In Proceedings of the 1995 International Symposium on
Multimedia Communications and Video Coding

# EFFECTS OF SMOOTHING ON END-TO-END PERFORMANCE GUARANTEES FOR VBR VIDEO

Edward W. Knightly[†] and Paola Rossaro[‡]

The Tenet Group
[†]EECS Department, U.C. Berkeley
[‡]International Computer Science Institute, Berkeley, CA

**Abstract**

Integrated services networks have the framework needed to provide performance guarantees to multimedia traffic sources such as Variable Bit Rate (VBR) video. This paper investigates the effects of smoothing VBR traffic sources on their end-to-end delay bounds and on the achievable utilization inside the network. We provide a set of rules that determine if smoothing results in a net advantage to the source and use experiments with traces of MPEG-compressed video to quantify the effectiveness of smoothing in practical networking situations.

## INTRODUCTION

Many distributed multimedia applications are delay- and loss-sensitive and have stringent requirements in terms of network performance. A resource reservation scheme together with priority service disciplines inside the network, as in (Ferrari and Verma, 1990), provide the means for giving network clients end-to-end Quality of Service (QoS) guarantees on delay, loss, and throughput. The guaranteed service can be *deterministic*, in which all packets of a connection are guaranteed to meet the promised QoS, or *statistical* in which probabilistic performance bounds are guaranteed. The two services can be viewed as trading off QoS for network utilization in that deterministic service provides the best QoS with its no-loss, no-delay-violation guarantee, but it cannot employ statistical multiplexing to achieve higher network utilization since that would result in a reduced (non-deterministic) QoS.

An important issue in providing end-to-end performance guarantees is the effect of *smoothing* traffic sources on QoS and network utilization. In this paper, we use analysis and traces of MPEG compressed video to explore the effects of smoothing VBR video sources on end-to-end deterministic performance guarantees. We consider the case of smoothing by a FIFO (a first-in first-out queue) so that before the packets of a VBR video source are transmitted to the network, they are sent through a FIFO at the source. This source-FIFO services packets at a smoothing rate $R_S$ where $R_S$ is less than the unsmoothed source's peak rate and greater than its long term average rate. The smoothing rate can be viewed within the context of the Deterministic Bounding Interval Dependent (D-BIND) traffic model (Knightly and Zhang, 1995) in which the bounding rate $R_S$ represents an upper bound on the unsmoothed source's rate over every interval of length $I_S$. The D-BIND model provides a more accurate traffic specification via multiple rate-interval pairs which in turn allows a higher utilization than is possible with a peak-rate-allocation scheme.

The traffic shaping implemented by the FIFO results in a traffic stream that is "less bursty" than the original traffic stream, where "less bursty" is defined in terms of the D-BIND model in a manner similar to (Low and Varaiya, 1993). When the smoothed traffic sources are multiplexed at queues inside the network, the resulting bound in queueing delay will be reduced. Or in other words, for a given queueing delay bound, smoother traffic results in a higher achievable network utilization. However, the smoothing-FIFO at the source has introduced additional end-to-end delay by adding a smoothing delay.

Thus, we are considering the effect of smoothing VBR video sources on the total end-to-end delay bound where the end-to-end delay bound considers both smoothing and queueing delays at all hops along the path. Since smoothing decreases the bound on queueing delay, but increases the bound on smoothing delay, or the time that packets spend in the traffic shaper before entering the network, smoothing can be considered a tradeoff between buffering at the source and buffering inside the network.

In this paper, we propose a set of conditions for determining if traffic smoothing results in a net improvement in end-to-end QoS for a source. We then investigate these results in a practical environment by analyzing traces of MPEG compressed video. The results show that over a single congested hop with homogeneous sources, smoothing is an ineffective means for achieving higher network utilization. That is, the savings in queueing delay achieved by transmitting smoother traffic are outweighed by the smoothing delay incurred at the source. Alternatively, if a source traverses multiple congested nodes, smoothing can result in considerable improvements in achievable network utilization for a given QoS guarantee. Intuitively, the reason for this is that the smoothing delay is incurred only once at the source, whereas the queueing delay may be incurred at each hop along the path. Thus, in many cases, a higher network utilization or lower end-to-end delay can be achieved by smoothing. Moreover, the analysis indicates that the advantages of smoothing increase with the number of congested hops traversed by the connection.

This work differs from other approaches to transporting VBR video (e.g., (Kanakia, Mishra, and Reibman, 1993; Zhang and Knightly, 1995)) in that once a connection is established, the network provides constant quality video transmission without being affected by the current network congestion. Thus, we do not adjust a coder's "Q-factor" (a quality factor in many compression algorithms to reduce the bit-rate), nor do we drop packets. Other related works include (de Veciana and Walrand, 1992; Shroff and Schwartz, 1994), in which smoothing is considered in the realm of statistical performance guarantees. Our work considers a *lossless smoothing*, in which we provide an end-to-end deterministic QoS and introduce only delay at the source in order to better shape the traffic for the network.

In the remainder of this paper, we first describe the D-BIND model and the associated admission control test for a FCFS scheduler. Next, we provide a definition of smoothing and introduce a set of rules to determine when a source should smooth. Finally, we use traces of MPEG-compressed video to examine the rules in a realistic scenario and we introduce various performance metrics to further investigate the effectiveness of smoothing and the tradeoffs involved.

## DETERMINISTIC QOS

Compared to statistical service, deterministic service provides better QoS in that it provides a no-loss, no-delay-violation service. While this does preclude statistical multiplexing, as shown below, it does not require a peak-rate-allocation scheme. For the network to deliver such a service, it needs a deterministic upper bound on all sources receiving the service. This approach has the added advantage that a source's traffic specification can be enforced. For example, if a source promises that its minimum packet inter-arrival time is $Xmin$, this may be easily verified and enforced by the network. Alternatively, statistical models of the source are inherently much more difficult to enforce.

### The Deterministic-BIND Model

As shown in (Knightly and Zhang, 1995), previous deterministic traffic models such as the $(\sigma, \rho)$ model (Cruz, 1991) and the $(Xmin, Xave, I, Smax)$ model (Ferrari and Verma, 1990) cannot capture the property that sources exhibit burstiness over a wide variety of interval lengths. The Deterministic Bounding Interval Dependent traffic model was introduced to address this issue. The key components of the D-BIND model are that it is *bounding*, required to provide deterministic QoS guarantees, and *interval-dependent*, needed to capture important burstiness properties of sources which in turn allows for a higher network utilization (see (Knightly and Zhang, 1995)).

Each deterministic traffic model uses parameters to define a traffic constraint function $b(t)$ which constrains or bounds the source over every interval of length $t$. Denoting $A[t_1, t_2]$ a connection's arrivals in the interval $[t_1, t_2]$, the traffic constraint function $b(t)$ requires that $A[s, s+t] \le b(t)$, $\forall s, t > 0$. Note that $b(t)$ is a time-invariant deterministic bound since it constrains the traffic source over every interval of length $t$. For example, the $(\sigma, \rho)$ is defined such that $A[s, s+t] \le \sigma + \rho t$ for all $t$.

The D-BIND model is defined via rate-interval pairs $\{(R_k, I_k) | k = 1, 2, \cdots, P\}$ so that the constraint function is given by a piece-wise linear function

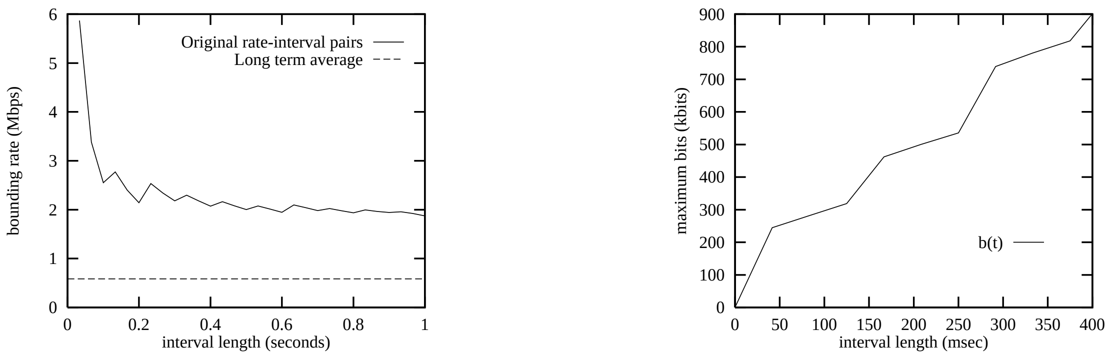
$$b(t) = \frac{R_k I_k - R_{k-1} I_{k-1}}{I_k - I_{k-1}}(t - I_k) + R_k I_k, \quad I_{k-1} \le t \le I_k \tag{1}$$

with $b(0) = 0$. Thus the rates $R_k$ can be viewed as an upper bound on the rate over every interval of length $I_k$ so that

$$A[t, t + I_k]/I_k \le R_k \quad \forall\, t > 0, k = 1, 2, \cdots, P. \tag{2}$$

Figure 1(a) shows a plot of the D-BIND rate-interval pairs for a 30 minute trace of an MPEG-compressed action movie. Plotting the bounding rate $R_k$ vs. interval length $I_k$, the figure shows that the model captures the source's burstiness over multiple interval lengths. For example, for small interval lengths, $R_k$ approaches the source's peak rate (5.87 Mbps in this case) while for longer interval lengths it approaches the long term average rate (the total number of bits in the MPEG sequence divided by the length of the sequence, 583 kbps).

Figure 1(b) shows the movie's D-BIND constraint function $b(t)$ as described by Equation (1) using the rate-interval pairs of Figure 1(a). The figure shows maximum bits (in kbits) that the source transmits over any interval of length $t$ (shown in seconds) and indicates that the D-BIND model is capturing the temporal properties of the MPEG video. For example, the peak rate shown in Figure 1(a) is caused by transmission of the largest I frame of the sequence. This can be seen in the portion of the constraint function with the largest slope (slopes indicating rates) between $t = 0$ and $t = 42$ msec (the frame rate is 24 frames per second). Importantly, even in the worst case, a large I frame is followed by two typically smaller B frames, which is captured by the constraint curve's slope decreasing in the interval $t = 42$ msec to $t = 125$ msec. Next, a P frame is transmitted (even in the worst case) and these tend to be smaller than I frames but larger than B frames. In (Knightly and Zhang, 1995) it was shown that the D-BIND model's ability to capture both micro- and macro-level burstiness of the video sequence, lead to considerably higher utilization than that achieved with previous models.




(a) Rate-Interval Pairs (b) Constraint Function

Figure 1: D-BIND Characterization for Action Movie

## Connection Admission Control

Deterministic admission control conditions rely on the delay analysis techniques of (Cruz, 1991; Zhang and Ferrari, 1994). These works show that if a FCFS scheduler with link speed $l$ is serving $N$ connections each bounded by its respective constraint functions $b_j(t)$, then an upper bound on delay for all connections is given by:

$$d = \frac{1}{l} \max_{t \ge 0} \{ \sum_{j=1}^{N} b_j(t) - lt + \overline{s} \}. \tag{3}$$

where $\overline{s}$ is the transmission time of the largest sized packet. The proof starts with an expression for the work in the system at time $t$, $W(t) = \max_{s \le t}(\sum_j A_j[s, s+t] - l(t-s))$ and uses bounds on individual sources ($b_j(t) \ge A_j[s, s+t]\ \forall s, t > 0$) to bound the aggregate. Delay bounds for priority service disciplines that are more suitable for providing integrated services may be expressed in a similar manner (Knightly et al., 1995).

# SMOOTHING OF VBR TRAFFIC

No-loss smoothing of VBR traffic in the case of the D-BIND model may be viewed from either of two equivalent view-points: the $(R_k, I_k)$ rate-interval pairs or the $b(t)$ constraint function. In the first case, smoothing can be viewed as transforming a source with upper bounds $\{(R_k, I_k) \mid k = 1, 2, \cdots, P\}$ to $\{(\hat{R}_k, \hat{I}_k) \mid k = 1, 2, \cdots, P\}$, with $\hat{R}_k \le R_k$ if $\hat{I}_k = I_k$. This transformation is realized with the smoothing or buffering at the traffic shaper.

A second view of traffic shaping may be seen from the smoothed source's new D-BIND constraint function $\hat{b}(t)$:

$$\hat{b}(t) = \frac{\hat{R}_k \hat{I}_k - \hat{R}_{k-1}\hat{I}_{k-1}}{\hat{I}_k - \hat{I}_{k-1}}(t - \hat{I}_k) + \hat{R}_k \hat{I}_k, \quad \hat{I}_{k-1} \le t \le \hat{I}_k. \tag{4}$$

With $\hat{R}_k \le R_k$ and Equations (1) and (4), we have that $\hat{b}(t) \le b(t)\ \forall t$. More formally, we define smoother traffic in the following manner (similar to the definition in (Low and Varaiya, 1993) for "burstiness curves"):

**Definition 1** *If $lim_{t\to\infty} b(t)/t = lim_{t\to\infty} \hat{b}(t)/t$, then $\hat{A}(t)$ can be considered smoother or less bursty than $A(t)$ if $\hat{b}(t) \le b(t)\ \forall t$.*

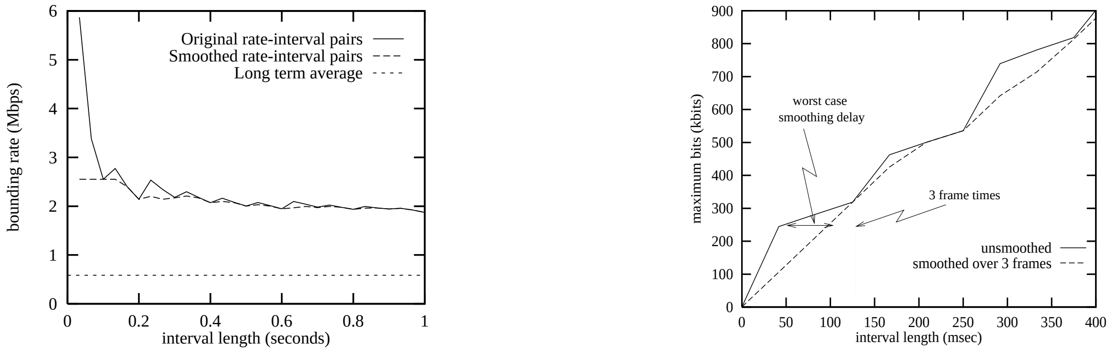
In this paper, we consider smoothing over $S$ video frames to be defined in the following manner: a source smoothed over $S$ frames is shaped by a FIFO queue with rate $R_S$ (a bound on the source's rate over an interval of length $I_S$) and maximum buffer size $B$. $B$ can be chosen with Equation (3) so that no packets are dropped by the traffic shaper. The buffer occupancy of the smoothing FIFO is then given by

$$q(t) = \sup_{0\le s\le t} \{A[s,t] - R_S \cdot (t-s)\}. \tag{5}$$

The smoothed arrival process $\hat{A}$ is therefore

$$\hat{A}[t, t+\delta] = \min(R_S \cdot \delta, A[t, t+\delta] + q(t)). \tag{6}$$

Note that while the smoothed process will be less bursty than the original process (as in Definition 1), it will not in general be constant bit rate (see (Low and Varaiya, 1993) for a proof).




(a) (b)

Figure 2: Smoothed D-BIND Rate-Interval and Constraint Curves

An example of the transformation of smoothing on a source's rate-interval pairs is shown in Figure 2(a). The figure shows the original and smoothed D-BIND rate-interval curve using empirical data from the MPEG trace of an action movie. In this example, the source's original peak rate is $R_1 = 5.87Mbps$. From $I_1 = 1$ frame time, the bounding rate decreases towards the long-term average rate. After smoothing over $S = 3$ frame times (via a FIFO with rate $R_3 = 2.55Mbps$), for a given interval length, the bounding rate of the smoothed source $\hat{R}_j$ is less than or equal to the bounding rate of the unsmoothed source $R_j$. As well, for interval lengths of up to $S$ frame times, $\hat{R}_j$ tends to reach the smoothing rate $R_S$.

Figure 2(b) shows the effect of smoothing on a source's constraint function. As shown, $\tau_S$, the worst case smoothing delay when smoothing over $S$ frames (using a FIFO with rate $R_S$) may be calculated as the maximum horizontal time distance between the two curves $b$ and $\hat{b}$ (Cruz, 1991). That is,

$$\tau_S = \max_{t_2 > t_1}\{t_2 - t_1 | \hat{b}(t_2) = b(t_1)\}. \tag{7}$$

## Tradeoffs

Intuitively, a less bursty source can better utilize network resources. This is stated in the following Lemma. We use $d$ to represent the delay bound for an *unsmoothed* source $A(t)$ with constraint function $b(t)$, and $\hat{d}$ to represent the delay bound for the *smoothed* source with smoothed arrivals $\hat{A}(t)$ and smoothed constraint function $\hat{b}(t)$.

**Lemma 1** *If source $j$ is smoothed so that arrival process $\hat{A}(t)$ is less bursty than $A(t)$, then the queueing delay bound for a FCFS scheduler is reduced. Equivalently, smoother or less bursty sources require fewer network resources.*

Proof: The FCFS queueing delay bound is given by Equation (3) as $d = \frac{1}{l}\max_{t\geq 0}\{\sum_{j=1}^{N} b_j(t) - lt + \overline{s}\}$. If source $j$ is smoothed, then the $j^{th}$ term of the summation $b_j(t)$ is replaced with $\hat{b}_j(t)$. Since $\hat{b}(t) \leq b(t)$ for all $t$, it can only reduce the delay bound to $\hat{d} \leq d$. □

However, while smoothing does reduce the queueing delay bound, it also introduces an additional delay $\tau_S$, due to buffering at the FIFO queue. This delay contributes to the total end-to-end delay perceived by the source. Therefore, from the perspective of delay bound, a source should smooth if the additional smoothing delay bound is less than the reduction in the queueing delay bound, i.e., if the total delay bound, smoothing plus queueing, is decreased. Below, we present a set of "rules" for determining whether smoothing is advantageous to a source by reducing its end-to-end delay bound. Proofs are given in (Knightly and Rossaro, 1995) and the investigation below focuses on describing the intuition behind these rules. We then present an empirical investigation into these rules in a networking scenario by using several traces of MPEG compressed video.

## The Single-Hop Case

In a single hop network, the tradeoffs between smoothing and queueing delay affect the total end-to-end delay bound as stated in the following theorem.

**Theorem 1** *In the single-hop network with homogeneous sources and deterministic delay bounds, the reduction in queueing delay bound for a FCFS server introduced by smoothing is outweighed by the additional smoothing delay. That is, smoothing will not cause the total delay bound (smoothing plus queueing) to decrease, or $\tau_S > d - \hat{d}$.*

The proof is given in (Knightly and Rossaro, 1995). The theorem states that from a delay bound perspective, smoothing of homogeneous sources should never be performed in the single hop case since it will result in a net increase in total end-to-end delay bound. However, other factors may influence the decision to smooth such as the respective prices of buffers on an end system and inside the network. As well, the theorem is not always true for *heterogeneous* sources.

## The Multi-Hop Case

Theorem 1 showed that over a single hop, smoothing can be an ineffective means for achieving higher network utilization. However, over multiple hops, the analysis has an important additional component: while the smoothing delay is incurred only once at the source's traffic shaper, in a congested network, queueing delays may be incurred at multiple nodes. Thus, a smoother source can reduce its queueing delay at each congested hop resulting in a considerable decrease in its end-to-end delay bound. Then, in many cases, over multiple congested hops this reduction in queueing delay outweighs the added delay caused by smoothing so that smoothing can often result in a net benefit to a network client.

The following proposition provides a rule for determining if smoothing provides a net advantage in terms of end-to-end delay bound for networks that use rate controlled service disciplines such as Rate Controlled Static Priority, Earliest Deadline First, or Hierarchical Round Robin (Zhang and Ferrari, 1994). The rate-controlling aspect of these service disciplines reshapes the traffic at each hop by using (for example) leaky buckets at each node rather than only at the entrance of the network.

**Proposition 1** *Consider a network in which each node $i$ has a rate-controlled service discipline that can provide an upper bound on queueing delay. If $\hat{d}_i$ is the queueing delay bound at hop $i$ for the smoothed source and $d_i$ is the original queueing delay bound at hop $i$, a source will obtain a net reduction in end-to-end delay bound due to smoothing if the following condition holds:*

$$\tau_S < \sum_{i=1}^{H} (d_i - \hat{d}_i) \tag{8}$$

*where $H$ is the number of hops between the source and destination.*

Proof: The proposition states that if the bound on smoothing delay $\tau_S$ is less than the total reduction in queueing delay across multiple hops, then smoothing is advantageous to the source. Because of the decoupling effect of per-hop rate control, the end-to-end delay bound $D$ may be calculated by summing delay bounds $d_i$ at individual nodes (Ferrari and Verma, 1990). Thus, for a smoothed source, we can bound the end-to-end delay as $\hat{D} = \tau_S + \sum_{i=1}^{H} \hat{d}_i + \sum_{i=1}^{H} \pi_i$ where $\pi_i$ is the propagation delay of the $i$'th hop. Without smoothing, the end-to-end delay is given by $D = \sum_{i=1}^{H} d_i + \sum_{i=1}^{H} \pi_i$. A source will have a reduction in end-to-end delay if $\hat{D} < D$. □

Equation 8 show the importance of the number of *congested* hops in determining smoothing conditions. The additional benefit of smoothing can only increase with $H$, since the summands in the right hand side of the inequality are always positive (see Lemma 1). If only a single hop is congested, then for all but one of the hops, $d_i \approx \hat{d}_i$, so that the situation will be similar to the single hop result of Theorem 1.

# EXPERIMENTAL RESULTS

In this section, we use two 30-minute traces of MPEG compressed video to quantify the effects of smoothing on the total end-to-end delay bounds guaranteed to a source. One trace is of an action movie and the other is of a newscast. Both were digitized and compressed at 24 frames per second with frame pattern **IBBPBBPBBPBB**. We consider only a deterministic guarantee in which the network provides a no-loss, no-delay-violation guarantee. The smoothing is also deterministic in that the smoothing FIFO's parameters are selected such that no packets are dropped and the smoothing delay is bounded.

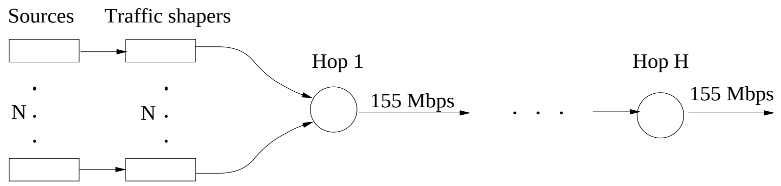



Figure 3: Network Topology

The experiments consider both the single and the multiple hop cases with a simple network topology shown in Figure 3. $N$ connections are shaped by their respective FIFOs and then they are multiplexed at the network nodes. The connections then traverse $H$ hops before reaching their destination.

The experiments are conducted as follows. First, we use the traces to calculate a source's D-BIND parameters.[1] We also calculate the D-BIND parameters under various smoothing intervals $S$. Next, for the given source characterization, we use the admission control test of Equation 3 to determine the maximum number of homogeneous admissible connections on the 155 Mbps link for a given delay bound. We then vary $S$, the number of frames smoothed over, $H$, the number of hops, and use the two different traces. To evaluate the effectiveness of smoothing, we use the following performance metrics: the utilization of the network which is directly proportional to the number of schedulable connections $N$; the net end-to-end savings in delay bound from smoothing connection $j$, $D_j - \hat{D}_j = \tau_{S,j} + \sum_{i=1}^{H}(d_{i,j} - \hat{d}_{i,j})$; and the total end-to-end delay bound $\hat{D}_j = \sum_{i=1}^{H} \hat{d}_{i,j} + \tau_{S,j}$.

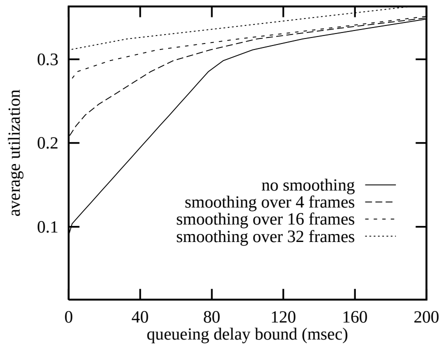


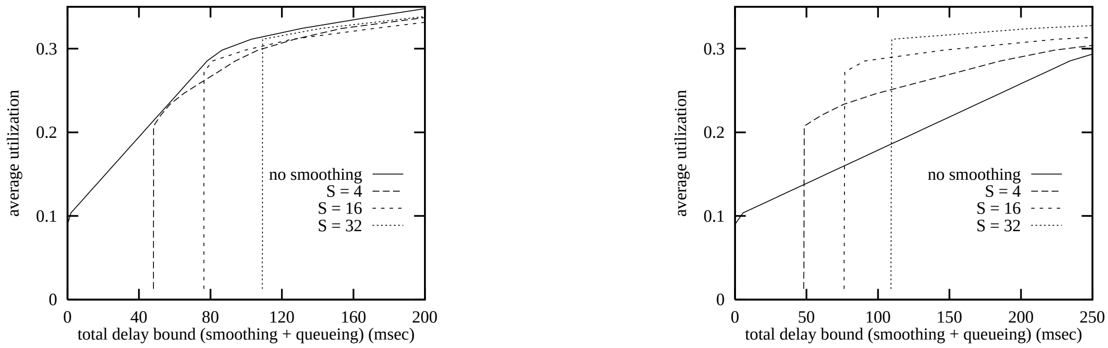

Figure 4: Utilization vs. Queueing Delay for Various Smoothing Policies

The first performance index used in the experiments is average utilization of the multiplexer. It is determined by calculating the maximum number of admissible connections and computing the resulting average network utilization for deterministic service as the sum of the admitted sources' long-term average rates divided by the link speed. Figure 4 shows the effect of smoothing on queueing

[1] In the more practical case where the trace is not available in advance, algorithms in (Zhang and Knightly, 1995) can be used to obtain the D-BIND parameters for a "live" source.

delay only, that is, the delay experienced at a network node (ignoring for the moment smoothing delay). The horizontal axis is queueing delay bound and the vertical axis is the average utilization of the node. The general shape of the curves indicates that as the queueing delay bound increases, more connections are admissible so that a higher utilization is possible. As stated in Lemma 1, as the smoothing interval increases (in this case, from 0 to 4, 16, and 32 frames), or equivalently as the FIFO smoothing rate decreases (to $R_4$, $R_{16}$, and $R_{32}$), the traffic transmitted to the network becomes smoother so that the queueing delay bound is reduced. Equivalently, for a given queueing delay, higher utilizations are achievable inside the network.

However, when the total end-to-end delay bound is considered including both smoothing and queueing delay, the resulting utilization is quite different. As predicted by Theorem 1, smoothing never results in an increased network utilization for a given delay bound in the single-hop homogeneous case. This is shown in Figure 5(a) which depicts average utilization versus total delay bound for various smoothing rates in a single-hop network. For example, for an *unsmoothed* source and a delay bound of 80 msec, a 29% utilization is achievable. Alternatively, when the source is smoothed over $S = 16$ frames, a smoothing delay of 76 msec is introduced so that for a total delay bound of 80 msec, the achievable utilization for 16-frame smoothing is reduced to 27%.



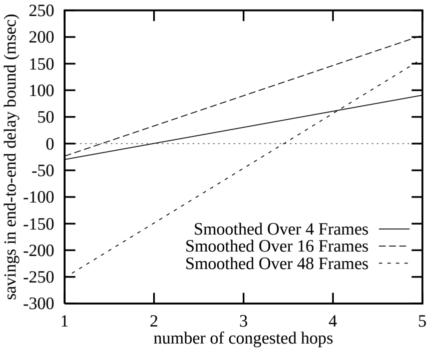
(a) One Hop (b) Three Hops

Figure 5: Average Utilization vs. Total Delay Bound for Various Smoothing Rates

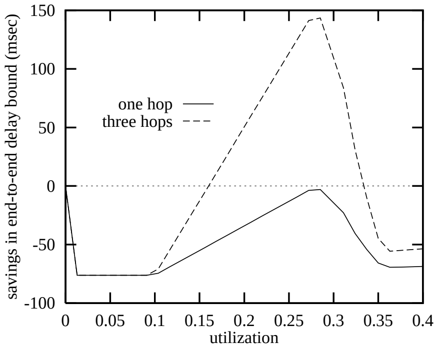
In the case of networks with *multiple* congested hops, the results can be considerably different from the above scenario. Figure 5(b) shows the same experiment described above except that the sources traverse three hops rather than one. In this case, smoothing results in a substantial reduction in a source's end-to-end delay bound. For example, for a total end-to-end delay bound of 100 msec, without smoothing, an average utilization of 17% is achieved. With smoothing over $S = 16$ frame times (with a FIFO of rate $R_{16} = 2.07$ Mbps), this utilization is improved to 29%. Thus, in this case, smoothing has resulted in a 70% increase in the number of admissible connections for the same end-to-end delay bound. This utilization improvement is explained by the rule presented in Proposition 1: while the smoothing delay occurs only once at the source, there is a queueing delay at each of the congested network nodes, thus favoring smoothing over multiple congested hops.

Figure 6(a) shows the effect of the number of congested *hops* (for the network topology of Figure 3) on the savings in end-to-end delay bound. That is, the figure shows the reduction in total delay ($D-\hat{D}$) versus the number of hops ($H$). The number of connections is fixed to 24 so that the average network utilization is 31% in all cases. In the homogeneous case, $D - \hat{D}$ reduces to $H(d_j - \hat{d}_j) - \tau_S$ where $d_j$ and $\hat{d}_j$ are the unsmoothed and smoothed delay bounds at node $j$. Thus, the savings in delay bound $D - \hat{D}$ increases linearly with the number of hops. The lines start negative at one hop (as predicted by Theorem 1) indicating that the smoothing delay $\tau_S$ outweighs the savings in queueing delay. However, since the lines are increasing, this savings in queueing delay eventually becomes positive indicating that as more and more congested hops are traversed, smoothing becomes more advantageous.

Figure 6(b) depicts the *savings* in end-to-end delay bound due to smoothing versus network utilization. The smoothing interval is fixed to $I = 16$ frames and the two curves depict the cases of one and three hops. In the one-hop case, as expected, smoothing does not result in a positive delay savings for any network utilization. Alternatively, for three hops, there is a wide range of utilizations where 16-frame smoothing results in a benefit to the network client with a positive delay savings. Note that even for a three hop network, smoothing is not beneficial if the network is lightly loaded (i.e., operating under low utilization). The reason for this is that in a lightly loaded switch $d \approx \hat{d}$ for any reasonable range of smoothing delays. This emphasizes that our results for a single hop are also an approximation for the case of a single *congested* hop. Also, for higher loads above 0.34, the savings in end-to-end delay again is less than zero even for the three-hop case. The reason for this is that the




(a) Effect of Number of Hops

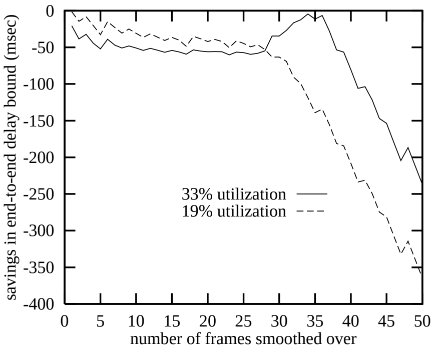



(b) Effect of Load

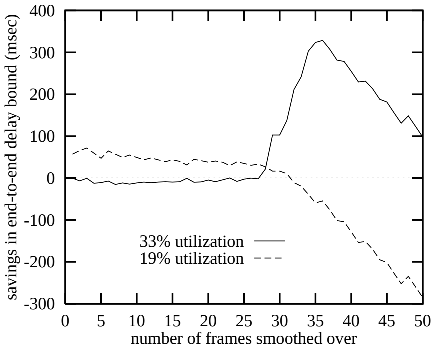
Figure 6: Effect of the Number of Congested Hops and Load on End-to-end Delay Savings

fixed smoothing interval $I = 16$ frames is inappropriate for higher utilizations and a larger smoothing interval would be required.

Figure 7 shows the effect of the smoothing interval (or equivalently, the smoothing rate or number of frames smoothed over) on savings in end-to-end delay $D - \hat{D}$. For the one-hop case, Figure 7(a) shows that more smoothing only worsens the end-to-end delay (decreasing the savings) with the two curves depicting utilizations of 33% and 19%. Figure 7(b) shows the same experiment for a network with three hops. At 33% utilization, there is a clear range of smoothing intervals from $S = 34$ and $S = 39$ where smoothing results in a significant savings in end-to-end delay of up to 320 msec. For the lower network utilization of 19%, smoothing is less effective as described above. The curves of Figure 7 indicate the importance of choosing the correct smoothing interval for a given network load according to the rules of Proposition 1 since improperly chosen smoothing policies can be far worse than doing no smoothing at all.




(a) One Hop

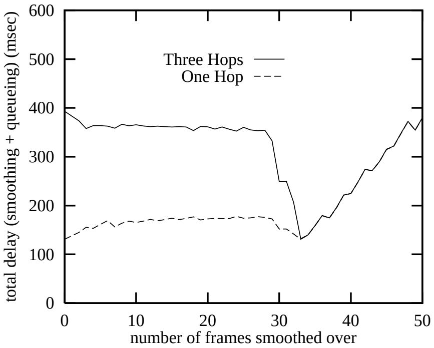



(b) Three Hops

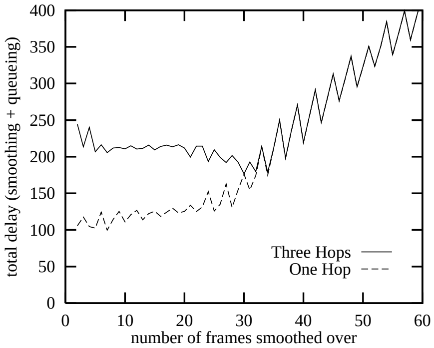
Figure 7: Effect of Smoothing Interval on Savings in Delay Bound

A network client is ultimately interested in its achievable end-to-end delay bound and Figure 8 shows the total end-to-end delay bound including smoothing and queueing delay versus the number of frames smoothed over. Figure 8(a) shows the curves for the MPEG trace of the action movie (the trace used thus far) with 32% network utilization and Figure 8(b) shows the curves for a trace of a newscast with 28% node utilization. The figures indicate that the total delay bound achieved by the network client is quite sensitive to the smoothing interval. For example, if a 150 msec end-to-end delay bound is required for the action movie sequence traversing three congested hops, Figure 8(a) shows that if the network load is 32%, the only two admissible smoothing intervals are 33 and 34 frames. Any other smoothing interval would result in a rejection of connections by the admission control algorithm indicating that the requested 150 msec delay bound cannot be guaranteed at a utilization of 32% for the "improperly" shaped sources.

Finally, we note that the utilizations presented here are for deterministic service only. Unused network resources can always be utilized by sources receiving a statistical performance guarantee or a best effort service. We also note that most of the curves above did not tend to be as smooth as one may expect (i.e., they are jagged in many regions). The primary reason for this can be seen from the rate-interval pairs of Figure 1 which shows that for these traces of MPEG video, the bounding rate does not monotonically decrease with interval length, primarily because of the IBP frame structure of the MPEG algorithm.




(a) Action Movie




(b) Newscast

Figure 8: Effect of Smoothing Rate on Total End-to-end Delay

## CONCLUSIONS

In this paper we have investigated the effects of traffic smoothing on end-to-end deterministic performance guarantees to VBR video. With analysis and experimentation with traces of MPEG-compressed video, we have shown that over a single congested hop, smoothing of homogeneous sources never results in a net reduction in end-to-end delay bound, i.e., the reduction in queueing delay is outweighed by the addition of a smoothing delay. Alternatively, over multiple congested hops, we showed that smoothing can indeed result in a considerable reduction in end-to-end delay bound in many cases. We provided a set of rules for determining whether or not a source should smooth under a given network load.

## ACKNOWLEDGEMENTS

The authors are grateful to Domenico Ferrari and Hui Zhang for their useful comments and suggestions and to Sandia National Laboratories and the University of Wuerzburg for providing the traces of MPEG video.

## REFERENCES

Cruz, R. 1991. A calculus for network delay, part I : Network elements in isolation. *IEEE Transactions on Information Theory*, 37(1):114–121, January.

de Veciana, G. and J. Walrand. 1992. Traffic shaping for ATM networks. Technical Report UCB/ERL M92/135, University of California at Berkeley, Berkeley, CA, December.

Ferrari, D. and D. Verma. 1990. A scheme for real-time channel establishment in wide-area networks. *IEEE Journal on Selected Areas in Communications*, 8(3):368–379, April.

Kanakia, H., P. Mishra, and A. Reibman. 1993. An adaptive congestion control scheme for real-time packet video transport. In *Proceedings of ACM SIGCOMM'94*, pages 20–31, San Francisco, CA, September.

Knightly, E. and P. Rossaro. 1995. Smoothing and multiplexing tradeoffs for deterministic performance guarantees to VBR video. Technical Report TR-95-033, International Computer Science Institute, Berkeley, CA, July.

Knightly, E., D. Wrege, J. Liebeherr, and H. Zhang. 1995. Fundamental limits and tradeoffs for providing deterministic guarantees to VBR video traffic. In *Proceedings of ACM SIGMETRICS'95*, Ottowa, Ontario, May.

Knightly, E. and H. Zhang. 1995. Traffic characterization and switch utilization using deterministic bounding interval dependent traffic models. In *Proceedings of IEEE INFOCOM'95*, pages 1137–1145, Boston, MA, April.

Low, S. and P. Varaiya. 1993. Burstiness bounds for some burst reducing servers. In *Proceedings of IEEE INFOCOM'93*, pages 2–9, San Francisco, CA, March.

Shroff, N. and M. Schwartz. 1994. Video modeling within networks using deterministic smoothing at the source. In *Proceedings of IEEE INFOCOM'94*, pages 342–349, Toronto, Ontario, June.

Zhang, H. and D. Ferrari. 1994. Rate-controlled service disciplines. *Journal of High Speed Networks*, 3(4):389–412.

Zhang, H. and E. Knightly. 1995. RED-VBR: A new approach to support VBR video in packet-switching networks. In *Proceedings of IEEE Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV'95)*, pages 275–286, Durham, NH, April.