Smoothing and Multiplexing Tradeoffs for Deterministic Performance Guarantees to VBR Video

Edward W. Knightly[†] and Paola Rossaro

[†]Also with EECS Department, U.C. Berkeley

TR-95-033

Abstract

The burstiness of variable bit rate traffic makes it difficult to both efficiently utilize network resources and provide end-to-end network performance guarantees to the traffic sources. Generally, smoothing *or shaping traffic sources at the entrance of the network reduces their burstiness to allow higher utilization within the network. However, this buffering introduces an additional delay so that, in effect, lossless smoothing trades queueing delay inside the network for smoothing delay at the network edge. In this paper, we consider the net effect of smoothing on end-to-end performance guarantees where a no-loss, no-delay-violation deterministic guarantee is provided with the D-BIND traffic model. We analytically quantify these tradeoffs and provide a set of general rules for determining under which conditions smoothing provides a net gain. We also empirically investigate these tradeoffs using traces of MPEG compressed video.*

1 Introduction

Future packet-switching integrated services networks must support applications with diverse traffic characteristics and performance requirements. Of the many traffic classes in integrated services networks, delay- and loss-sensitive Variable Bit Rate (VBR) video traffic poses a unique challenge. Indeed, providing good Quality of Service (QoS) to such bursty traffic sources is at odds with achieving high network utilization.

An important issue in providing end-to-end performance guarantees is the effect of *smoothing* traffic sources on network utilization and QoS. 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 performance guarantees. We consider the case of deterministic QoS in which the network provides a no-loss no-delay-violation guarantee so that all packets of a connection will meet their promised QoS. Deterministic performance guarantees are especially important given recent studies that indicate that even a single cell loss for an MPEG video stream can have easily detectable and persistent effects on perceptual quality [10]. Moreover, recent studies in deterministic service indicate that, in many cases, reasonable network utilizations can be achieved even while providing worst-case guarantees [6, 8].

Within this context, we analyze smoothing via 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 that shapes the traffic. 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 FIFO's buffer is sized so that no cells are dropped, thus keeping the *deterministic* end-to-end performance guarantee. We use the Deterministic Bounding Interval Dependent (D-BIND) traffic model to characterize traffic using bounding rates over multiple interval lengths. This model allows for a higher network utilization by providing a more accurate traffic specification than previous models [8].

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 [9]. When the smoothed, less bursty traffic sources are multiplexed at queues inside the network, the resulting bound on queueing delay will be reduced. Equivalently, with less bursty traffic, more admissible connections or higher utilization is achievable inside the network for a given queueing delay bound. However, the smoothing FIFO at the source has increased the end-to-end delay bound by adding a smoothing delay. Thus, a bound on the smoothing delay must also be accounted for when considering the total end-to-end delay bound.

With the D-BIND model and associated admission control techniques [8] used to bound queueing delay, and the techniques described in this paper to bound smoothing delay, we analyze the effect of smoothing VBR video sources on the *total* end-to-end delay bound considering 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 show analytically and demonstrate empirically that for homogeneous sources traversing a *single* hop, smoothing at the source 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. The same result can be expected to hold in networks where the total queueing delay is dominated by a single "bottleneck node."

We then consider the case of *multi-hop* networks and present a set of general rules to determine if smoothing will result in a net reduction in end-to-end delay bound. Analysis and experimentation indicate that in many cases of multi-hop networks with multiple congested hops, smoothing may be an *effective* means of reducing total end-to-end delay. 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, a higher network utilization or lower end-to-end delay can often be achieved by smoothing. Moreover, the analysis indicates that the advantages of smoothing increase with the number of congested hops traversed by the connection.

The remainder of this paper is organized as follows. In Section 2, we review the D-BIND model and its application to providing deterministic performance guarantees. In Section 3, we provide a definition of smoothing. Next, in Sections 4 and 5 we use analysis to provide a set of rules for when a source should smooth in the single-hop and multi-hop case respectively. We also use traces of MPEG-compressed video to evaluate the rules in a realistic scenario. Finally, in Sections 6 and 7 we discuss related work and conclude.

2 A Deterministic Approach

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.

2.1 D-BIND Model

As shown in [8], previous deterministic traffic models such as the (σ, ρ) model [1] and the $(Xmin, Xave, I, Smax)$ model cannot capture the property that sources exhibit burstiness over a wide variety of interval lengths. The Deterministic Bounding Interval Dependent (D-BIND) 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 [8]).

Each deterministic traffic model defines 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 $A[s, s+t] \leq b(t),\ \forall s, t > 0$. Note that $b(t)$ is a time-invariant deterministic bound since it constrains the traffic source over every interval of length t.

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

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

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

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

Figure 1 shows a plot of the D-BIND rate-interval pairs for a 28 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 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).

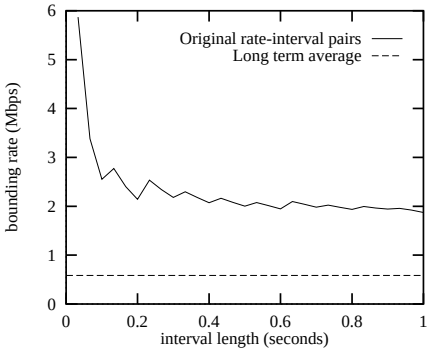

Figure 1: D-BIND Curve for Action Movie

Figure 2 shows the D-BIND constraint function $b(t)$ (described by Equation (1)) for the movie. The figure shows maximum bits (in kbits) that the source transmits over any interval of length t (shown in seconds). The figure indicates that the D-BIND model is capturing the temporal properties of the MPEG video. For example, the peak rate shown in Figure 1 is caused by transmission of the largest I frame of the sequence. This can be seen in the constraint function with the large 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 [8] 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.

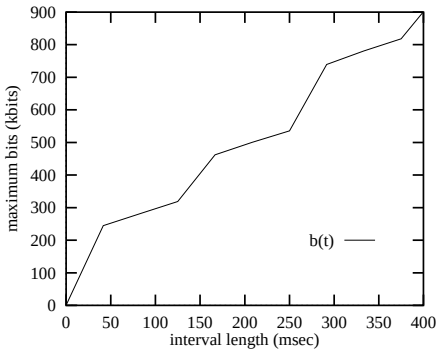

Figure 2: D-BIND Constraint Function for News Sequence

2.2 Connection Admission Control

Deterministic admission control conditions rely on the delay analysis techniques of [1, 14] which are illustrated in Figure 3. The horizontal axis is time and the vertical axis is bits. The upper curve represents the total number of cumulative bits that have arrived from all sources into the queue by time t; the lower curve represents the total number of bits transmitted by time t. The difference between the two curves is the number of bits currently in the queue, or the *backlog* function. When the backlog function returns to zero (the two curves meet) there are no bits in the queue and thus a busy period has ended. The key to this analysis is that if the upper curve is a deterministic bounding curve (the sum of the $b(t)$ constraint curves), then the maximum delay can be expressed as a function of the two curves. For example, the maximum backlog divided by the link speed provides an upper bound on delay for a FCFS server. Delay bounds for priority service disciplines that are more suitable for providing integrated services may be expressed in a similar manner [6].

The constraint function provides the required bound on arrivals in any interval of length t, so that with the aggregate of individual source's respective $b(t)$ constraint functions forming the upper curve of Figure 3, admission control conditions for deterministic delay and throughput bounds may be derived. For example, for a FCFS scheduler with $j = 1, ..., n$ multiplexed connections constrained by their respective constraints $b_j(t)$, and with a link speed l, and a maximum packet size $\overline{s}$, a deterministic upper bound on delay for all

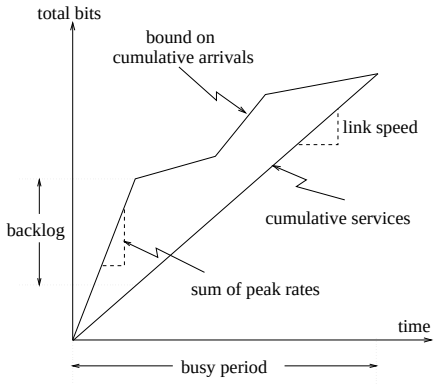

Figure 3: Admission Control Algorithm

connections is given by

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

The proof is given in Theorem 1 of [14]. In the homogeneous case, the maximum number of connections that can be multiplexed for a delay bound d is therefore given by

$$N(d) = max\{n \mid \frac{1}{l} \max_{t \geq 0} \{ nb(t) - lt + \overline{s} \} \leq d \}. \tag{4}$$

2.3 Achievable Network Utilization

Figure 4 shows the achievable network utilization for the case of multiplexing a number of these action movies. That is, for a 155 Mbps link and the FCFS admission control test of Equation (4), the figure shows the average utilization that corresponds to the maximum number of homogeneous connections that can be multiplexed so that all connections obtain the given deterministic delay bound shown on the horizontal axis. For example, for a delay bound of 40 msec, 51 connections can be multiplexed so that no packets are dropped and none violate their delay bounds. This situation corresponds to an average network utilization of 19%. Alternatively, for a peak-rate-allocation scheme, 26 connections can be multiplexed which results in an average network utilization of 9%. Thus network utilizations well above those achieved with peak-rate-allocation are possible even though there is a fundamental limitation to the achievable utilization that a deterministic service can provide in that both the source characterization and multiplexing behavior are calculated in a worst-case manner.

In the next section, we investigate the impact of smoothing on achievable utilization, still considering the case of deterministic bounds on delay and loss.

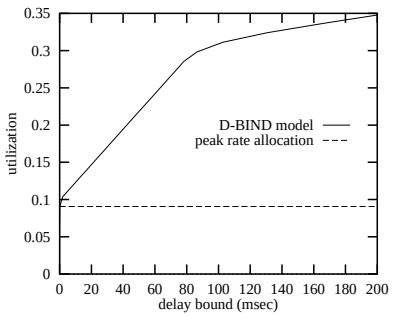

Figure 4: Achievable Utilization for Action Movie

3 Definition of Smoothing

Smoothing traffic in the case of the D-BIND model may be viewed from either of two equivalent viewpoints: the (R_k, I_k) rate-interval pairs or the $b(t)$ constraint function. As shown in Figure 1, a source may be described by bounding rates over multiple interval lengths. Equation (1) indicates that a lower rate for a given interval length will result in a lower constraint curve and thus, from Equation (3) possibly more admissible connections in the network. Thus, 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 $\hat{R}_k \leq R_k$ if $\hat{I}_k = I_k$. This transformation is realized with the smoothing or buffering at the traffic shaper. The mechanisms for achieving this are similar to those used for policing such as multiple leaky buckets or multiple non-concave buckets as described in [8]. The primary difference is that we are considering smoothing policies that only buffer, and do not drop, incoming packets.

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

$$\hat{b}(t) = \frac{\hat{R}_k \hat{I}_k - \hat{R}_{k-1} \hat{I}_{k-1}}{\hat{I}_k - \hat{I}_{k-1}} (t - \hat{I}_k) + \hat{R}_k \hat{I}_k, \quad \hat{I}_{k-1} \leq t \leq \hat{I}_k. \tag{5}$$

With $\hat{R}_k \leq R_k$ and Equations (1) and (5), we have that $\hat{b}(t) \leq b(t)\ \forall t$. More formally, we define smoother traffic in the following manner (similar to the definition in [9] 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 $\hat{b}(t) \leq b(t)\ \ \forall t$.*

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. A description of the network and the FIFO traffic shaper is given in Figure 5.

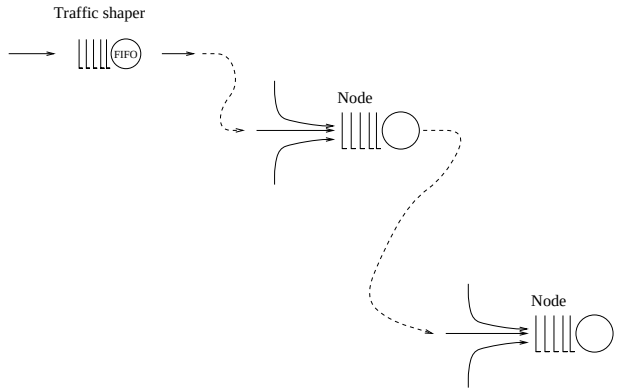

Figure 5: Description of the FIFO traffic shaper

The buffer occupancy of the smoothing FIFO is then given by

$$q(t) = \sup_{0 \le s \le t} \{A[s,t] - R(I_S) \cdot (t-s)\}. \tag{6}$$

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

$$\hat{A}[t, t+\delta] = \min(R(I_S) \cdot \delta, A[t, t+\delta] + q(t)). \tag{7}$$

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.

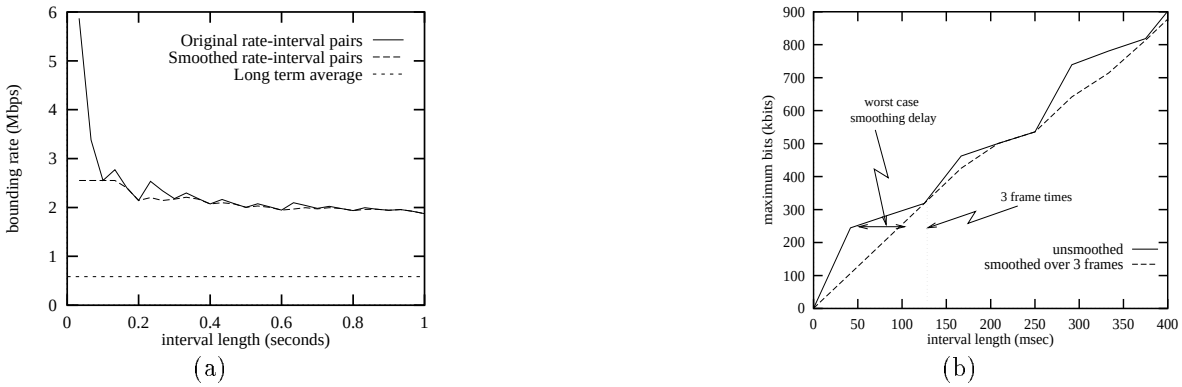

Figure 6: Smoothed D-BIND rate-interval and constraint curves

The transformation of smoothing on a source's rate-interval pairs is shown in Figure 6(a) which 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 $R_1 = 5.87 Mbps$. 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 $R_3 = 2.55 Mbps$), 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 6(b) shows the effect of smoothing on a source's constraint function. As shown, τ_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 $\hat{b}$ [1]. That is,

$$\tau_S = \max_{t_2 > t_1} \{t_2 - t_1 | \hat{b}(t_2) = b(t_1)\}. \tag{8}$$

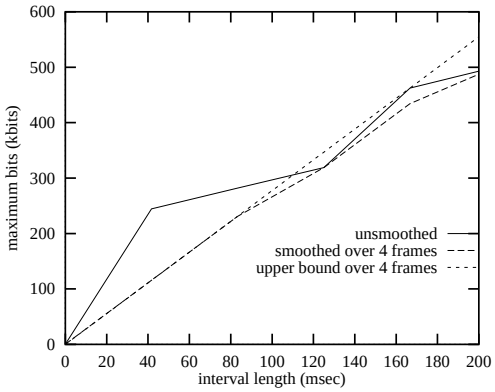

Figure 7: Upper bound on smoothing delay for $S = 4$ frames

From the definition of smoothing, $\hat{R}_k \leq R_S$, and in most cases, $\hat{R}_{j=1,2,\cdots,S-1} = R_S$. If this is not the case (i.e., $\hat{R}_{j=1,2,\cdots,S-1} \neq R_S$), then $\hat{b}(t)$ can be upper bounded by $R_S\, t$ for $0 \leq t \leq I_S$. Figure 7 shows an example of the smoothed curve $R_S\, t$, for $S = 4$ frames. We can then define τ' as the distance between the $b(t)$ and $R_S\, t$ so that τ' represents a lower bound on the smoothing delay τ_S, since it is the minimum distance between the two curves.

We note that other smoothing functions are possible. That is, any desired piece–wise linear constraint curve $\hat{b}(t)$ can be constructed with the D-BIND model. The traffic could be shaped to conform to the new $\hat{b}(t)$ with the policing or traffic shaping mechanism of the D-BIND model as described in [8]. In this work, we are considering transformations from b to $\hat{b}$ that are caused by smoothing with FIFO's, where b and $\hat{b}$ are derived from the D-BIND model.

4 The Single Hop Case

In this section, we evaluate the smoothing and multiplexing tradeoffs for the single hop case via analysis and experiments with traces of MPEG compressed video.

4.1 General Rules

Intuitively, a less bursty sources can better utilize network resources. This is stated in the following Lemma.

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 $\hat{d} \leq d$. □

Since network utilization is related to the number of admissible connections under the delay bound constraint, the above Lemma equivalently states that for the same delay bound d, potentially more sources could be multiplexed (i.e., N could be increased to $\hat{N}$) resulting in higher network utilization.

Thus, for a given set of connections, smoothing over S frames may decrease the queueing delay d to $\hat{d}$. On the other hand, smoothing introduces a delay τ_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 delay it incurs by smoothing is less than the reduction in the queueing delay bound, i.e., if the total delay bound, smoothing plus queueing, is decreased. The following Theorem shows that for a single-hop network and homogeneous sources, smoothing does not result in a net decrease in delay bound.

Theorem 1 *In the single-hop case for homogeneous sources and deterministic delay bounds, the reduction in queueing delay bound 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 $\tau_S > d - \hat{d}$.*

Proof: Consider a source described by the D-BIND model parameters $(R_k, I_k)_{k=1}^{P}$. Let $R_S t$ be the upper bound on the smoothed constraint function as described in section 3 with the maximum distance between these two curves defined as τ'. If the proof is valid for τ' it will be valid for all the $\tau_S \geq \tau'$. Thus, in the rest of the proof, we assume that $\hat{b}(t)$ and τ_S correspond to their upper and lower bound $R_S t$ (for $0 \leq t \leq I_S$) and τ' respectively.

Because of subadditivity of the constraint function $b(t)$ (see [8]), the worst case smoothing delay τ_S (defined in Equation (8)) occurs between $t = 0$ and $t = I_S$. From Equation (8), the maximum time difference between b and $\hat{b}$ is given by:

$$\tau_S = \max_{0\leq t\leq I_S}\{\frac{t}{R_S} - b^{-1}(t)\} \tag{9}$$

Since this maximum will occur at t_j for $1 \leq j \leq S$,

$$\tau_S = \max_{1\leq j\leq S}\{\frac{R_j I_j}{R_S} - I_j\} \tag{10}$$

$$= \max_{1 \le j \le S} \{ I_j (\frac{R_j}{R_S} - 1) \}. \tag{11}$$

From the perspective of a source's total delay bound, the source should smooth if the delay it incurs by smoothing, τ_S, is less than the reduction in the queueing delay bound $d - \hat{d}$. For this reason, we next investigate the *savings* in queueing delay, $d - \hat{d}$, achieved by transmitting the smoother traffic.

When multiplexing N homogeneous sources, each with D-BIND parameters $(R_k, I_k)_{k=1}^{P}$, for a FCFS scheduler with link speed l, the queueing delay bound for all unsmoothed sources can be obtained by Equation (3) as:

$$d = \frac{1}{l} \max_{1 \le j \le P} \{ N\, b(I_j) - l\, I_j \} \tag{12}$$

If this delay bound occurs at the interval length I_α (i.e., Equation (12) is maximized at $j = \alpha$), then Equation (12) can be rewritten as

$$d = (N\, R_\alpha\, I_\alpha - l\, I_\alpha)/l \tag{13}$$

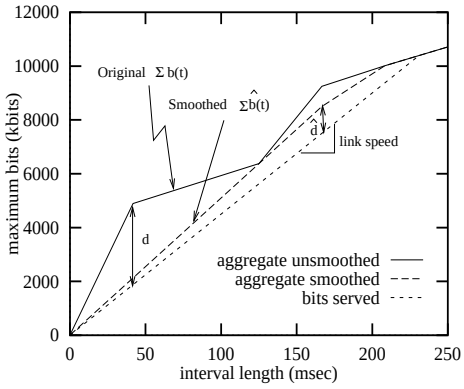

Figure 8: Effect of Smoothing on Delay Bound

Since the worst case smoothing delay will also occur at this point [1], Equation (11) becomes:

$$\tau_S = I_\alpha (\frac{R_\alpha}{R_S} - 1) \tag{14}$$

As in the case of Figure 8, with smoothing, the worst case queueing delay is reduced and shifted from the interval I_α to I_S. If this were not the case, there will be a j such that $\hat{d}_j > \hat{d}_S$. Since we want to prove that $\tau_S > d - \hat{d}$ it is sufficient to show that $\tau_S > d - \hat{d}_S > d - \hat{d}_j$. The new delay bound for the smoothed source is therefore

$$\hat{d} = (N\, \hat{R}_S\, I_S - l\, I_S)/l. \tag{15}$$

The savings in queueing delay due to smoothing is thus

$$d - \hat{d} = I_\alpha (\frac{N\, R_\alpha}{l} - 1) - I_S (\frac{N\, R_S}{l} - 1) \tag{16}$$

since $R_S = \hat{R}_S$.

A source should smooth if the smoothing delay is less than the savings in queueing delay, i.e., if

$$\tau_S < d - \hat{d} \tag{17}$$

Substituting with Equations (14) and (16), the source should smooth if:

$$\begin{aligned} I_\alpha(\frac{R_\alpha}{R_S} - 1) &< I_\alpha(\frac{N\ R_\alpha}{l} - 1) - I_S(\frac{N\ R_S}{l} - 1) \\ I_S R_S &< I_\alpha R_\alpha \\ b(I_S) &< b(I_\alpha) \end{aligned}$$

The latter inequality is never true since $b(\cdot)$ is strictly increasing by definition and $I_S > I_\alpha$ $\Box$.

Theorem 1 therefore states that from a delay bound perspective, smoothing should never be performed in the single hop case since it will result in a net increase in total 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.

4.2 Experiments with MPEG Traces

The experiments below use 28 minute trace of the MPEG-compressed action movie together with the analysis developed above.

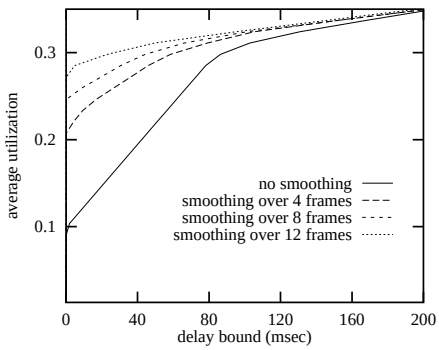

Figure 9: Average Utilization vs. Queueing Delay Bound for Various Smoothing Rates

Figure 9 demonstrates Lemma 1. For the MPEG trace data and a 155 Mbps link, the figure shows the average network utilization versus queueing delay. This utilization is determined by calculating the maximum number of admissible connections using the admission control condition of Equation (3) and computing the resulting average network utilization for deterministic service as the sum of the sources' long-term average rates divided by the link speed. As stated in the lemma, as the smoothing interval increases from 0 to 4,

8, and 12 frames (or equivalently as the FIFO smoothing rate decreases to R_4, R_8, and R_{12}), the traffic becomes smoother so that for a given queueing delay, higher utilizations are achievable inside the network.

Figure 10 shows the other side of the tradeoff – that smoothing has introduced an additional delay in the smoothing FIFO. Figure 10 depicts smoothing delay versus S, or the number of frames smoothed over. As expected, the general trend is that smoothing over more frames introduces a larger smoothing delay. This curve is not monotonic though since the rate-interval curve shown in Figure 1 is not monotonic and the FIFO rates are determined by these D-BIND bounding rates.

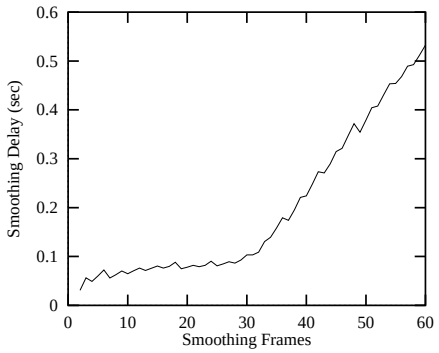

Figure 10: Smoothing Delay vs. Number of Frames

A source is ultimately interested in the combined effects of Figures 9 and 10. Thus, Figure 11 shows the net effect of smoothing on a source's total delay bound depicting average utilization versus total delay bound (smoothing plus queueing) for various smoothing rates. As predicted by Theorem 1, smoothing never results in a decrease in total delay for a given number of admissible connections (or utilization). Equivalently, smoothing never results in an increased network utilization for a given delay bound in the single-hop homogeneous case.

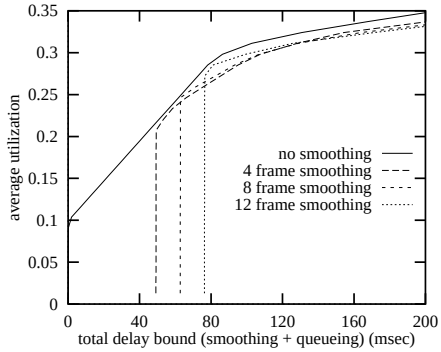

Figure 11: Average Utilization vs. Total Delay Bound for Various Smoothing Rates

5 The Multi-hop Case

The previous section showed that over a single hop, smoothing is 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.

5.1 General Rules

As described in Section 4, smoothing can be considered advantageous to the source if the incurred smoothing delay τ_S is less than the reduction in queueing delay. The analysis of [7] shows that per-hop rate control is required to achieve reasonable network utilization in multi-hop networks. That is, if rate control, such as the multi-level leaky bucket, is not used at each hop of the network (as opposed to only at the network edges), there will be a severe utilization penalty. For this reason, we consider the case in which rate control is performed at each hop as in the RCSP scheduler. With per-hop rate control, the original constraint function $b(t)$ (or $\hat{b}$ if the source is smoothed) is reconstructed at each node via the same policing mechanism that enforces the original curve.

The following proposition provides a rule for determining if smoothing provides a net advantage in terms of end-to-end delay bound.

Proposition 1 *Consider a network where 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:*

$$\tau_S < \sum_{i=1}^{H} (d_i - \hat{d}_i) \tag{18}$$

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

Proof: The proposition states that if the bound on smoothing delay τ_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 then be calculated by summing delay bounds d_i at individual nodes [3] (while this gives a sufficient condition for end-to-end delay bound, tighter bounds may be possible by extending techniques such as those in [5]). Thus, for a smoothed source, we can bound the end-to-end delay as:

$$\hat{D} = \tau_S + \sum_{i=1}^{H} \hat{d}_i + \sum_{i=1}^{H} \pi_i \tag{19}$$

where π_i is the propagation delay of the i'th hop. Without smoothing, the end-to-end delay is given by

$$D = \sum_{i=1}^{H} d_i + \sum_{i=1}^{H} \pi_i \tag{20}$$

A source will have a reduction in end-to-end delay if $\hat{D} < D$ which is the inequality stated in the proposition. □

Equation 18 indicates that as the number of *congested* hops increases, smoothing becomes more important. The reason for this is that the summands in the right hand side of the inequality are always positive (see Lemma 1) and therefore the additional benefit of smoothing can only increase with H. If only a single hop is congested, then for all but one of the hops, $d_i \approx \hat{d}_i$, and the situation will be similar to the single hop result of Theorem 1.

Because rate controlled service disciplines decouple the network nodes, Proposition 1 applies to heterogeneous networks consisting of a wide range of service disciplines at the nodes [15]. The local delay bounds that constitute the sum in Equation (18) may then be calculated by the admission control tests that correspond to the local service discipline (see [6] for several examples).

In practice, the smoothing decision could be made if the source specifies both its original and smoothed D-BIND parameters to the network. During the connection establishment, the admission control algorithms can test both traffic specifications so that both d and $\hat{d}$ can be calculated at each hop. The source can then make its choice on whether or not to smooth based on Equation (18). Additional criteria are also possible. For example, a source could choose I_S based on information about the state of the network including both current load and a pricing policy. Ideally, a source would specify its D-BIND parameters, an end-to-end delay bound, and a desired price, and the network would return the appropriate smoothing rate R_S to meet those constraints. An appropriate pricing policy could then encourage more optimal use of network resources.

The following subsection uses a trace of MPEG-compressed video to evaluate Proposition 1 under a realistic workload.

5.2 Experiments with MPEG Traces

The experiments below use a network topology as shown in Figure 12 to evaluate the smoothing criteria of Proposition 1. The network consists of nodes with 155 Mbps (OC-3) and 1 Gbps (OC-24) link speeds. All sources are assumed to have the same D-BIND rate-interval pairs as obtained from the MPEG trace. As shown in the figure, the sources are shaped at the entrance to the network with a FIFO as described in Section 3. After being smoothed, the sources traverse five hops as shown with the dashed line in the figure: the first and last hop have a 155 Mbps link speed and the intermediate hops have a 1 Gbps link speed. In the experiment, connections are established across the network until further connections can not be admitted without violating the admission control conditions of Section 2.

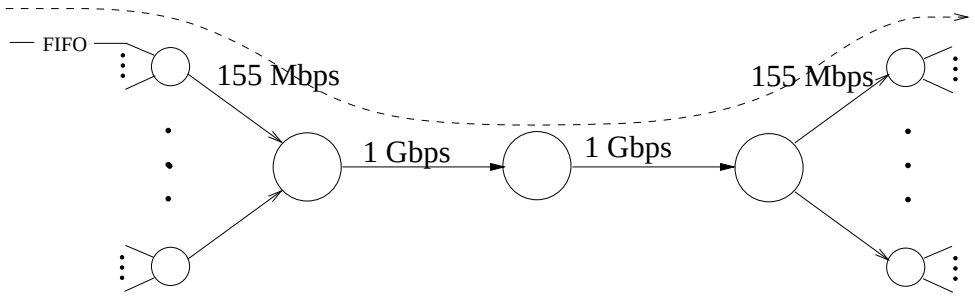

Figure 12: Network Topology for Multi-hop Experiment

For the above network, Figure 13 shows average utilization versus total end-to-end delay bound for various smoothing rates of R_5, R_{20}, R_{30}, and R_{40} as well as for no smoothing. The end-to-end delay bound includes the bounds on smoothing delay as well as queueing delay at each hop (propagation delay could be included by shifting all curves to the right by the total propagation delay). As shown in the figure, smoothing can result in a significant increase in utilization in many cases. This increase depends on both the smoothing rate and the sources' desired end-to-end delay bound. For example, if the source requires an end-to-end delay bound of 125 milliseconds, it cannot smooth with rate $R_{40} = 1.64$ Mbps since this rate is too low and would cause end-to-end delays beyond 125 msec (the maximum smoothing delay for R_{40} is 224 msec). Thus, for a 125 msec end-to-end delay bound a smaller smoothing interval or equivalently, a higher smoothing rate must be used. In the figure, a smoothing rate of $R_{30} = 1.87Mbps$ will result in a smoothing delay bound of 103 msec. Together with a 22 msec bound in total queueing delay, the resulting end-to-end delay bound of 125 msec results in an average network utilization of 31.1%. Comparatively, without smoothing, the resulting utilization is 15.6% for the same delay bound. Thus, smoothing has resulted in a 99% increase in utilization in this case.

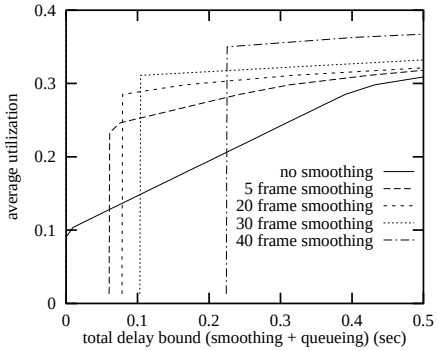

Figure 13: Average Utilization vs. Total End-to-end Delay Bound for the Five Hop Network

6 Related Work

In the recent literature, traffic shaping or smoothing has received much attention in different contexts. For example, in [4], the authors use similar techniques to those used here to show that with smoothing at the source, end-to-end delays with rate controlled service disciplines are strictly less than with Rate Proportional Processor Sharing (RPPS).

In a related work on deterministic guarantees, the authors of [13] consider a traffic shaper for a peak-rate, bounded-burst-length deterministic traffic model. The shaping they consider affects only the peak rate or cell spacing. For a single hop they argued that with minimal delay (less than 1 msec), significant utilization gains are possible. However, this gain is overstated since a peak-rate-allocation is assumed which was shown in [1, 14] to be an unnecessary condition for providing deterministic service.

Several other works consider the case of traffic shaping in terms of statistical performance guarantees. For example, [2] considers smoothing via periodic averaging of a sources rate (PARing). Large deviations techniques are then used to determine buffering requirements at the source and loss probabilities inside the network. The bounds on smoothing delay considered here are considerably less than for PARing in which a source's arrivals are buffered for a fixed interval and released smoothly into the network.

In [11], the authors consider "deterministic smoothing" of MPEG traffic sources in which the cells of a video frame are transmitted in equally spaced intervals over the frame time. They then consider end-to-end statistical guarantees using the histogram techniques introduced in [12]. In this work, we have considered a *deterministic* approach that does not allow either the shaper or the network to discard packets. Thus, we are considering the case of end-to-end deterministic performance guarantees and we quantify the tradeoffs involved in smoothing in terms of both the user's QoS and the network's utilization.

7 Conclusions

Smoothing via lossless buffering at the source introduces a tradeoff between smoothing delay at the entrance of the network and queueing delay inside the network. We have considered the effects of smoothing for the case of end-to-end deterministic performance guarantees using the D-BIND traffic model to characterize the sources.

For the case of a network with a single congested hop, we showed that with homogeneous sources, smoothing never results in a net reduction in total delay bound. That is, while smoothing does reduce the bound in queueing delay, this reduction in delay is outweighed by the added smoothing delay. We confirmed this result and provided insights into typical delays incurred by a realistic workload with experiments involving a 28 minute trace of MPEG-compressed video.

Over multiple congested hops, the situation changes to favor smoothing primarily because while the

smoothing delay is incurred only once, the queueing delay may be incurred at each congested hop. We provided a set of general rules that can be used to determine if smoothing will result in a net decrease in a source's total end-to-end delay bound. Using the MPEG trace, we demonstrated that in many cases smoothing indeed resulted in a net reduction in end-to-end delay bound. For example, for a 5 hop network and an end-to-end delay bound of 125 milliseconds, smoothing by an appropriate rate resulted in a 99% increase in average network utilization.

As a final note, we caution that since the end-to-end delay bound calculations are not tight (i.e., necessary and sufficient), the analysis presented here may excessively favor smoothing. Thus, with more efficient calculations of end-to-end delay bounds, the rules presented in Section 5 can be improved. As well, the issue of the relative prices of buffers at the end system and inside the network (and other charging policies) will also affect a source's decision whether or not to smooth.

References

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

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

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

[4] L. Georgiadis, R. Guerin, and V. Peris. The effect of traffic shaping in efficiently providing end-to-end performance guarantees. Technical Report RC 20014, IBM Research Division, Yorktown Heights, NY, April 1995.

[5] P. Goyal, S. Lam, and H. Vin. Determining end-to-end delay bounds for heterogeneous networks. In *Proceedings of IEEE Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV'95)*, pages 287–298, Durham, NH, April 1995.

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

[7] E. Knightly and H. Zhang. D-BIND: A new traffic model for providing deterministic QoS to VBR video, April 1995. Submitted for publication.

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

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

[10] I. Richardson and M. Riley. Usage parameter control cell loss effects on MPEG video. In *Proceedings of ICC'95*, pages 970–974, Seattle, WA, June 1995.

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

[12] 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.

[13] N. Yamanaka, Y. Sato, and K. Sato. Traffic shaping for VBR traffic in ATM networks. *IEICE Transactions Communications*, E75-B(10):1105–1108, October 1992.

[14] H. Zhang and D. Ferrari. Improving utilization for deterministic service in multimedia communication. In *Proceedings of 1994 International Conference on Multimedia Computing and Systems*, pages 295–304, Boston, MA, May 1994.

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