# Guaranteeing Statistical QoS to Regulated Traffic: The Single Node Case *

Martin Reisslein †
University of Pennsylvania
reisslei@seas.upenn.edu

Keith W. Ross
Eurecom, France
ross@eurecom.fr

Srini Rajagopal
Manugistics
srajagop@manu.com

June 1998

**Abstract**

Multimedia traffic can tolerate some loss but has rigid delay constraints. A natural QoS requirement for a multimedia connection is a prescribed bound on the the fraction of traffic that exceeds an end–to–end delay limit. We propose and analyze a traffic management schemes which guarantees QoS to multimedia traffic while simultaneously allowing for a large connection-carrying capacity. We study our traffic management scheme in the context of a single node. In order for the node to guarantee QoS, each connection's traffic is regulated. In order to support many connections, the link statistically multiplexes the connections' traffic. The scheme consists of (i) cascaded leaky-buckets for traffic regulation, (ii) smoothers at the ingresses, and (iii) bufferless statistical multiplexing within the node. For this scheme we show that loss probabilities are minimized with simple one–buffer smoothers which operate at specific minimum rates. We also show that the worst–case input traffic is extremal on–off traffic for all connections. These two results lead to a straightforward scheme for guaranteeing QoS to regulated traffic. Using MPEG video traces, we present numerical results which demonstrate the methodology. Finally, we compare the bufferless scheme with buffered statistical multiplexing.

## 1 Introduction

Over the past ten years, significant research effort has addressed the important problem of guaranteeing QoS to multimedia traffic in a packet–switched network. The goal has been to develop traffic management schemes that allow for high link utilizations while simultaneously guaranteeing that the QoS requirements of the ongoing connections are met. It is generally agreed that high link utilizations can only be achieved by allowing traffic to be statistically multiplexed, i.e., by allowing each connection's traffic to have a small amount of loss and exploiting the statistical independence of the connections' traffic [24][13][14][11]. It is also the view of many researchers that QoS can only be guaranteed by requiring the traffic to be regulated (e.g., by leaky buckets) at the edges of the network [15][28] [10][20][8][18][21].

*Supported partially by NSF grant NCR96–12781.

†Corresponding Author: Martin Reisslein, Department of Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104–6315, phone: (215) 898-4772, fax: (215) 573-2065.

In recent years the problem of providing QoS guarantees to regulated sources which are statistically multiplexed in a shared buffer has been carefully studied [8][18][21]. The existing solutions, however, do not extend to the network environment in a satisfactory manner. Also in recent years, the problem of providing end–to–end deterministic guarantees to regulated traffic in *networks* has been adequately solved [30][29][10][20]. The deterministic QoS guarantees, however, typically imply a small connection–carrying capacity for networks with bursty multimedia traffic. In this paper we lay the groundwork for a traffic–management architecture that provides end–to–end statistical QoS guarantees. We focus our attention to a network consisting of a single node in this paper. We extend the traffic mangement to networks in a subsequent paper [23].

In this paper we view traffic as fluid. The fluid model, which closely approximates a packetized model with small packets, permits us to focus on the central issues and significantly simplifies notation. We suppose that the traffic sent into the node by each connection is regulated by a connection–specific cascade of leaky buckets. A cascade of leaky buckets is more general than the two–leaky–bucket regulator, commonly used in the literature [8][18], and can more accurately characterize a source's traffic. Moreover, cascaded–leaky–bucket traffic can easily be policed. For admission control, all that we know about a connection's traffic is its regulator constraint defined by its cascade of leaky buckets; in particular, we do not have available statistical characterizations of the traffic.

We also assume that the following natural QoS requirement is in force: the fraction of traffic that exceeds a specific delay limit must be below a prescribed bound. Traffic which overflows at a buffer is considered as having infinite delay, and therefore violates the QoS requirement. Importantly, we permit each connection to have its own limit on the nodal delay and its own bound on the fraction of traffic that exceeds this delay limit. This QoS requirement is particularly appropriate for multimedia traffic, whereby timestamping and a playout buffer can ensure the continuous playout of video or audio without jitter.

Given each connection's traffic characterization and its QoS requirement, we address the following problem: How should we manage the traffic and perform admission control in order to guarantee QoS while maintaining a large connection-admission region? We advocate the following simple and pragmatic scheme: (*i*) smooth each connection's traffic at the connection's input as much as allowed by the connection's delay constraint; (*ii*) employ bufferless statistical multiplexing within the node; (*iii*) base admission control on the worst-case assumption that sources are adversarial to the extent permitted by the connection's regulator, while concurrently assuming the connections generate traffic independently. This scheme enjoys the following features:

- Admission control is solely based on the connections' regulator parameters, which are policable. It is not based on more complex, difficult–to–police statistical characterizations.
- It allows for statistical multiplexing at the node while meeting the QoS requirements. The smoothing at the input increases the statistical multiplexing gain.
- It allows for per-connection QoS requirements: the connections can have vastly different delay and loss requirements.

- Because the multiplexing is bufferless, the switch requires only small input buffers (when traffic is packetized), thereby reducing switch cost.
- A connection's traffic characterization does not change as the traffic passes through the bufferless multiplexer.

It is this last feature that is particularly useful when extending the traffic management scheme to a multihop network [23]. With our scheme the traffic leaving the network node conforms to the same regulator constraints as the traffic entering the node. With shared buffer multiplexers it is difficult (if not impossible) to tightly characterize a connection's traffic once the traffic passes through a shared buffer.

This paper is organized as follows. In Section 2 we formally define the cascaded leaky-bucket regulators and the QoS requirement. In Section 3 we determine the worst-case traffic for a single-link and outline our smoothing and admission control procedure. We also consider general smoothers and show that the optimal smoother is a single-buffer smoother which smoothes traffic as much as the delay limit permits. In Section 4 we present numerical results using MPEG–encoded traces. In Section 5 we compare our scheme to designs based on buffered statistical multiplexing. We conclude in Section 6.

## 2 Regulated Traffic and the QoS Requirement

In this paper we focus on a single node consisting of a bufferless multiplexer that feeds into a link of capacity $C$. We view traffic as fluid, i.e., packets are infinitesimal. Consider a set of $J$ connections. Each connection $j$ has an associated regulator function, denoted by $\mathcal{E}_j(t)$, $t \geq 0$. The regulator function constrains the amount of traffic that the $j$th connection can send into the node over all time intervals. Specifically, if $A_j(t)$ is the amount of traffic that the $j$th connection sends to the node over the interval $[0, t]$, then $A_j(\cdot)$ is required to satisfy

$$A_j(t+\tau) - A_j(\tau) \leq \mathcal{E}_j(t) \text{ for all } \tau \geq 0, \quad t \geq 0. \tag{1}$$

A popular regulator is the simple regulator, which consists of a peak–rate controller in series with a leaky bucket; for the simple regulator, the regulator function takes the following form:

$$\mathcal{E}_j(t) = \min\{\rho_j^1 t,\ \sigma_j^2 + \rho_j^2 t\}.$$

For a given source type, the bound on the traffic provided by the simple regulator may be loose and lead to overly conservative admission control decisions. For many source types (e.g., for VBR video), it is possible to get a tighter bound on the traffic and dramatically increase the admission region. In particular, regulator functions of the form

$$\mathcal{E}_j(t) = \min\{\rho_j^1 t,\ \sigma_j^2 + \rho_j^2 t, \ldots,\ \sigma_j^{L_j} + \rho_j^{L_j} t\} \tag{2}$$

are easily implemented with cascaded leaky buckets; it is shown in (see [28]) that the additional leaky buckets can lead to substantially larger admission regions for deterministic multiplexing. We

shall show that this is also true for statistical multiplexing. Throughout this paper we assume that each regulator has the form (2). Without loss of generality we may assume that $\rho_j^1 > \rho_j^2 > \cdots > \rho_j^{L_j}$ and $\sigma_j^2 < \sigma_j^3 < \cdots < \sigma_j^{L_j}$. For ease of notation, we set $\rho_j = \rho_j^{L_j}$. Note that for connection–$j$ traffic, the long–run average rate is no greater than $\rho_j$ and the peak rate is never greater than $\rho_j^1$.

Each connection also has a QoS requirement. In this paper we consider a QoS requirement that is particularly appropriate for multimedia traffic, such as audio and video traffic. Specifically, each connection has a connection–specific delay limit and a connection–specific loss bound. Denote $d_j$ and $\epsilon_j$ for the delay limit and loss bound for the $j$th connection. Any traffic that overflows at a buffer is considered to have infinite delay, and therefore violates the delay limit. The QoS requirement is as follows: for each connection $j$, the long–run fraction of traffic that is delayed by more than $d_j$ seconds must be less than $\epsilon_j$.

This QoS requirement can assure continuous, uninterrupted playback for a multimedia connection as follows. Each bit (or packet for packetized traffic) is time–stamped at the source. If a bit from connection $j$ is time–stamped with value $x$, the bit (if not lost in the node) arrives at the receiver no later than $x + d_j$. The receiver delays playout of the bit until time $x + d_j$. Thus, by including a buffer at each receiver, the receiver can playback a multimedia stream without jitter with a fixed delay of $d_j$ and with bit loss probability of at most $\epsilon_j$.

The strategy that we take in this paper is to pass each connection's traffic through a smoother at the connection's input to the node. We design the smoother for the $j$th connection so that the $j$th connection's traffic is *never* delayed at the smoother by more than $d_j$. After having smoothed a connection's traffic, we pass the smoothed traffic to the node. At the link the connection's traffic is multiplexed with traffic from other connections. The second aspect of our strategy is to remove all of the buffers in the node; that is, we use bufferless statistical multiplexing rather than buffered multiplexing before the link. In our fluid model, a connection's traffic that arrives to a bufferless link either flows through the link without any delay or overflows at the link, and therefore has infinite delay. In order to satisfy the $j$th connection's QoS requirement, it therefore suffices that the fraction of connection–$j$ traffic that overflows the link be less than $\epsilon_j$. Also, if the loss at the link is small, we can reasonably approximate a connection's traffic at the output of the multiplexer as being identical to its traffic at the input to the multiplexer. In other words, a connection that satisfies the regulator constraint $\mathcal{E}_j(t)$ at the input of the node satisfies the same regulator constraint $\mathcal{E}_j(t)$ at the output of the node. Our scheme extends therefore in a straightforward manner from a single node to a general network of bufferless multiplexers with smoothers at the network ingresses [23]. Our approach is illustrated in Figure 1.

For the smoother at the $j$th connection's input, initially we use a buffer which serves the traffic at rate $c_j^*$. When the smoother buffer is nonempty, traffic is drained from the smoother at rate $c_j^*$. When the smoother buffer is empty and connection–$j$'s traffic is arriving at a rate less than $c_j^*$, traffic leaves the smoother exactly at the rate at which it enters the buffer. For the fluid model and QoS criterion of this paper we shall show that more complex smoothers consisting of cascaded leaky buckets do not improve performance.

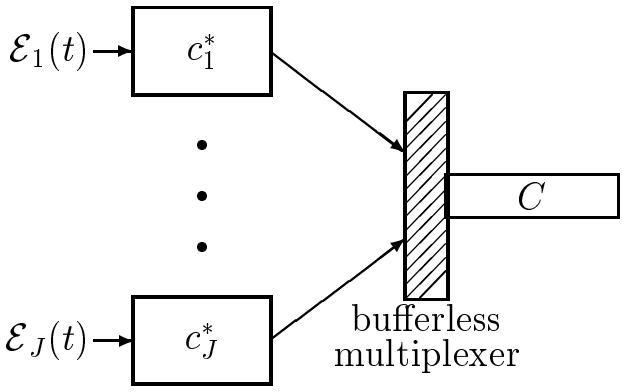



Figure 1: The traffic of the $j$th connections is characterized by the regulator function $\mathcal{E}_j(t)$. The traffic is passed through a smoother with rate $c_j^*$ and then multiplexed onto a link with capacity $C$.

Using the theory developed in [2], it can be shown that the maximum delay at the smoother is

$$\max_{t\geq 0}\left\{\frac{\mathcal{E}_j(t)}{c_j^*}-t\right\}. \tag{3}$$

Also, because the bufferless multiplexer and link introduce no delays, traffic from the $j$th connection that flows through the node without loss has the maximum delay of the smoother. We set the smoother rate to

$$c_j^* = \min\left\{c_j \geq 0 : \max_{t\geq 0}\left\{\frac{\mathcal{E}_j(t)}{c_j}-t\right\}\leq d_j\right\}\ , \tag{4}$$

so that the traffic that passes through the node (i.e., traffic which does not overflow at the link) is not be delayed by more than $d_j$. It is straightforward to show from (4) that the smoother rate can be expressed as

$$c_j^* = \max_{t\geq 0}\frac{\mathcal{E}_j(t)}{d_j+t}. \tag{5}$$

## 3 Guaranteeing Statistical QoS

We focus in this paper on a single link with $J$ connections. Connection $j$ has a regulator constraint function $\mathcal{E}_j(t)$ and QoS parameters $d_j$ and $\epsilon_j$. Now regard the $j$th arrival process as a stochastic process. Let $(A_j(t),\ t\geq 0)$ denote the $j$th arrival process, and let $(A_j(t,\omega),\ t\geq 0)$ denote a realization of the stochastic process. Also let $\mathbf{A}(t)=(A_1(t),\ldots,A_J(t))$, and let $(\mathbf{A}(t),\ t\geq 0)$ be the associated vector stochastic arrival process. We say that the vector arrival process $(\mathbf{A}(t),\ t\geq 0)$ is *feasible* if $(i)$ the component arrival processes $(A_j(t),\ t\geq 0)$, $j=1,\ldots,J$, are independent, and $(ii)$ for each $j=1,\ldots,J$, each realization $(A_j(t,\omega),\ t\geq 0)$ satisfies the regulator constraint

$$A_j(t+\tau,\omega)-A_j(\tau,\omega)\leq \mathcal{E}_j(t) \text{ for all } \tau\geq 0,\quad t\geq 0. \tag{6}$$

Denote $\mathcal{A}$ for the set of all feasible vector arrival processes $(\mathbf{A}(t),\ t\geq 0)$.

Our first goal is to develop a straightforward procedure to determine whether the QoS requirements are met for all possible feasible stochastic arrival processes. For a fixed feasible vector arrival process $(\mathbf{A}(t),\ t \geq 0)$, let $U_j(t)$ be the rate at which traffic from the $j$th connection leaves the associated smoother at time $t$, and let $U_j$ be the corresponding steady–state random variable. Consider multiplexing $(\mathbf{A}(t),\ t \geq 0)$ traffic streams onto a bufferless multiplexer of rate $C$. The long–run average fraction of traffic lost by connection $j$ is

$$P_{\text{loss}}^{\text{info}}(j) = \frac{E\left[(\sum_{k=1}^{J} U_k - C)^+ \frac{U_j}{\sum_{k=1}^{J} U_k}\right]}{E[U_j]}. \tag{7}$$

In the definition of $P_{\text{loss}}^{\text{info}}(j)$ we make the natural assumption that traffic loss at the bufferless multiplexer is split between the sources in a manner proportional to the rate at which the sources send traffic into the multiplexer. Note that $P_{\text{loss}}^{\text{info}}(j)$ keeps track of loss for each individual connection.

Although $P_{\text{loss}}^{\text{info}}(j)$ is an appealing performance measure, we have found it to be mathematically unwieldy. Instead of $P_{\text{loss}}^{\text{info}}(j)$ we shall work with a bound on $P_{\text{loss}}^{\text{info}}(j)$ which is more tractable and which preserves the essential characteristics of the original performance measure. Noting that the term in the expectation of the numerator of equation (7) is non–zero only when $\sum_{k=1}^{J} U_k > C$, we obtain:

$$P_{\text{loss}}^{\text{info}}(j) \leq \frac{E\left[(\sum_{k=1}^{J} U_k - C)^+ U_j\right]}{C \cdot E[U_j]} := P_{\text{loss}}(j). \tag{8}$$

In most practical circumstances the QoS requirement specifies traffic loss to be miniscule, on the order of $\epsilon_j = 10^{-6}$ or less. Thus we expect the bound to be very tight: during the rare event when $\sum_{j=1}^{J} U_j$ exceeds $C$, we expect $\sum_{j=1}^{J} U_j$ to be very close to $C$. Henceforth, we focus on the bound $P_{\text{loss}}(j)$, and we refer to $P_{\text{loss}}(j)$ as the *loss probability for the $j$th connection* . In Section 5 we provide numerical results which show that $P_{\text{loss}}(j)$ is very nearly equal to the actual loss probability $P_{\text{loss}}^{\text{info}}(j)$.

By taking the supremum over all the feasible vector stochastic processes, we obtain the following worst–case loss probability of the $j$th connection:

$$\phi_j^* = \sup_{\mathcal{A}} \frac{E\left[(\sum_{k=1}^{J} U_k - C)^+ U_j\right]}{C \cdot E[U_j]} \tag{9}$$

If $\phi_j^* \leq \epsilon_j$ for all $j = 1, \ldots, J$, then the QoS requirements are guaranteed to be met for all feasible vector arrival processes, that is, for all independent arrival processes whose sample paths satisfy the regulator constraints. In our strategy, at connection admission we determine whether $\phi_j^* \leq \epsilon_j$ for all $j = 1, \ldots, J$ will continue to hold when adding the new connection. If not, the connection is rejected. Thus, we need to develop an efficient method to compute the bounds $\phi_1^*, \ldots, \phi_J^*$. As a first step in computing these bounds, we need to explicitly determine the random variables $U_1, \ldots, U_J$ that attain the supremum in (9).

**Lemma 1** *Let $U_1^*, \ldots, U_J^*$ be independent random variables, with $U_j^*$ having distribution*

$$U_j^* = \begin{cases} c_j^* & \textit{with probability } \frac{\rho_j}{c_j^*} \\ 0 & \textit{with probability } 1 - \frac{\rho_j}{c_j^*} \end{cases}.$$

*There exists a feasible vector arrival process which produces the steady–state rate variables $U_1^*, \ldots, U_J^*$ at the smoother outputs.*

**Proof.** The proof is by construction. For each $j = 1, \ldots, J$, let

$$t_j = \frac{\sigma_j^2}{\rho_j^1 - \rho_j^2}$$

and

$$T_j = \frac{\rho_j^1 \sigma_j^2}{(\rho_j^1 - \rho_j^2)\rho_j}.$$

Also let $\theta_1, \ldots, \theta_J$ be independent random variables with $\theta_j$ uniformly distributed over $[0, T_j]$. Let $b_j(t)$ be a deterministic periodic function with period $T_j$ such that

$$b_j(t) = \begin{cases} \rho_j^1 & 0 \le t < t_j \\ 0 & t_j \le t \le T_j \ . \end{cases}$$

Define the $j$th arrival stochastic process as

$$A_j(t) = \int_0^t b_j(s + \theta_j) ds.$$

Thus each component arrival process $(A_j(t),\ t \ge 0)$ is generated by a periodic on–off source; the $j$th process has peak rate $\rho_j^1$ and average rate $\rho_j$. By sending each component process $(A_j(t),\ t \ge 0)$ into its respective smoother, we obtain an on–off process whose peak rate is $c_j^*$ and whose average rate is $\rho_j$. Also, the component processes are independent; thus the vector arrival process produces the steady–state random variables $U_1^*, \ldots, U_J^*$ at the smoother outputs.

It remains to show that each realization of $(A_j(t),\ t \ge 0)$ satisfies the regulator constraint (6). It follows immediately from the definition of $b_j(t)$ that

$$\int_0^t b_j(s) ds \le \mathcal{E}_j(t) \text{ for all } 0 \le t \le T_j. \tag{10}$$

We can, in fact, show that

$$\int_0^t b_j(s) ds \le \mathcal{E}_j(t) \text{ for all } t \ge 0. \tag{11}$$

To see this consider any arbitrary $t = nT_j + s$, where $n$ is some non–negative integer and $0 \le s \le T_j$. We have

$$\begin{aligned} \int_0^t b_j(s) ds &= \int_0^{T_j} b_j(s) ds + \ldots + \int_{(n-1)T_j}^{nT_j} b_j(s) ds + \int_{nT_j}^{nT_j + s} b_j(s) ds \\ &\le nT_j \rho_j + \mathcal{E}_j(s) \\ &\le [\mathcal{E}_j(nT_j + s) - \mathcal{E}_j(s)] + \mathcal{E}_j(s) \\ &= \mathcal{E}_j(t) \ . \end{aligned}$$

The first inequality follows from (10) and from the fact that the average rate of $b_j(t)$ over any period of length $T_j$ is $\rho_j$. The second inequality follows because the slope of $\mathcal{E}_j(t)$ is never less than $\rho_j$. This establishes (11). Finally because $b_j(t)$ is non–increasing over each of its periods, we have

$$\int_\tau^{t+\tau} b_j(s) ds \le \int_0^t b_j(s) ds \text{ for all } \tau \ge 0, \quad t \ge 0. \tag{12}$$

Combining (11) and (12) proves that each realization of $(A_j(t),\ t \geq 0)$ satisfies the regulator constraint (6).

■

We now show that the random variables $U_1^*, \ldots, U_J^*$ attain the supremum in (9). This result will lead to a simple procedure for calculating the worst–case loss probabilities $\phi_1^*, \ldots, \phi_J^*$. To this end, we will need to make use of a concept from stochastic ordering. A random variable $X$ is said to be smaller than a random variable $Y$ in the sense of the *increasing convex stochastic (ics) ordering*, written as $X \leq_{icx} Y$, if $E[h(X)] \leq E[h(Y)]$ for all increasing, convex functions $h(\cdot)$.

**Theorem 1** *For each $j = 1, 2, \ldots, J$, worst–case loss probability for the $j$th connection is*

$$\phi_j^* = \frac{E\left[(\sum_{k=1}^J U_k^* - C)^+ U_j^*\right]}{C \cdot E[U_j^*]}$$

**Proof.** Let $\mathcal{U}$ be the set of all random vectors $(U_1, \ldots U_J)$ such that

1. $U_j$, $j = 1, 2, \ldots, J$ are independent.
2. $0 \leq E[U_j] \leq \rho_j$ and $0 \leq U_j \leq c_j^*$ for all $j = 1, 2, , \ldots, J$.

All feasible vector arrival processes in $\mathcal{A}$ give steady–state rate variables that belong to $\mathcal{U}$. Let $(U_1, \ldots, U_J)$ be a random vector in $\mathcal{U}$. Let $U = U_1 + \cdots + U_J$ and $U^* = U_1^* + \cdots + U_J^*$. We need to show that

$$\frac{E[(U - C)^+ U_j]}{CE[U_j]} \leq \frac{E[(U^* - C)^+ U_j^*]}{CE[U_j^*]}. \tag{13}$$

Fix $i$, with $1 \leq i \leq J$, and consider the random vector $(\hat{U}_1, \ldots, \hat{U}_J)$ such that $\hat{U}_i = U_i^*$ and $\hat{U}_j = U_j$ for $j \neq i$. Note that $(\hat{U}_1, \ldots, \hat{U}_J) \in \mathcal{U}$. We first show that for each fixed $j$,

$$\frac{E[(U - C)^+ U_j]}{CE[U_j]} \leq \frac{E[(\hat{U} - C)^+ \hat{U}_j]}{CE[\hat{U}_j]}. \tag{14}$$

Consider the case $i \neq j$. Let $V = U - U_i - U_j$. Let $dF_V(\cdot)$ and $dF_{U_j}(\cdot)$ be the distribution functions for $V$ and $U_j$. Noting that $U_i$, $U_j$ and $V$ are independent, we have

$$\begin{aligned} E[(U - C)^+ U_j] &= E[(U_i + V + U_j - C)^+ U_j] \\ &= \int_0^\infty \int_0^\infty E[(U_i + v + u - C)^+ u] dF_V(v) dF_{U_j}(u) \end{aligned}$$

The function $f(x) = (x + v + u - C)^+ u$ within the expectation is an increasing, convex function in $x$ for each fixed $v$ and $u$. Thus, because $U_i \leq_{icx} \hat{U}_i$ (e.g., see Proposition 1.5.1 in [27]), we have

$$E[(U_i + v + u - C)^+ u] \leq E[(\hat{U}_i + v + u - C)^+ u]$$

for all $v$ and $u$. Combining the above two equations gives

$$E[(U - C)^+ U_j] \leq E[(\hat{U} - C)^+ \hat{U}_j],$$

which, when combined with $E[\hat{U}_j] = E[U_j]$, gives (14).

Now consider the case $i = j$. Let $W = U - U_i$. Using $U_i \le c_i^*$, the independence of $W$ and $U_i$, and the independence of $W$ and $\hat{U}_i$, we obtain

$$\begin{aligned}
\frac{E[(U-C)^+ U_i]}{CE[U_i]} &= \frac{E[(W+U_i-C)^+ U_i]}{CE[U_i]} \\
&\le \frac{E[(W+c_i^*-C)^+]}{C}\frac{E[U_i]}{E[U_i]} \\
&= \frac{E[(W+c_i^*-C)^+]}{C}\frac{E[\hat{U}_i]}{E[\hat{U}_i]} \\
&= \frac{E[(W+c_i^*-C)^+ \hat{U}_i]}{CE[\hat{U}_i]}.
\end{aligned}$$

Also

$$\begin{aligned}
E[(\hat{U}-C)^+ \hat{U}_i] &= E[(W+\hat{U}_i-C)^+ \hat{U}_i] \\
&= E[(W+c_i^*-C)^+ \hat{U}_i].
\end{aligned}$$

Combining the above two equations gives (14) for $i = j$.

Thus (14) holds for all $i = 1, \ldots, J$. Therefore, starting with the original vector $(U_1, \ldots, U_J) \in \mathcal{U}$ we can replace $U_1$ with $U_1^*$ and obtain a new vector in $\mathcal{U}$ such that (14) holds. Rename this new vector as $(U_1, \ldots, U_J)$. We can repeat the procedure, this time replacing $U_2$ with $U_2^*$, and again obtaining a new vector in $\mathcal{U}$ such that (14) holds. Performing this procedure for all $i = 1, \ldots, J$ gives (13). ■

Using the fact that $U_j^*$ is a Bernoulli random variable, we obtain from Theorem 1 the following expression for the bound of $P_{\text{loss}}(j)$:

$$\phi_j^* = \frac{E\left[(\sum_{k\neq j} U_k^* + c_j^* - C)^+\right]}{C} \tag{15}$$

We can compute these bounds directly by convolving the distributions of the independent random variables. An efficient approximate convolution algorithm is presented in [16]. We can also obtain an accurate approximation for the right–hand side of (15) by applying large deviation theory to the expectation in the numerator: To this end let

$$\mu_{U_k^*}(s) := \ln E[e^{sU_k^*}].$$

Note that $\mu_{U_k^*}(s)$ is the logarithm of the moment generating function for $U_k^*$. We define

$$U^* = \sum_{k\neq j} U_k^*.$$

Note that

$$\mu_{U^*}(s) = \sum_{k\neq j} \mu_{U_k^*}(s)$$

by the independence of the $U_k^*$'s. The large deviation (LD) approximation gives the following approximation for $\phi_j^*$ [24, p. 146]

$$\frac{1}{Cs^{\star 2}\sqrt{2\pi\mu''_{U^*}(s^\star)}}e^{-s^\star(C-c_j^*)+\mu_{U^*}(s^\star)},$$

where $s^\star$ is the unique solution to

$$\mu'_{U^*}(s^\star) = C - c_j^*.$$

The LD approximation is known to be very accurate [24, 11, 7, 8, 22] and is also computationally very efficient. We use the LD approximation for the numerical studies in this paper.

In summary, (15) is a simple expression for the worst–case loss probability $\phi_j^*$; this simple expression involves the independent Bernoulli random variables $U_1^*, \ldots, U_J^*$, whose distributions we know explicitly. The LD approximation for (15) is highly accurate and is easily calculated. For admission control, we advocate using the LD approximation to calculate $\phi_j^*$ and then verifying the QoS requirement, i.e., verifying in real–time whether $\phi_j^* \leq \epsilon_j$ for all $j = 1, \ldots, J$.

At this juncture we note some important related work by Doshi [5, 6]. He studies worst–case, unsmoothed traffic that maximizes an aggregate loss ratio, where the aggregation is taken over all sources. For this criterion he discovers a number of anomalies; in particular, extremal on–off sources are not always worst case. With our bound $P_{\text{loss}}(j)$ (8) the loss is maximized by the extremal on–off sources, which greatly simplifies admission control. Furthermore, as we show in this paper, smoothing of traffic can significantly expand the admission region.

### 3.1 The Optimal Smoother

Up to this point we have assumed that the smoother for each connection $j$ consists of a single buffer that limits the peak rate of the smoother output to $c_j^*$. In this subsection we study more general smoothers, namely, smoothers that consist of a cascade of leaky buckets. The smoother for connection $j$, defined by a function $S_j(t)$, constrains the amount of traffic that can enter the network over any time interval. Specifically, if $B_j(t)$ is the amount of traffic leaving smoother $j$ over the interval $[0, t]$, then $B_j(t)$ is required to satisfy

$$B_j(t+\tau) - B_j(\tau) \leq S_j(t) \quad \text{for all } t \geq 0,\ \tau \geq 0.$$

We assume throughout this section that the smoother functions are of the form

$$S_j(t) = \min_{1 \leq k \leq M_j} \{s_j^k + r_j^k t\} \tag{16}$$

with $r_j^1 > r_j^2 > \cdots > r_j^{M_j}$ and $0 = s_j^1 < s_j^2 < \cdots < s_j^{M_j}$. These piecewise linear, concave smoother functions can be easily implemented by a cascade of leaky buckets. The single–buffer smoother defined in Section 2 is a special case with $M_j = 1,\ s_j^1 = 0$ and $r_j^1 = c_j^*$.

We say that a set of smoothers $(S_1(t), ..., S_J(t))$ is *feasible* if the maximum delay incurred at smoother $j$ is $\leq d_j$ for all $j = 1, \ldots, J$. By definition the set of smoothers $(c_1^* t, \ldots, c_J^* t)$ studied

earlier is feasible. Now fix a feasible set of smoothers $(S_1(t), \ldots, S_J(t))$, and let the regulated traffic from the $J$ connections pass through these smoothers. Let

$$\phi_j = \sup_{\mathcal{A}} \frac{E\left[(\sum_{k=1}^{J} U_k - C)^+ U_j\right]}{C \cdot E[U_j]} \tag{17}$$

be the associated worst–case loss probability. Recall that $\phi_j^*$ is the same worst–case loss probability but with the traffic passing through the set of smoothers $(c_1^* t, \ldots, c_J^* t)$. The proof of the following result is provided in the appendix.

**Theorem 2** *$\phi_j^* \leq \phi_j$ for all $j = 1, \ldots, J$. Thus the single–buffer smoothers with $c_j = c_j^*$ minimize the worst–case loss probability over all feasible sets of smoothers.*

It follows from Theorem 2 that the more complex smoothers consisting of cascaded leaky buckets do not increase the connection carrying capacity of the network. Thus without loss of performance, we may use the simple smoothers of the form $(c_1 t, \ldots, c_J t)$. Furthermore, Theorem 2 verifies the intuition that in order to maximize the admission region the smoother rates are as small as the delay constraints permit, that is, $c_j = c_j^*$ for $j = 1, \ldots, J$.

### 3.2 A Heuristic for Finding a Leaky Bucket Characterization of Prerecorded Sources

In this subsection we discuss how to obtain a good characterization $\mathcal{E}_j(t)$ of a source for a given restriction $L_j$ on the number of leaky buckets. For any given characterization $\mathcal{E}_j(t)$ we use at the network edge a single–buffer smoother with rate $c_j^*$ given by (5). Our goal is to find a characterization $\mathcal{E}_j(t)$ that has at most $L_j$ slopes (i.e., $L_j$ cascaded leaky buckets) and attempts to minimize both $\rho_j$ and $c_j^*$. From Theorem 2 we know that minimizing $\rho_j$ and $c_j^*$ minimizes the worst–case loss probabilities, and thereby maximizes the connection–carrying capacity of the network.

We develop the heuristic for determining the characterization $\mathcal{E}_j(t)$ in the context of prerecorded sources. These sources include full–length movies, music video clips and educational material for video–on–demand (VoD) and other multimedia applications. It is well known how to compute the empirical envelope for prerecorded sources [12, 28, 17]. The empirical envelope gives the tightest bound on the amount of traffic that can emanate from a prerecorded source over any time interval. The empirical envelope is however not necessarily concave, and therefore we may not be able to be characterize it by a cascade of leaky buckets. However, applying the algorithms of Knightly *et al.* [28] or Grahams Scan [1], we can compute the concave hull of the empirical envelope. The concave hull for connection–$j$ traffic, denoted by $\mathcal{H}_j(t)$, takes the form

$$\mathcal{H}_j(t) = \min_{1 \leq i \leq K_j} \{\sigma_j^i + \rho_j^i t\}. \tag{18}$$

Here, $K_j$ denotes the number of piecewise linear segments in the concave hull. Without loss of generality we may assume $\sigma_j^1 < \sigma_j^2 < \cdots < \sigma_j^{K_j}$ and $\rho_j^1 > \rho_j^2 > \cdots > \rho_j^{K_j}$.

The number of segments in the concave hull can be rather large. The *"Silence of The Lambs"* video segment used in our numerical experiments, for instance, has a concave hull consisting of

39 segments. This implies that 39 leaky bucket pairs are required to police the tightest concave characterization of the *"Silence of The Lambs"* video segment. Our goal is to find a more succinct characterization of prerecorded sources in order to simplify call admission control and traffic policing.

Suppose that a source is allowed to use $L_j$ ($L_j < K_j$) leaky buckets to characterize its traffic. We now present a heuristic for the following problem: Given a source's concave hull $\mathcal{H}_j(t) = \min_{1\le i\le K_j}\{\sigma_j^i + \rho_j^i t\}$ and the delay limit $d_j$, find $L_j$ leaky buckets (out of the $K_j$ leaky bucket pairs in the concave hull) that maximize the admission region.

We illustrate our heuristic for the case $L_j = 2$. For $L_j = 2$ the traffic constraint function takes the form

$$\mathcal{E}_j(t) = \min\{\sigma_j^{a_j} + \rho_j^{a_j} t,\ \sigma_j^{b_j} + \rho_j^{b_j} t\} \ \text{ with } \ 1 \le a_j, b_j \le K_j, \tag{19}$$

where the indices $a_j$ and $b_j$ are yet to be specified. Our strategy is to first choose the leaky bucket that has the tightest bound on the average rate (i.e., minimize $\rho_j$), and then choose another leaky bucket which minimizes the smoother rate $c_j^*$. Let $r_j^{\text{ave}}$ denote the average rate of the prerecorded source. We found in our numerical experiments that some of the leaky bucket pairs in the concave hull (particularly those with high indices) may have slopes $< r_j^{\text{ave}}$. We set $b_j = \max\{i : \rho_j^i \ge r_j^{\text{ave}},\ 1 \le i \le K_j\}$. In words, we use the highest indexed leaky bucket with a slope larger than $r_j^{\text{ave}}$ to specify the sources' average rate.

In order to find the leaky bucket indexed by $a_j$ we consider all leaky buckets $(\sigma_j^i,\ \rho_j^i)$ with $1 \le i < b_j$. We compute the smoother rates obtained by combining each of the leaky buckets $(\sigma_j^i,\ \rho_j^i),\ 1 \le i < b_j$ with the leaky bucket $(\sigma_j^{b_j}, \rho_j^{b_j})$ and select the index $i$ that gives the smallest smoother rate — and thus the largest admission region. More formally, let $c_j^{*i},\ 1 \le i < b_j$, denote the minimal smoother rate for traffic with regulator function $\mathcal{E}_j(t) = \min\{\sigma_j^i + \rho_j^i t,\ \sigma_j^{b_j} + \rho_j^{b_j} t\}$ and delay bound $d_j$. By (5) we have

$$c_j^{*i} \ = \ \max_{t\ge 0} \frac{\min\{\sigma_j^i + \rho_j^i t,\ \sigma_j^{b_j} + \rho_j^{b_j} t\}}{d_j + t}.$$

We can obtain a more explicit expression for $c_j^{*i}$. Since

$$\min\{\sigma_j^i + \rho_j^i t,\ \sigma_j^{b_j} + \rho_j^{b_j} t\} = \begin{cases} \sigma_j^i + \rho_j^i t & \text{for } 0 \le t \le t_i \\ \sigma_j^{b_j} + \rho_j^{b_j} t & \text{for } t \ge t_i \end{cases}$$

with $t_i = (\sigma_j^{b_j} - \sigma_j^i)/(\rho_j^i - \rho_j^{b_j})$, we have

$$c_j^{*i} \ = \ \max\left[\max_{0\le t\le t_i} \frac{\sigma_j^i + \rho_j^i t}{d_j + t},\ \max_{t\ge t_i} \frac{\sigma_j^{b_j} + \rho_j^{b_j} t}{d_j + t}\right].$$

The expressions inside the max[·] can be further simplified. It can be shown that

$$\max_{0\le t\le t_i} \frac{\sigma_j^i + \rho_j^i t}{d_j + t} = \begin{cases} \frac{\sigma_j^i}{d_j} & \text{if } d_j \le \frac{\sigma_j^i}{\rho_j^i} \\ \frac{\sigma_j^i + \rho_j^i t_i}{d_j + t_i}, & \text{if } d_j \ge \frac{\sigma_j^i}{\rho_j^i} \end{cases}$$

and

$$\max_{t \geq t_i} \frac{\sigma_j^{b_j} + \rho_j^{b_j} t}{d_j + t} = \begin{cases} \frac{\sigma_j^i + \rho_j^i t_i}{d_j + t_i}, & \text{if } d_j \leq \frac{\sigma_j^{b_j}}{\rho_j^{b_j}} \\ \frac{\sigma_j^{b_j}}{d_j} & \text{if } d_j \geq \frac{\sigma_j^{b_j}}{\rho_j^{b_j}}. \end{cases}$$

We set the smoother rate to $\min_{1 \leq i < b_j} c_j^{*i}$ and set $a_j$ to the index that attains this minimum.

We now briefly discuss how to find the optimal regulator function consisting of 3 or more leaky buckets. First, note that there are $\begin{pmatrix} b_j - 1 \\ L_j - 1 \end{pmatrix}$ combinations of leaky bucket pairs to consider. This can be computationally prohibitive. The heuristic can be sped up by considering only regulator functions consisting of $L_j - 1$ consecutive leaky buckets of the concave hull and the leaky bucket $(\sigma_j^{b_j} \ \rho_j^{b_j})$. In the case $L_j = 3$, for instance, we compute the minimal smoother rates only for the regulator functions $\mathcal{E}_j(t) = \min\{\sigma_j^i + \rho_j^i t, \ \sigma_j^{i+1} + \rho_j^{i+1} t, \ \sigma_j^{b_j} + \rho_j^{b_j} t\}$ with $1 \leq i < b_j - 1$. This speed–up of the heuristic can produce a suboptimal regulator function. Our numerical experiments (see Section 4), however, indicate that it works surprisingly well.

### 3.3 Interaction between Application and Network

In this subsection we discuss how the responsibilities of smoothing, call admission control and traffic policing can be shared by the application and the network. Call admission control is the responsibility of the network. Before accepting a new connection, the network has to ensure that the QoS requirements continue to hold for all established connections and the new connection. Policing is also a network responsibility. The network edge has to police all established connections in order to ensure that all connections comply with their respective regulator function advertised at connection establishment. While call admission control and traffic policing are responsibilities of the network, smoothing can be performed by either the application or the network. We refer to the case where the application performs the smoothing and sends the smoothed traffic to the network edge as *application smoothing*. The case where the application sends its unsmoothed traffic to the network edge and the network edge performs the smoothing is referred to as *network smoothing*.

With application smoothing the application internally smoothes its traffic. Based on the regulator function of its traffic and the maximum delay it can tolerate, the application finds the minimum smoother rate by applying (5). Since the smoothing is done by the application, there is no need to reduce the number of leaky buckets used to characterize the traffic by applying the heuristic outlined in Section 3.2. Instead, the concave hull of a prerecorded source is used directly for dimensioning its smoother. The application advertises the regulator function $\mathcal{E}_j(t) = \min\{c_j^* t, \ \sigma_j^{L_j} + \rho_j^{L_j} t\}$ and the delay bound $d_j = 0$ to the network. We remark that this dual leaky bucket regulator function has been adopted by the ATM Forum [9] and is being proposed for the Internet [26]. The network does not have to be aware of the smoothing done by the application. The network edge dimensions its own smoother based on $\mathcal{E}_j(t)$ and $d_j = 0$. Since $d_j = 0$ the networks' smoother degenerates to a server with rate $c_j^*$ preceded by a buffer of size zero.

With network smoothing the application advertises its regulator function and maximum tolerable delay to the network. Prerecorded sources apply the heuristic of Section 3.2 when the network

| Trace | Mean (bit) bits/frame | Mean kbits/sec | Peak/Mean |
|---|---|---|---|
| lambs | 7,312 | 171.2 | 18.4 |
| mr.bean | 17,647 | 423.5 | 13.0 |

Table 1: Statistics of MPEG–1 traces.

restricts the number of leaky buckets to a number smaller than the number of segments in the concave hull. The network edge dimensions the smoother based on the regulator function and delay bound supplied by the application. Call admission control is based on the assumption of worst–case on–off traffic at the smoother output. The network edge polices the applications' traffic before it enters the smoother and drops violating traffic.

# 4 Numerical Experiments

In this section we evaluate the smoothing/bufferless multiplexing scheme proposed in this paper using traces from MPEG encoded movies. In all experiments we consider a single bufferless multiplexer which feeds into a 45 Mbps link. We obtained the frame size traces, which give the number of bits in each video frame, from the public domain [25]. (We are aware that these are low resolution traces and some critical frames are dropped; nevertheless, the traces are extremely bursty.) The movies were compressed with the Group of Pictures (GOP) pattern IBBPBBPBBPBB at a frame rate of $F = 24$ frames/sec [25]. Each of the traces has $N = 40{,}000$ frames, corresponding to about 28 minutes. The mean number of bits per frame and the peak–to–mean ratio are given in Table 1. Let $x_n,\ n = 1, \ldots, N$, denote the size of the $n$th frame in bits. We convert the discrete frame size trace to a fluid flow by transmitting the $n$th frame at rate $x_n F$ over the interval $[n - 1/F, n/F]$.

We first evaluate the heuristic of Section 3.2. We compute the empirical envelope and the concave hull of each trace using the algorithms of Knightly *et al.* [28]. Based on the concave hull of each video we compute the minimal smoother rate $c_j^*$. We also apply the heuristic of Section 3.2 to the concave hull in order to find the optimal leaky bucket characterization with 2 and more leaky buckets. (We apply the speed–up described in Section 3.2 for the leaky bucket characterizations with 3 or more leaky buckets.)

The heuristic of Section 3.2 produced the optimal leaky bucket characterizations given in Table 2 for the lambs trace. The table gives the index $a_{\text{lambs}}$ and the parameters of the leaky bucket $(\sigma_{\text{lambs}}^{a_{\text{lambs}}},\ \rho_{\text{lambs}}^{a_{\text{lambs}}})$ for various delay bounds. The average rate is characterized by the 34th leaky bucket in the concave hull, i.e., $b_{\text{lambs}} = 34$, for all delay bounds. The table also gives the minimal smoother rates for the various delay bounds. For a delay bound of zero, the smoother rate is set to the rate of the first leaky bucket, i.e., the peak rate of the trace. For $d_{\text{lambs}} = 0.042$ sec $(= 1/F)$ the trace is characterized by the 2nd and 34th leaky bucket of the concave hull ($a_{\text{lambs}} = 2,\ b_{\text{lambs}} = 34$). Note that $d_{\text{lambs}} < \sigma_{\text{lambs}}^{a_{\text{lambs}}}/\rho_{\text{lambs}}^{a_{\text{lambs}}}$ in this case and $c_{\text{lambs}} = \sigma_{\text{lambs}}^{a_{\text{lambs}}}/d_{\text{lambs}}$. For $d_{\text{lambs}} \geq 0.125$ sec we have $d_{\text{lambs}} > \sigma_{\text{lambs}}^{a_{\text{lambs}}}/\rho_{\text{lambs}}^{a_{\text{lambs}}}$ and $c_{\text{lambs}}^* = (\sigma_j^{a_{\text{lambs}}} + \rho_j^{a_{\text{lambs}}} t_{a_{\text{lambs}}})/(d_j + t_{a_{\text{lambs}}})$.

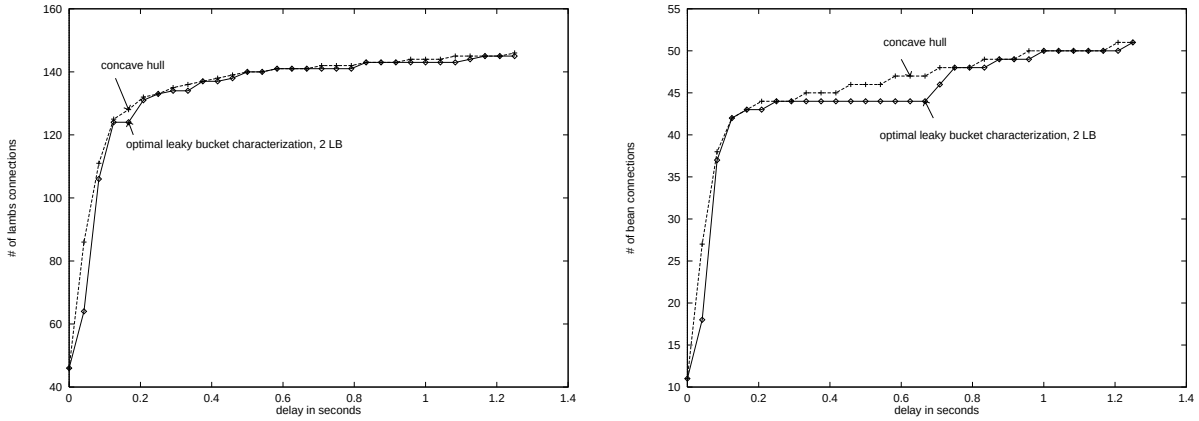
| $d_{\text{lambs}}$ sec. | $a_{\text{lambs}}$ | $\sigma_{\text{lambs}}^{a_{\text{lambs}}}$ kByte | $\rho_{\text{lambs}}^{a_{\text{lambs}}}$ kbit/sec | $c_{\text{lambs}}^{*}$ kbit/sec |
|---|---|---|---|---|
| 0 | 1 | 0 | 3474.8 | 3474.8 |
| 0.042 | 2 | 13.3 | 939.3 | 2535.5 |
| 0.125 | 2 | 13.3 | 939.3 | 939.0 |
| 0.250 | 4 | 23.5 | 802.2 | 801.9 |
| 0.500 | 8 | 43.5 | 711.0 | 710.8 |
| 1.000 | 10 | 69.9 | 676.9 | 674.7 |

Table 2: Parameters of the optimal leaky bucket characterization with 2 leaky buckets as a function of the delay bound for the lambs trace. The average rate is characterized by the 34th leaky bucket, i.e., $b_{\text{lambs}} = 34$, with parameters $\sigma_{\text{lambs}}^{b_{\text{lambs}}} = 3,157.8$ kByte and $\rho_{\text{lambs}}^{b_{\text{lambs}}} = 208.8$ kbit/sec for all delay bounds.

Assuming worst–case on–off traffic, the smoother outputs are statistically multiplexed onto the bufferless link as discussed in the previous sections. We set $\epsilon_j = 10^{-7}$ for all connections. In Figure 2 we plot the number of admissible video connections as a function of the delay bound. The graph gives the number of admissible video connections when the videos are characterized by the concave hull or the optimal leaky bucket characterization with 2 leaky buckets. We observe from the plots that the optimal leaky bucket characterization with 2 leaky buckets admits almost as many video connections as the more accurate concave hull characterization. The curves for 3 or more leaky buckets coincide with the curve for the concave hull.

In the next experiment we compare the admission region of our approach with the admission region obtained with the deterministic admission control condition of Knightly *et al.* [28]. Note that the deterministic approach of Knightly *et al.* is lossless and guarantees that no bit is delayed by more than the prespecified delay limit in the multiplexer buffer. Our approach, on the other hand, exploits the independence of traffic emanating from the $J$ connections. The videos are passed through simple smoothers with $c_j = c_j^*$. The smoother outputs — assuming worst–case on–off traffic — are then statistically multiplexed onto the bufferless link, as discussed in the preceding sections. We set $\epsilon_j = 10^{-7}$ for all connections. Losses this small have essentially no impact on the perceived video quality and can be easily hidden by error concealment techniques [19].

In Figure 3 we plot the number of admissible lambs connections as a function of the delay bound. The graph gives the number of lambs connections that are admitted with the our approach (RRR) when 2 or 3 leaky buckets (LB) are used to characterize the video trace. As we just saw in Figure 2 the optimal leaky bucket characterization with 3 leaky buckets admits as many connections as the concave hull, the most accurate, concave characterization of the video; using more leaky buckets does not increase the admission region. We also plot the number of lambs connections that are admitted with the approach of Knightly *et al.* (KLZ) when 2, 3, 8 or 16 leaky buckets are used to characterize the trace. We observe that for delays on the order of .5 seconds or more, the number




(a) lambs (b) bean

Figure 2: Number of video connections as a function of the delay bound. The videos are characterized by the concave hull or the optimal leaky bucket characterization with 2 leaky buckets. The bound on the loss probability is $10^{-7}$.

of admissible connections significantly increases as the number of leaky buckets used to describe the trace increases. The approach of Knightly *et al.* thus greatly benefits from a more accurate characterization of the video — achieved by more leaky buckets.

The main result of this experiment, however, is that our approach allows for more than twice the number connections than does the approach of Knightly *et al.* For example, for a delay bound of 1.1 seconds, Knightly *et al.* admit 69 connections ( = 29.6 % average link utilization) with 16 leaky buckets while our approach admits 146 connections ( = 62.7 % average link utilization) with 3 leaky buckets. We obtain this dramatic increase in the admission region by exploiting the independence of the sources and allowing for a small loss probability.

In Figure 4 we consider multiplexing two different movies, beans and lambs, each with its own delay constraint. We again assume a 45 Mbps link. We use delay bounds of $d_{\text{lambs}} = 125$ ms or 1.25 seconds and $d_{\text{bean}} = 125$ ms or 1.25 seconds, giving four combinations. Both videos are characterized by 3 leaky buckets. We assume that both video connections have the QoS requirement that the fraction of traffic that is delayed by more than the imposed delay limit is less than $10^{-7}$. For the Knightly *et al.* plot we use Earliest Deadline First (EDF) scheduling. We see that for all four cases, the admission region for our approach is dramatically larger.

In Figure 5 we compare the actual loss probability, $P_{\text{loss}}^{\text{info}}(j)$ given by (7) with our bound for loss probability, $P_{\text{loss}}(j)$, given by (8). We obtain $P_{\text{loss}}^{\text{info}}(j)$ and $P_{\text{loss}}(j)$ by simulation, and assume worst–case on-off traffic. We also verify the accuracy of the large deviation approximation for $P_{\text{loss}}(j)$. In Figure 5 we plot the loss probabilities as a function of the number of connections being multiplexed over a 45 Mbps link. We consider the scenario where the videos have a delay bound of 1 second and are characterized by 3 leaky buckets. We observe that the bound on the loss probability $P_{\text{loss}}(j)$ (solid line) tightly bounds the actual loss probability $P_{\text{loss}}^{\text{info}}(j)$ (dotted line). We also observe that

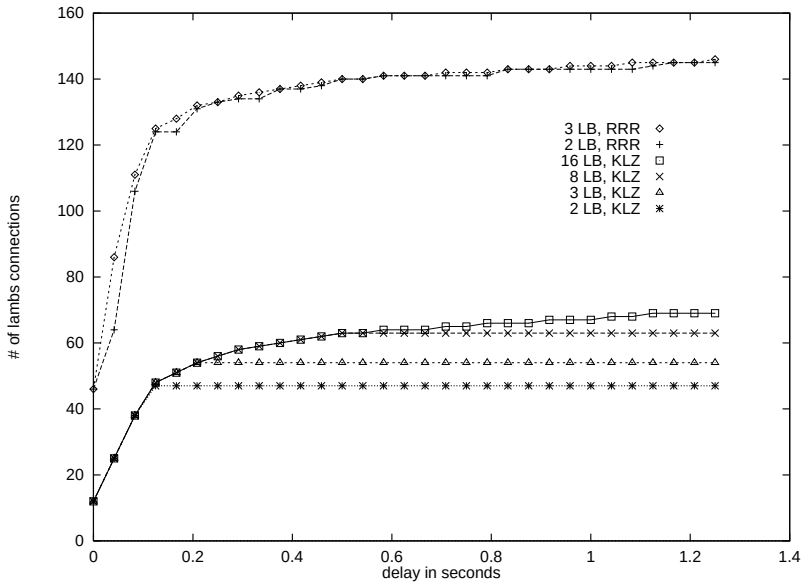



Figure 3: Number of lambs connections as a function of the delay bound and the number of leaky buckets (LB). Plots shown are for Knightly *et al.* (KLZ) [28] and our approach(RRR). The bound on the loss probability is $10^{-7}$.

the LD approximation (dashed line) closely approximates the simulation results.

# 5 Comparison with Buffered Statistical Multiplexing

The numerical results of the previous section show that our approach allows for dramatically more connections than buffered deterministic multiplexing. In this section we briefly consider buffered multiplexing with an allowance of small loss probabilities, which we refer to as *buffered statistical multiplexing*. Consider the buffered analogy of the single-link bufferless system studied in Section 3. The link has capacity $C$ and is preceded by a finite buffer of capacity $B$. Let the same $J$ connections arrive to this system; specifically the $J$ connections are independent and the $j$th connection is regulated by a given regulator function $\mathcal{E}_j(t)$. The traffic from the $J$ connections passes directly into the buffered multiplexer, i.e., the traffic is not pre–smoothed before arriving at the buffer. This buffered system is illustrated in Figure 6. Assuming that traffic is served FIFO, the maximum delay in this system is $d = B/C$. Suppose that the buffer overflow probability is constrained to be no greater than $\epsilon$.

It is a difficult and challenging problem to accurately characterize the admission region for a buffered multiplexer which multiplexes regulated traffic and which allows for statistical multiplexing. Elwalid *et al.* in [8] made significant progress in this direction. They consider the buffered multiplexer for the *special case of regulators with two leaky buckets*, i.e., for $\mathcal{E}_j(t) = \min\{\rho_j^1 t,\ \sigma_j + \rho_j t\}$. (In our numerical comparisons, we extend their theory to the case of multiple cascaded leaky buckets.) In order to make the buffered multiplexer mathematically tractable they assign each connection its own virtual buffer/trunk system. Each virtual buffer/trunk system is allocated buffer

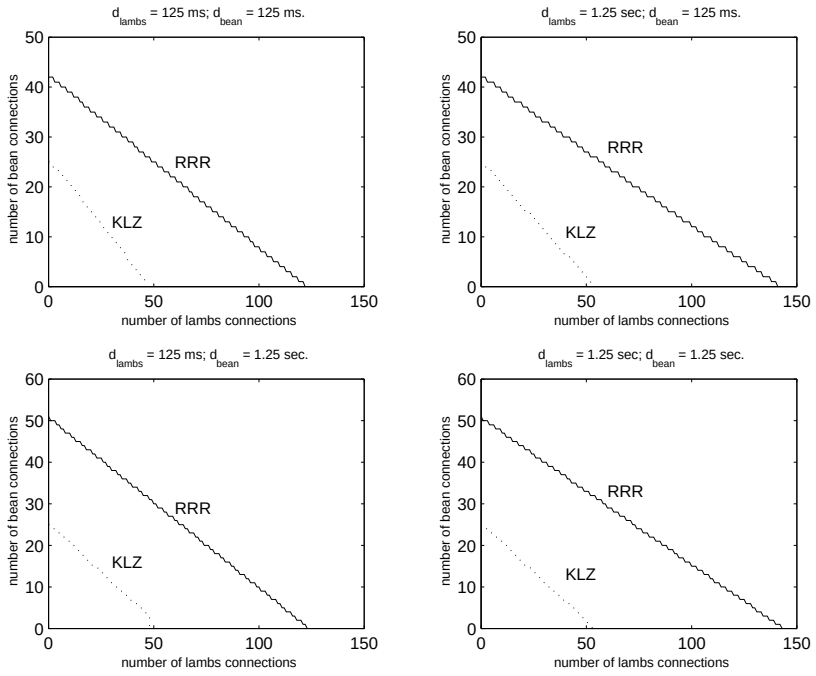



Figure 4: Admission region for the multiplexing of lambs and bean connections over a 45 Mbps link.

$b_{0,j}$ and bandwidth $e_{0,j}$. The allocations are based on the buffer and bandwidth resources ($B$ and $C$, respectively) and on the regulator parameters ($\rho_j$, $\rho_j^1$, and $\sigma_j$) for the input traffic. It turns out that the bandwidth $e_{0,j}$ is exactly the $c_j^*$ obtained by setting $d_j = d = B/C$ in (4). After some analysis Elwalid *et al.* obtain the following bound on the fraction of time during which loss occurs at the buffered multiplexer:

$$P_{\text{loss}}^{\text{EMW}} = P(U_1^* + \cdots + U_J^* \leq C).$$

where $U_1^*, \ldots, U_J^*$ are exactly the same random independent random variables that occur in Theorem 1. (To calculate the associated $c_1^*, \ldots, c_J^*$, set $d_j = d = B/C$ for each connection $j$.)

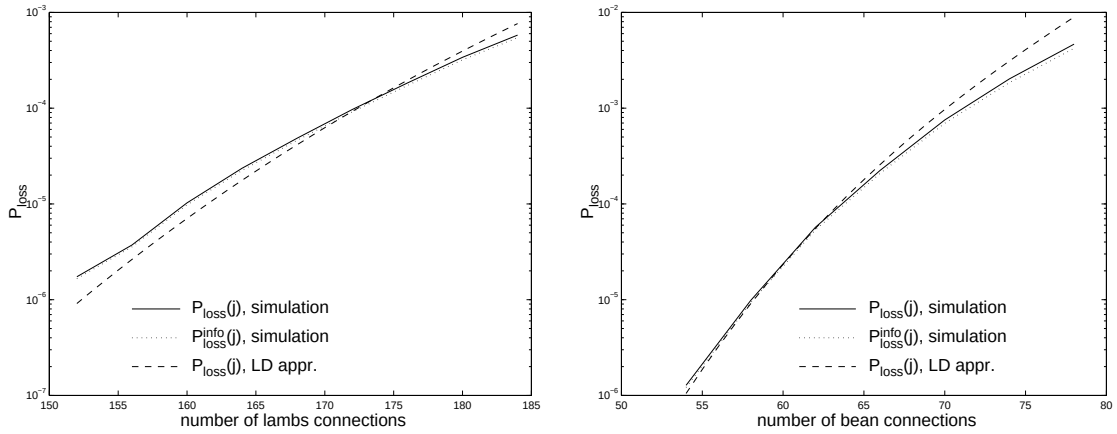
This observation indicates that the bufferless system of this paper has remarkably similarities with the buffered system in [8]. Specifically, for a fixed maximum delay $d$ in the buffered system, we can design a bufferless system with pre–smoothers which has the same maximum delay and which has an admission region based on the same set of independent random variables $U_1^*, \ldots, U_J^*$. The pre–smoothers essentially implement the virtual buffer/trunk systems introduced by Elwalid *et al.* For a maximum loss probability of $\epsilon$ the admission region for the buffered multiplexer is defined by

$$P(U_1^* + \cdots + U_J^* \leq C) \leq \epsilon$$

whereas the admission region for the bufferless system is

$$\frac{E[(\sum_{k=1}^J U_k^* - C)^+ U_j^*]}{C \cdot E[U_j^*]} \leq \epsilon.$$

Although these admission regions are different, they are based on exactly the same independent random variables $U_1^*, \ldots, U_J^*$. The difference in these admission regions is an artifact of using two




(a) lambs (b) bean

Figure 5: The simulation verifies that the bound on the loss probability $P_{\text{loss}}(j)$ tightly bounds the actual loss probability $P_{\text{loss}}^{\text{info}}(j)$. The plots further confirm the accuracy of the Large Deviation (LD) approximation. We use a delay bound of 1 second and characterize the videos by 3 leaky buckets. The link rate is 45 Mbps. The plots give the loss probability as a function of the number of ongoing connections.

different notions of loss probability: while in this paper we use "fraction of traffic lost", the paper [8] uses "the fraction of time during which loss occurs". If the same notions of loss were used, then the admission regions would be identical. Figure 7 gives the number of lambs connections that are admitted with the approach of Elwalid *et al.* (EMW) [8] and our approach (RRR) when 3 leaky buckets are used to characterize the trace. We assume a 45 Mbps link and set $\epsilon_j = 10^{-7}$ for all connections.

Thus, our bufferless system has essentially the same admission region as the buffered system in [8] for a fixed worst–case delay $d$ and loss probability $\epsilon$. While being no more difficult to perform call admission, we believe that the bufferless system has some important advantages over the buffered system: (*i*) no buffer is needed at the multiplexer (for packetized traffic, a relatively small buffer would be needed); (*ii*) the bufferless approach allows for a per–connection QoS requirement, whereas the buffered system imposes the same QoS requirement on all connections; and (*iii*), perhaps most importantly, networks are quite tractable for bufferless links, as we can reasonably approximate a connection's traffic at the output of the multiplexer as being identical to its traffic at the input to the multiplexer.

On the other hand, the buffered system does have some advantages over the bufferless system. First, although both systems have the same worst–case delay, the buffered system will have a lower average delay. Second, the admission region of [8] can be increased using the techniques in [18] and [21] (at the expense of a much more complicated admission procedure). Because multimedia applications are typically designed for a delay bound, and because the aforementioned increase in admission region is typically small, we feel that the advantages of the bufferless approach outweigh

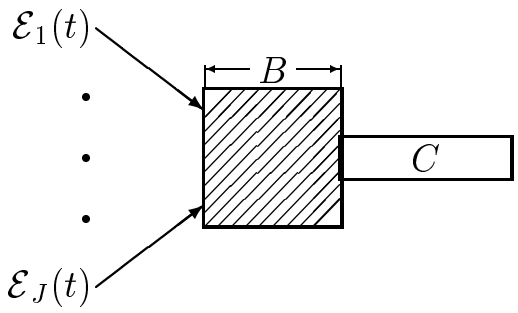



Figure 6: The traffic of connection $j$ is characterized by the regulator function $\mathcal{E}_j(t)$ and fed directly, i.e. unsmoothed, into a buffered multiplexer.

the advantages of the buffered approach.

# 6 Final Remarks

In this paper we have considered traffic management for multimedia networking applications which permit a small amount of loss and some bounded delay. We have argued that it is preferable to smooth the traffic at the ingress and to perform bufferless statistical multiplexing within the node than to use shared–buffer multiplexing. For our scheme we have determined the worst–case traffic and have outlined an admission control procedure based on the worst–case traffic. We have also explicitly characterized the optimal smoother.

As pointed out in Section 3.3 the smoothing can be performed by either the network (at the network edge) or by the applications themselves. If the applications perform the smoothing, then an application should smooth the traffic as much as permitted by the delay constraint, and the network should offer a service to the application which guarantees queueing–free delays (delays only due to propagation and nodal processing) and allows the application to specify a maximum tolerable loss rate. The network node should perform statistical multiplexing in order to maximize its connection–carrying capacity. To guarantee QoS, admission control should suppose that the traffic is adversarial to the extent permitted by the regulators and smoothers.

Throughout this paper we have studied a single–node network. A subsequent paper addresses how the scheme can be extended to more general networks [23].

**Acknowledgements:** We gratefully acknowledge interactions with Jim Roberts at the early stages of this research.

# Appendix

The purpose of this appendix is to provide a proof for Theorem 2. But first we need to establish two lemmas.

**Lemma 2** *A necessary condition for* $(S_1(t), \ldots, S_J(t))$ *to be feasible is* $r_j^1 \geq c_j^*$ *for all* $j = 1, \ldots, J$.

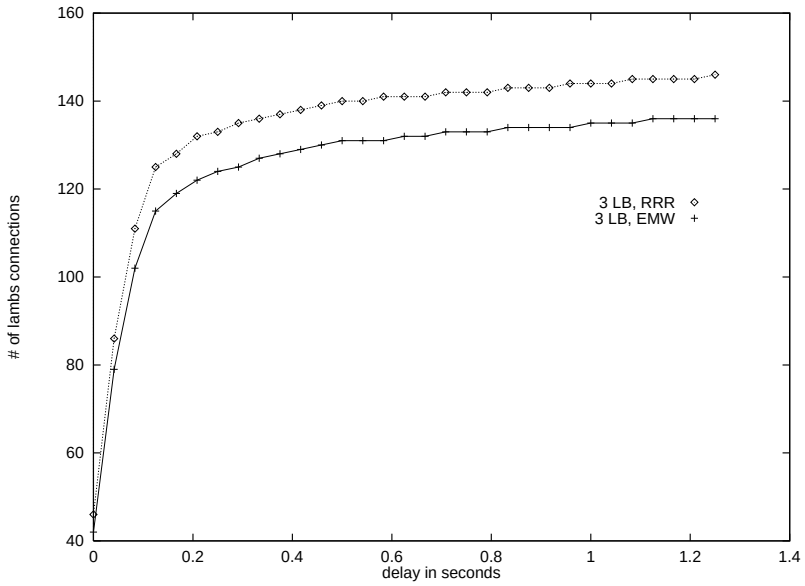



Figure 7: Number of lambs connections as a function of the delay bound. The lambs video is described by 3 leaky buckets. Plots shown are for Elwalid *et al.* (EMW) [8] and our approach(RRR). The difference in the number of admissible connections is due to the different notions of loss probability. Elwalid *et al.* use "fraction of time during which loss occurs" while we use "fraction of traffic lost".

**Proof.** From [3, 4, 10] the maximum delay at smoother $j$ is

$$\tilde{d}_j = \max_{t\geq 0}\{\max_{1\leq k\leq M_j} \frac{\mathcal{E}_j(t) - s_j^k}{r_j^k} - t\}. \tag{20}$$

Suppose $r_j^1 < c_j^*$ for some $j = 1, \ldots, J$. Because $s_j^k \geq 0$ and $r_j^k \leq r_j^1$ for all $k$, it follows from (20) that

$$\tilde{d}_j \geq \max_{t\geq 0}\{\frac{\mathcal{E}_j(t)}{r_j^1} - t\}. \tag{21}$$

And because, by assumption, $r_j^1 < c_j^*$, it follows from (21) that

$$\tilde{d}_j > \max_{t\geq 0}\{\frac{\mathcal{E}_j(t)}{c_i^*} - t\} = d_j,$$

where the last equality follows from (4). ■

**Lemma 3** *There exists a stochastic vector arrival process in $\mathcal{A}$ that produces the steady-state rate variables $\tilde{U}_1, \ldots, \tilde{U}_J$ with $\tilde{U}_j$ having distribution*

$$\tilde{U}_j = \begin{cases} \min(r_j^1, \rho_j^1) & \textit{with probability } \frac{\rho_j}{\min(r_j^1, \rho_j^1)} \\ 0 & \textit{with probability } 1 - \frac{\rho_j}{\min(r_j^1, \rho_j^1)} \end{cases}$$

*at the smoother outputs.*

**Proof.** For each $j = 1, \ldots, J$, let $t_j = \sigma_j^2/(\rho_j^1 - \rho_j^2)$ and $\delta_j = s_j^2/(r_j^1 - r_j^2)$. At $t = t_j$ the slope of $\mathcal{E}_j(t)$ changes form $\rho_j^1$ to $\rho_j^2 < \rho_j^1$. Consequently, $\mathcal{E}_j(t_j) = \rho_j^1 t_j$ is the maximum size burst that can be transmitted at rate

| | $\rho_j^1 \geq r_j^1$ | | $\rho_j^1 < r_j^1$ | |
|---|---|---|---|---|
| | $\mathcal{E}_j(t_j) \geq S_j(\delta_j)$ | $\mathcal{E}_j(t_j) < S_j(\delta_j)$ | $\mathcal{E}_j(t_j) \geq S_j(\delta_j)$ | $\mathcal{E}_j(t_j) < S_j(\delta_j)$ |
| $T_j$ | $S_j(\delta_j)/\rho_j$ | $\mathcal{E}_j(t_j)/\rho_j$ | $S_j(\delta_j)/\rho_j$ | $\mathcal{E}_j(t_j)/\rho_j$ |
| $t_{\text{on}_j}$ | $S_j(\delta_j)/\rho_j^1$ | $t_j$ | $S_j(\delta_j)/\rho_j^1$ | $t_j$ |
| $\tau_{\text{on}_j}$ | $\delta_j$ | $\mathcal{E}_j(t_j)/r_j^1$ | $S_j(\delta_j)/\rho_j^1$ | $t_j$ |

Table 3: On–times and periods of $\tilde{b}_j(t)$ and $\tilde{o}_j(t)$.

$\rho_j^1$, provided successive maximum size bursts are spaced at least $\mathcal{E}_j(t_j)/\rho_j - t_j$ apart. Similarly, at $t = \delta_j$ the slope of $S_j(t)$ changes form $r_j^1$ to $r_j^2 < r_j^1$. Consequently, $S_j(\delta_j) = r_j^1\delta_j$ is the maximum size burst the smoother can pass at rate $r_j^1$, provided successive maximum size bursts are spaced at least $S_j(\delta_j)/r_j^{M_j} - \delta_j$ apart.

Let $\tilde{b}_j(t)$ be a deterministic periodic function such that

$$\tilde{b}_j(t) = \begin{cases} \rho_j^1 & 0 \leq t < t_{\text{on}_j} \\ 0 & t_{\text{on}_j} \leq t \leq T_j \end{cases}.$$

with on–time $t_{\text{on}_j}$ and period $T_j$ given in Table 3. Also, let $\theta_1, \ldots, \theta_J$ be independent random variables with $\theta_j$ uniformly distributed over $[0, T_j]$ and define the $j$th stochastic arrival process as

$$\tilde{A}_j(t) = \int_0^t \tilde{b}_j(s + \theta_j)ds.$$

Thus each component arrival process $(\tilde{A}_j(t),\ t \geq 0)$ is generated by a periodic on-off source; the $j$th process has peak rate $\rho_j^1$ and average rate $\rho_j$. The argument in the proof of Theorem 1 shows that the vector process $(\tilde{\mathbf{A}}(t),\ t \geq 0)$ is a feasible process in $\mathcal{A}$.

It remains to show that by sending each component process $(\tilde{A}_j(t),\ t \geq 0)$ into its respective smoother we obtain an on-off process whose peak rate is $\min(r_j^1, \rho_j^1)$ and whose average rate is $\rho_j$. Specifically, we now show that $\tilde{A}_j(t)$ produces $\tilde{O}_j(t) = \int_0^t \tilde{o}_j(s + \theta_j)ds$ at the smoother output where

$$\tilde{o}_j(t) = \begin{cases} \min(r_j^1, \rho_j^1) & 0 \leq t < \tau_{\text{on}_j} \\ 0 & \tau_{\text{on}_j} \leq t \leq T_j \end{cases},$$

where the periods and on–times are given in Table 3.

First, consider the case $\rho_j^1 \geq r_j^1$ and $\mathcal{E}_j(t_j) \geq S_j(\delta_j)$. Clearly, $t_{\text{on}_j} \leq t_j$ since $t_{\text{on}_j} = S_j(\delta_j)/\rho_j^1$ and $t_j = \mathcal{E}_j(t_j)/\rho_j^1$ and by assumption $S_j(\delta_j) \leq \mathcal{E}_j(t_j)$. This implies that $\mathcal{E}_j(t_{\text{on}_j}) = \rho_j^1 t_{\text{on}_j}$. Hence

$$S_j(\tau_{\text{on}_j}) = \mathcal{E}_j(t_{\text{on}_j}). \tag{22}$$

Note furthermore that

$$t_{\text{on}_j} \leq \tau_{\text{on}_j} \tag{23}$$

since $t_{\text{on}_j} = S_j(\delta_j)/\rho_j^1 = r_j^1\delta_j/\rho_j^1$ and by assumption $r_j^1 \leq \rho_j^1$. Because of (22) and (23) and $\tau_{\text{on}_j} = \delta_j$ the smoother bursts at rate $r_j^1$ for a duration of $\tau_{\text{on}_j}$ when fed with an input burst at rate $\rho_j^1$ for a duration of $t_{\text{on}_j} \leq t_j$. Also, note that the smoother output has average rate $\mathcal{E}_j(t_{\text{on}_j})/T_j = \rho_j \leq r_j^{M_j}$, where the last inequality follows from the stability condition.

Next, consider the case $\rho_j^1 \geq r_j^1$ and $\mathcal{E}_j(t_j) < S_j(\delta_j)$. We have

$$\tau_{\text{on}_j} \leq \delta_j \tag{24}$$

since $\tau_{\text{on}_j} = \mathcal{E}_j(t_j)/r_j^1$ and $\delta_j = S_j(\delta_j)/r_j^1$ and by assumption $S_j(\delta_j) > \mathcal{E}_j(t_j)$. Thus $S_j(\tau_{\text{on}_j}) = r_j^1 \tau_{\text{on}_j}$. Hence

$$S_j(\tau_{\text{on}_j}) = \mathcal{E}_j(t_{\text{on}_j}). \tag{25}$$

Also,

$$t_{\text{on}_j} \leq \tau_{\text{on}_j} \tag{26}$$

since $t_{\text{on}_j} = \mathcal{E}_j(t_j)/\rho_j^1$ and $\tau_{\text{on}_j} = \mathcal{E}_j(t_j)/r_j^1$ and by assumption $\rho_j^1 > r_j^1$. Because of (24), (25) and (26) the smoother bursts at rate $r_j^1$ for a duration of $\tau_{\text{on}_j}$ when fed with an input burst at rate $\rho_j^1$ for a duration of $t_{\text{on}_j}$. The average rate of the smoother output is $\mathcal{E}_j(t_{\text{on}_j})/T_j = \rho_j \leq r_j^{M_j}$, where the last inequality follows from the stability condition.

Now consider the case $\rho_j^1 < r_j^1$ and $\mathcal{E}_j(t_j) \geq S_j(\delta_j)$. We have $t_{\text{on}_j} \leq t_j$ since $t_{\text{on}_j} = S_j(\delta_j)/\rho_j^1$ and $t_j = \mathcal{E}_j(t_j)/\rho_j^1$ and by assumption $S_j(\delta_j) \leq \mathcal{E}_j(t_j)$. This implies that $\mathcal{E}_j(t_{\text{on}_j}) = \rho_j^1 t_{\text{on}_j}$. Hence

$$S_j(\delta_j) = \mathcal{E}_j(t_{\text{on}_j}). \tag{27}$$

Note furthermore that

$$\delta_j \leq t_{\text{on}_j} \tag{28}$$

since $\delta_j = S_j(\delta_j)/r_j^1$ and $t_{\text{on}_j} = S_j(\delta_j)/\rho_j^1$ and by assumption $r_j^1 > \rho_j^1$. Because of (27), (28) and $\rho_j^1 < r_j^1$ (by assumption) the smoother passes the input burst at rate $\rho_j^1$ for a duration of $t_{\text{on}_j}$ unchanged. The average rate of the smoother output is $\mathcal{E}_j(t_{\text{on}_j})/T_j = \rho_j \leq r_j^{M_j}$, where the last inequality follows from the stability condition.

Finally, consider the case $\rho_j^1 < r_j^1$ and $\mathcal{E}_j(t_j) < S_j(\delta_j)$. These two assumptions imply that the smoother can pass the input burst of size $\mathcal{E}_j(t_j)$ at rate $\rho_j^1$. The average rate of the smoother output is $\mathcal{E}_j(t_{\text{on}_j})/T_j = \rho_j \leq r_j^{M_j}$, where the last inequality follows from the stability condition. ■

**Proof of Theorem 2:** Using Lemma 3 and mimicking the proof of Theorem 1 we obtain

$$\phi_j \quad = \quad \frac{E\left[(\sum_{k=1}^{J} \tilde{U}_k - C)^+ \tilde{U}_j\right]}{C \cdot E[\tilde{U}_j]},$$

where $\tilde{U}_1, \ldots, \tilde{U}_J$ are defined in Lemma 3. Using the fact that $\tilde{U}_j$ is a Bernoulli random variable, we obtain from the above expression

$$\begin{aligned} \phi_j \quad &= \quad \frac{E\left[(\sum_{k \neq j} \tilde{U}_k + \min(r_j^1, \rho_j^1) - C)^+\right]}{C} \\ &\geq \quad \frac{E\left[(\sum_{k \neq j} \tilde{U}_k + c_j^* - C)^+\right]}{C}, \end{aligned} \tag{29}$$

where the last inequality follows from Lemma 2.

From (15) and (29) it remains to show that

$$E[(\sum_{k\neq j} U_k^* + c_j^* - C)^+] \leq E[(\sum_{k\neq j} \tilde{U}_k + c_j^* - C)^+]. \tag{30}$$

From Lemma 2 and Proposition 1.5.1 in [27]

$$U_k^* \leq_{icx} \tilde{U}_k \quad \text{for all } k = 1, \ldots, J. \tag{31}$$

The inequality (30) follows from (31), the independence of $U_1^*, \ldots, U_J^*$ and an argument that parallels the argument in the proof of Theorem 1. □

# References


[1] T. Cormen, C. Leiserson, and R. Rivest. *Introduction to Algorithms*. MIT Press, Cambridge, MA, 1990.

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

[3] R. Cruz. A calculus for network delay, part II: Network analysis. *IEEE Transactions on Information Theory*, 37(1):122–141, January 1991.

[4] R. Cruz. Quality of service guarantees in virtual circuit switched networks. *IEEE Journal on Selected Areas in Communications*, 13(6):1048–1056, August 1995.

[5] B. T. Doshi. Deterministic rule based traffic descriptors for broadband ISDN: Worst case behavior and connection acceptance control. In J. Labetoulle and J. W. Roberts, editors, *Proceedings of International Teletraffic Congress (ITC) 14*, pages 591–600. Elsevier Science B. V., 1994.

[6] B. T. Doshi. Deterministic rule based traffic descriptors for broadband ISDN: Worst case behavior and its impact on connection acceptance control. *International Journal of Communication Systems*, 8:91–109, 1995.

[7] A. Elwalid, D. Heyman, T. Lakshman, D. Mitra, and A. Weiss. Fundamental bounds and approximations for ATM multiplexers with application to video teleconferencing. *IEEE Journal on Selected Areas in Communications*, 13(6):1004–1016, August 1995.

[8] A. Elwalid, D. Mitra, and R. H. Wentworth. A new approach for allocating buffers and bandwidth to heterogeneous regulated traffic in an ATM node. *IEEE Journal on Selected Areas in Communications*, 13(6):1115–1127, August 1995.

[9] ATM Forum. *ATM User–Network Interface Specification, Version 3.0*. Prentice–Hall, 1993.

[10] L. Georgiadis, R. Guerin, V. Peris, and K.N. Sivarajan. Efficient network QoS provisioning based on per node traffic shaping. *IEEE/ACM Transactions on Networking*, 4(4):482–501, August 1996.

[11] I. Hsu and J. Walrand. Admission control for ATM networks. In *IMA Workshop on Stochastic Networks*, Minneapolis, Minnesota, March 1994.

[12] E. Knightly, J. Liebeherr, D. Wrege, and H. Zhang. Fundamental limits and tradeoffs for providing deterministic guarantees to VBR video traffic. In *Proceedings of IEEE Infocom '95*, Boston, MA, April 1995.

[13] E. W. Knightly. H-BIND: A new approach to providing statistical performance guarantees to VBR traffic. In *Proceedings of IEEE Infocom '96*, San Francisco, CA, April 1996.

[14] E. W. Knightly and H. Zhang. Providing end–to–end statistical performance guarantees with bounding interval dependent stochastic models. In *Proceedings of ACM Sigmetrics '94*, 1994.

[15] E. W. Knightly and H. Zhang. D-BIND: an accurate traffic model for providing QoS guarantees to VBR traffic. *IEEE/ACM Transactions on Networking*, 5(2):219–231, April 1997.

[16] T. Lee, K. Lai, and S. Duann. Design of a real–time call admission controller for ATM. *IEEE/ACM Transactions on Networking*, 4(5):758–765, October 1995.

[17] J. Liebeherr and D. Wrege. Video characterization for multimedia networks with a deterministic service. In *Proceedings of IEEE Infocom '96*, San Francisco, CA, March 1996.

[18] F. LoPresti, Z. Zhang, D. Towsley, and J. Kurose. Source time scale and optimal buffer/bandwidth trade–off for regulated traffic in an ATM node. In *Proceedings of IEEE Infocom*, Kobe, Japan, April 1997.

[19] W. Luo and M. El Zarki. Analysis of error concealment schemes for MPEG-2 video transmission over ATM based networks. In *Proceedings of SPIE Visual Communications and Image Processing 1995*, Taiwan, May 1995.

[20] V. Peris. *Architecture for Guaranteed Delay Service in High Speed Networks.* PhD thesis, Institute for Systems Research, University of Maryland, College Park, 1997.

[21] S. Rajagopal, M. Reisslein, and K. Ross. Packet multiplexers with adversarial regulated traffic. In *Proceedings of IEEE Infocom '98*, pages 347–355, San Francisco, CA, April 1998.

[22] M. Reisslein and K. W. Ross. Call admission for prerecorded sources with packet loss. *IEEE Journal on Selected Areas in Communications*, 15(6):1167–1180, August 1997.

[23] M. Reisslein, K. W. Ross, and S. Rajagopal. Guaranteeing statistical QoS to regulated traffic: The multiple node case. In *in preparation.*

[24] J. Roberts, U. Mocci, and J. Virtamo (Eds.). *Broadband Network Traffic: Performance Evaluation and Design of Broadband Multiservice Networks, Final Report of Action COST 242, (Lecture Notes in Computer Science Vol. 1155).* Springer Verlag, 1996.

[25] O. Rose. Statistical properties of MPEG video traffic and their impact on traffic modelling in ATM systems. Technical Report 101, University of Wuerzburg, Insitute of Computer Science, Am Hubland, 97074 Wuerzburg, Germany, February 1995.
ftp address and directory of the used video traces:
ftp-info3.informatik.uni-wuerzburg.de /pub/MPEG/.

[26] S. Shenker and J. Wroclawski. Request for comments 2215: General characterization parameters for integrated service network elements. September 1997.

[27] D. Stoyan. *Comparison methods for queues and other stochastic models.* Wiley, 1983.

[28] D. Wrege, E. Knightly, H. Zhang, and J. Liebeherr. Deterministic delay bounds for VBR video in packet–switching networks: Fundamental limits and tradeoffs. *IEEE/ACM Transactions on Networking*, 4(3):352–362, June 1996.

[29] H. Zhang. Providing end-to-end performance guarantees using non-workconserving disciplines. *Computer Communications: Special Issue on System Support for Multimedia Computing*, 18(10), October 1995.

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