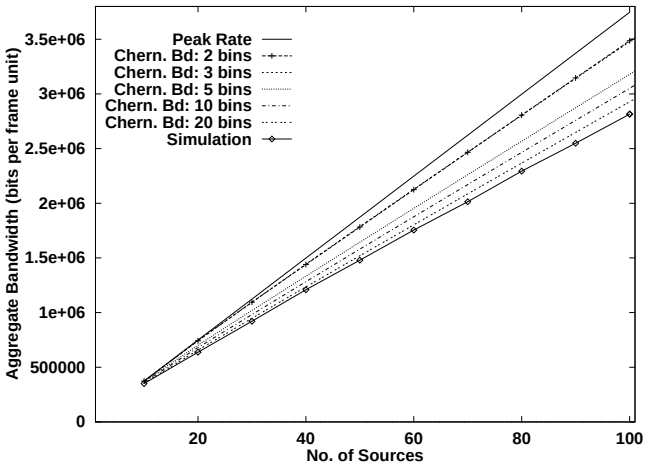
The results are shown in Figure 9 for the unsmoothed video streams, and in Figure 10 for the smoothed video streams (with 512 KB client buffers). In both figures, sources in case (a) are homogeneous (generated from the *Star Wars* trace), whereas sources in case (b) are generated from 10 different video traces with an equal number of sources of each type. In all cases, we see that as the number of bins used to describe 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.

In Figure 11, 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

[3]Choosing r_k this way results in a histogram that generally has a 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.



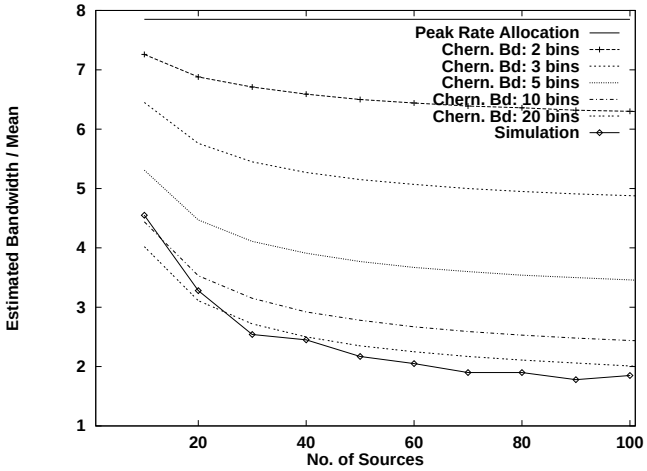
(a) *Star Wars*



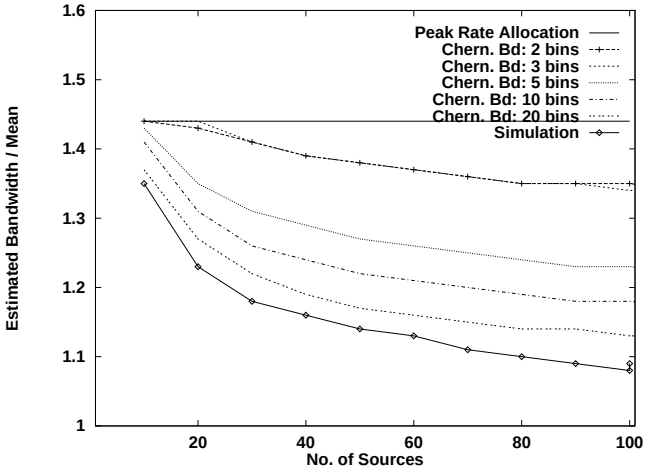
(b) 10 *different* videos

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

allocation and the simulation result. The figure indicates how much statistical multiplexing gain can be realized when the Chernoff-bound based call admission control scheme is employed in combination with the histogram method with various number of bins.



(a) 10 *different* videos: Unsmoothed



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

Figure 11: Comparison of Chernoff Bound Estimation for the Unsmoothed and Smoothed Streams: Loss Rate 10^{-6}

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 degrees of robustness of QoS guarantees. For example, by choosing $K = 3$, the network makes a rather conservative assumption about user behavior, in terms of its allocation of bandwidth to provide the requested service to the users. By choosing $K = 5$, or $K = 10$, or larger, the network makes increasingly optimistic and

aggressive assumptions about user behavior (Figures 9 and 10). 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, counter-balancing the benefits resulting from higher network utilization. To support multiple levels of QoS services with varying robustness, the network has to support a call admission control mechanism with a traffic specification scheme that requires a different number of bins to be provided by the user for each service level, therefore adding more complexity in the network resource management.

4.3 Parsimonious Bounding Models for Marginal Distribution Characterization

In this section we take a very different approach to the problem of characterizing the source marginal distribution for the purpose of call admission control. The key idea behind the approach is to have the network construct an approximation to the user's marginal distribution, from a very small number of parameters provided by the user. Thus we consider the following problem: given a user traffic specification described by a set of parameters such as the mean and peak rates of a source, how should the network construct a marginal distribution that matches the given user parameters? Clearly there are many possible 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 for stored video, given our results in Section 2. 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? In answering this question, we take a bounding approach 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 with this distribution yields an *upper* bound on the bandwidth estimate that would result from using *any* marginal distribution matching the given set of user-specified parameters.

To address this 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 [25]) 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 [25]. 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 $\|\cdot\|_\infty$ is the essential supremum of a random variable, defined as $\|X\|_\infty = \inf\{x : Pr\{X > x\} = 0\}$[4].

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*.

[4] 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.

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[5] (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 the definition of $\Lambda^*(c)$, this is equivalent to

$$\sup_{\theta \geq 0}\{\theta c - \hat{\Lambda}(\theta)\} \leq \sup_{\theta \geq 0}\{\theta c - \Lambda(\theta)\}. \tag{5}$$

Clearly, (5) 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 $\theta \geq 0$. ∎

4.3.1 Simple Parsimonious Models

Based on Theorem 1, we now construct two simple bounding models which require 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 is stated formally in Theorem 2 (for a similar result, see [19]).

As we shall see, the two-state model based only on the mean and peak rates of a source results in a rather conservative bandwidth estimate by the Chernoff bound method. In the following, we thus present a simple "three-state", five-parameter model to characterize the marginal distribution of a video: in addition to the two parameters representing the mean m and 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.

$$Pr\{X \geq \tilde{r}\} = \tilde{p} \text{ and } E[X|X \geq \tilde{r}] = \tilde{m}. \tag{6}$$

[5] Since the exponential term in (2) 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.

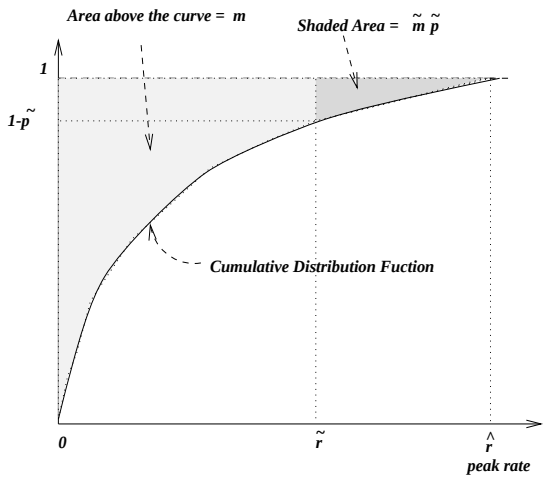


Figure 12: Illustration of the Parameters of the Three-State Model

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 12. 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 $0 < \tilde{p} < 1$,

$$\hat{X} = \begin{cases} 0 & \text{with probability } (1 - \frac{\tilde{m}'}{\tilde{r}-1})\tilde{q}; \\ \tilde{r}-1 & \text{with probability } \frac{\tilde{m}'}{\tilde{r}-1}\tilde{q}; \\ \tilde{r} & \text{with probability } (1 - \frac{\tilde{m}-\tilde{r}}{\hat{r}-\tilde{r}})\tilde{p}; \\ \hat{r} & \text{with probability } \frac{\tilde{m}-\tilde{r}}{\hat{r}-\tilde{r}}\tilde{p} \end{cases} \tag{7}$$

where $\tilde{q} = 1 - \tilde{p} = Pr\{X < \tilde{r}\}$ and $\tilde{m}' = E[X|X < \tilde{r}]$. As $m = E[X] = E[X|X < \tilde{r}]Pr\{X < \tilde{r}\} + E[X|X \geq \tilde{r}]Pr\{X \geq \tilde{r}\} = \tilde{m}'\tilde{q} + \tilde{m}\tilde{p}$, $\tilde{m}' = \frac{m-\tilde{m}\tilde{p}}{\tilde{q}}$. 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[6].

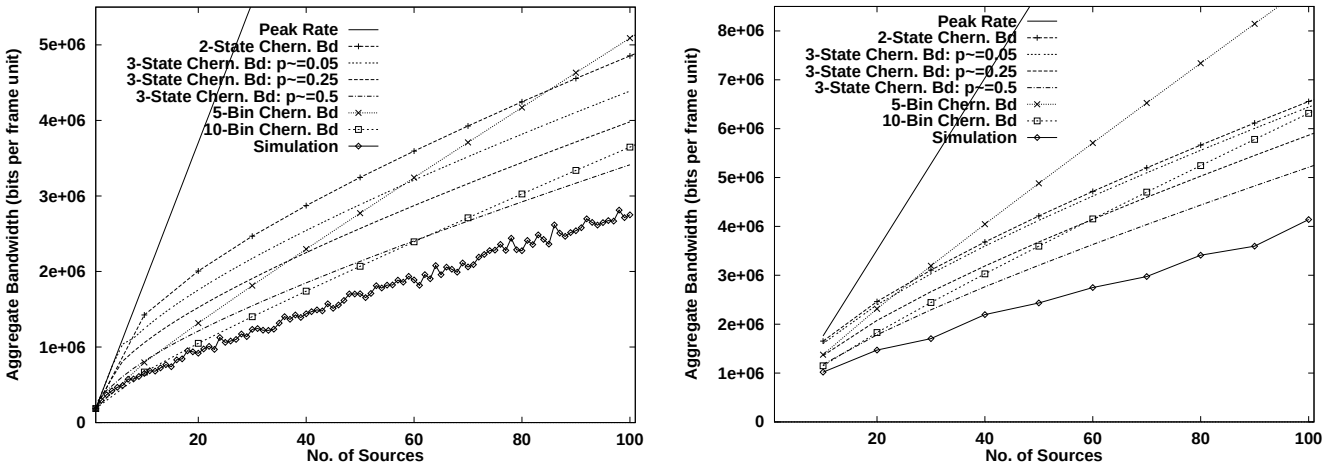
In the cases $\tilde{p} = 0$ or $\tilde{p} = 1$, the three-state model degenerates into the two-state model described earlier.

It is easy to check that $E[\hat{X}] = m$, $\|\hat{X}\|_\infty = \hat{r}$, $Pr\{\hat{X} \geq \tilde{r}\} = \tilde{p}$ and $E[\hat{X}|\hat{X} \geq \tilde{X}] = \tilde{m}$. Theorem 2 states that this 3-state model has the *most stochastically variable* marginal distribution among all discrete random variables X with the matching parameters.

Theorem 2

(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 $X \leq_{icx} \hat{X}$.

[6] 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: Unsmoothed Streams, Loss Rate 10^{-6}

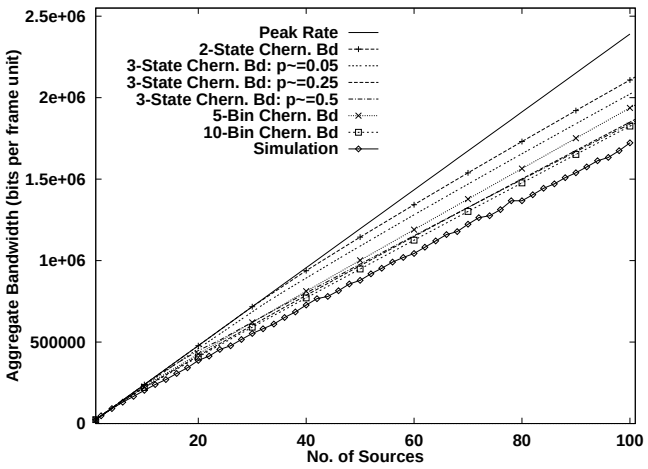
(2) If X is an arbitrary nonnegative discrete random variable such that $E[X] = m$,$\|X\|_\infty = \hat{r}$, $Pr\{X \geq \tilde{r}\} = \tilde{p}$ and $E[X|X \geq \tilde{r}] = \tilde{m}$, and $\hat{X}$ is defined as in (7), then $X \leq_{icx} \hat{X}$.

The proof of the theorem can be found in Appendix A.

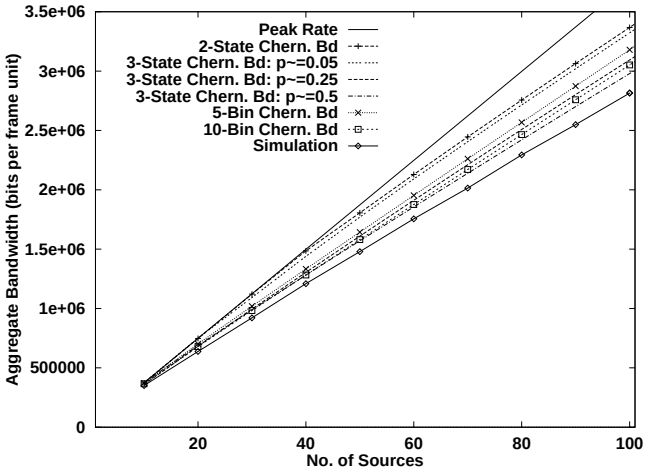
4.3.2 Evaluation

We now examine the performance of the two-state and three-state models as the parameter $\tilde{p}$ is varied. Figure 13 shows the performance for unsmoothed video streams, and Figure 14 for 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 as many parameters as 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 $\tilde{p} = 0.5$) to rather conservative (say, $\tilde{p} = 0.05$). 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.

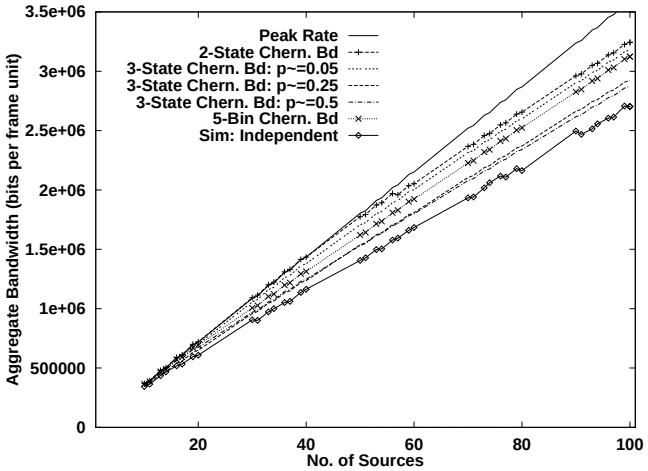


(a) *Star Wars*

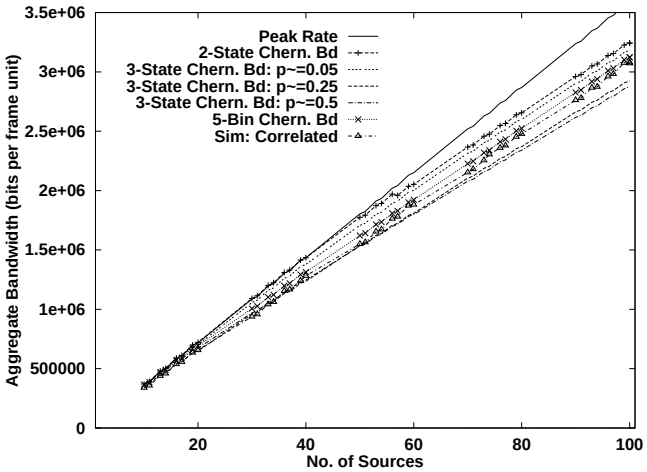


(b) 10 *different* videos

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



(a) Independent Streams



(b) Correlated Streams

Figure 15: Comparison of Marginal Distribution Models: Mixed Smoothed Streams: Loss Rate 10^{-6}

An additional example is shown in Figure 15, where a more diverse mix of video streams is considered. In this example, 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. Furthermore, the number of sources of each video type in this example 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. Figure 15(a) presents the results for the scenario where all video streams arrive at a network node *independently*. In Figure 15(b), we consider a scenario with *correlated* video streams to illustrate the need to provide different service levels to account for possible correlated user behaviors. In this scenario, the *Star Wars* sources all arrive within a period of 10 minutes, and the *Wizard of Oz* sources within a period of 1 minute. We see that the correlated video streams 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 video streams, 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 video streams with the target QoS guarantee, 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 provides a simple and flexible mechanism to balance these two concerns. Using this model, we can fix the call-level QoS guarantee while varying the robustness of the service by setting some parameters of the model, the values of which may be determined by extensive testing to assess the trade-off between the service robustness and the realization of statistical multiplexing gain. For example, for a given loss rate, by choosing $\tilde{p} = 0.05$, $\tilde{p} = 0.25$ and $\tilde{p} = 0.5$, the network can provide three levels of QoS service classes to trade off robustness and realization of statistical multiplexing gains in various degrees. These differential network services can be implemented in combination with a network pricing scheme that reflects the per-connection bandwidth allocation made by the network for these services: the more robust service (e.g., the service with $\tilde{p} = 0.05$) could charge more for setting up a connection than the less robust one (e.g., the service with $\tilde{p} = 0.5$). In any case, the appropriate choice of the parameters used in the three-state model plays a critical role in determining the robustness of the QoS services provided by the network.

The following high-level guidelines summarize the call admission control procedure based on the three-state model.

- **User:**
  1) Chooses the desired loss rate guarantee λ and appropriate level of QoS service (represented by $\tilde{p}$);
  2) Provides the additional four traffic parameters (m, $\hat{r}$, $\tilde{m}$ and $\tilde{r}$) describing the marginal distribution of the new video stream.

- **Network:** At each node along the route to be traversed by the video stream, the network
  1) Computes the worst-case distribution matching the user parameters using formula (7);

2) Estimates the bandwidth required to support the target loss rate λ (using equations (3) and (4));

3) Rejects the user connection set-up request if the available bandwidth at the node is *not* sufficient to support the new video stream, else proceeds to the next node on the route if there is any, accepting the new video stream if the current node is the last one.

In the above description, we have used the marginal distribution of a video stream at the network edge to approximate the distribution at a node within the network. Because of the discrete-time bufferless model used in our approach, we expect this approximation to be conservative, as losses occurring at the upstream nodes would in effect "re-shape" the marginal distribution so that losses are reduced at the downstream nodes. However, a study of the end-to-end behavior of the system is needed to verify this conjecture.

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 [9, 5, 8, 31]. In [5], 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 parsimonious DAR(1) model is employed to specify the source traffic. However, the parsimonious DAR(1) model relies on the fact that the marginal distribution of the video conferencing traces can be approximated by a negative geometrical distribution. Our experience shows that DAR(1) is not appropriate for both smoothed and unsmoothed MPEG compressed video streams because of the long-range dependence exhibited by the traces. A histogram-based call admission control scheme is proposed in [29], 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 [22], 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.

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 [8], and to VBR service with deterministic QoS guarantees [30]. In [8], 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.

Several methods have been used in characterizing the "heavy-tailed" marginal distribution of unsmoothed video traces (*see*, e.g., [7, 12, 24]), where a known distribution, such as Gamma, Pareto, or Lognormal, is used to approximate the marginal distribution in order to obtain a parsimonious characterization. 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 gains and its implication in network resource management and call admission control. We started by investigating the issue of statistical multiplexing gains when streams are smoothed and showed how statistical multiplexing gains 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 call admission control scheme based on the Chernoff-bound method that uses a simple five-parameter model for traffic specification. The scheme provides an effective and flexible mechanism to support different levels of QoS services with statistical QoS guarantees. We evaluated the efficacy of the scheme over a set of MPEG-1 coded video traces.

In summary, our work supports the contention that by explicitly exploiting statistical 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. Although our results are established solely through evaluation using a set of MPEG-1 coded video traces, we believe that they will also hold qualitatively for other VBR coded video streams.

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 such as the impact of VCR functionality on network service and network resource control and management. 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. Careful evaluation of the computational cost of our call admission control scheme in a "real" environment, along with those of other methods proposed in the literature, is also an important research subject.

Acknowledgments

We would like to thank the researchers who generously shared their MPEG video traces. In particular, the contributions of Mark Garrett [7], Ed Knightly [11], Marwan Krunz [12] and Oliver Rose [24] are gratefully acknowledged. We would also like to thank Jayanta Dey and Francesco Lo Presti for many insightful discussions. We are particularly grateful to the anonymous reviewers for many helpful comments that have greatly improved the presentation of this paper.

A Appendix

In this appendix, we prove Theorem 2. 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 3 *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 $a \geq 0$,*

$$\int_a^\infty \bar{F}(x)dx \leq \int_a^\infty \bar{G}(x)dx. \tag{8}$$

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

For a proof, see Proposition 8.5.1 of [25].

Lemma 4 *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:*

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

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 3, 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 2:
(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 3, it suffices to show that for any $a \geq 0$, (8) holds.

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 $x \geq a$, $\bar{F}(x) \leq \bar{G}(x)$. Therefore,

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

For any $0 \leq a < \alpha$, if $0 \leq x \leq a$, then $F(x) < 1 - p = G(x)$. Thus $\int_0^a F(x)dx \leq \int_0^a G(x)dx$. 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 4, we have that $X \leq_{icx} \hat{X}$. ∎

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] 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)*, 15–24, San Francisco, CA, October 1994.

[5] 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 Journal of Selected Areas in Communications*, 13(6):1004–1016, August 1995.

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

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

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

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

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

[11] E. W. Knightly, D. E. Wrege, J. Liebeherr, and H. Zhang. Fundamental limits and tradoffs of providing deterministic guarantees to VBR video traffic. In *Proc. ACM SIGMETRICS*, 98–107, Ottawa, Canada, May 1995.

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

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

[14] 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.

[15] F. Lo Presti, Z.-L. Zhang, J. Kurose, and D. Towsley. Source time scale and optimal buffer/bandwidth trade-off for regulated traffic in an ATM node. Technical Report UM-CS-96-38, Computer Science Department, University of Massachusetts at Amherst, June 1996. A revised version to appear in *Proc. of IEEE INFOCOM*, Kobe, Japan, March, 1997.

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

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

[18] 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.

[19] D. Mitra and J. A. Morrison. Multiple time scale regulation and worst case processes for ATM network control. In *Proceedings of the 34th Conference on Decision and Control*, New Orleans, December 1995.

[20] 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.

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

[22] 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.

[23] J. W. Roberts. *COST 224 Final Report, Performance evaluation and design of multiservice networks*. Commission of the European Commmunities, Luxembourg, 1992.

[24] 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.

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

[26] B. H. Ryu and A. Elwalid. The importance of long-range dependence of VBR video traffic in ATM traffic engineering: Myths and realities. In *Proc. ACM SIGCOMM*, 3–14, Stanford, CA, August 1996.

[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)*, 222–231, Philadelphia, PA, May 1996.

[28] N. Shroff and M. Schwartz. Video modeling within networks using deterministic smoothing at the source. In *Proc. IEEE INFOCOM*, 342–349, 1994.

[29] 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.

[30] 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*, 275–286, Durham, NH, April 1995.

[31] Z.-L. Zhang, D. Towsley, and J. Kurose. Statistical analysis of the generalized processor sharing scheduling discipline. *IEEE Journal of Selected Areas in Communications*, 13(6):1071–1080, August 1995.