INTERNATIONAL JOURNAL OF COMMUNICATION SYSTEMS
Int. J. Commun. Syst. 2001; **14**:735–761 (DOI: 10.1002/dac.503)

Measurement-based admission control for bufferless multiplexers

Martin Reisslein*,†

GMD FOKUS, Kaiserin–Augusta–Allee 31, 10589 Berlin, Germany

SUMMARY

In this paper we study measurement-based admission control for bufferless multiplexing, which is very attractive for real-time traffic. We first discuss a novel large deviations (LD) approach to measurement-based admission control. We then provide an extensive review of the existing literature on measurement-based admission control. We conduct simulation studies with traces of MPEG 1 encoded movies to compare the performance of the admission rules in the literature with that of our large deviations approach. We demonstrate that our LD approach achieves higher link utilizations. Finally, we compare the performance of measurement-based admission control with that of traditional admission control, which relies on *a priori* traffic descriptors. Our numerical work indicates that measurement-based admission control achieves significant gains in link utilizations over traditional admission control. Copyright © 2001 John Wiley & Sons, Ltd.

KEY WORDS: bufferless multiplexing; large deviations; measurement-based admission control; statistical QoS

1. INTRODUCTION

Call admission control is performed in integrated services networks to ensure that the connections' quality of service (QoS) requirements are met. A call admission test is performed before a new connection is accepted. The new connection is accepted if and only if the network is able to meet the QoS requirements of all already existing connections as well as the new connection.

Traditional call admission tests are based on *a priori* characterizations (e.g. leaky bucket characterizations) of the connections' traffic [1]. Oftentimes, however, it is difficult, if not impossible, to provide an accurate *a priori* characterization of a connection's traffic. This is especially true for traffic emanating from live sources, such as the video traffic from the live coverage of a sporting event. Even if accurate *a priori* characterizations are available, however, traditional call admission tests typically over-provision networking resources. This is because

* Correspondence to: Martin Reisslein, Department of Electrical Engineering, Arizona State University, P.O. Box 877206, Tempe, AZ85287-7206, U.S.A.

† E-mail: reisslein@asu.edu

Published online 4 June 2001

Revised September 2000
Accepted January 2001

traditional call admission tests usually assume that the connections are adversarial to the extent permitted by the *a priori* characterizations and transmit worst-case traffic patterns [2–6]. This assumption, however, is often overly conservative, as in most practical circumstances connections do not transmit worst-case traffic patterns. As a consequence traditional call admission tests typically over-provision networking resources and thus underutilize the network.

Measurement-based admission control is a promising alternative to admission control based on *a priori* traffic descriptors. Instead of relying on *a priori* traffic characterizations, measurement–based admission control bases admission decisions primarily on traffic measurements. Admission decisions are based on measurements of the actual traffic from the already existing connections and an *a priori* characterization of the connection requesting establishment. The *a priori* characterization of the connection requesting establishment can be very simple, such as a peak rate specification. An overly conservative *a priori* characterization does not result in an over-provisioning of resources for the entire lifetime of the new connection, as the new connection—once admitted—is included in the measurements and is no longer characterized by its *a priori* specification. Thus, measurement-based admission control is able to exploit the statistical multiplexing effect and achieves high network utilizations.

Our focus is on measurement-based admission control for bufferless multiplexing. Bufferless multiplexing is very attractive for real-time streaming traffic since it ensures that the traffic incurs minimal delay and preserves the traffic characteristics throughout the network [7]. In this paper we discuss measurement-based admission rules that base admission decisions on measurements of the aggregate traffic from the already existing connections. We do not consider admission rules that require per-flow traffic measurements, as it is difficult to conduct per-flow measurements accurately and cost-efficiently in practice. Admission tests based on aggregate measurements cannot enforce per-flow QoS; these would require per-flow measurements which are not practicable. Therefore, in this paper we focus on measurement-based admission rules that provide aggregate QoS. We study the measurement-based admission rules within the smoothing/bufferless multiplexing framework [5,6]. The key aspects of the smoothing/bufferless multiplexing framework are to (i) pass each connection's traffic through a *buffered* smoother (peak rate limiter) at the connection's input to the network, and (ii) use *bufferless* statistical multiplexing inside the network. The bufferless multiplexing inside the network has the advantage that a new connection's *a priori* characterization (e.g. peak rate) does not change as it passes through a bufferless node. Thus, the same *a priori* characterization can be used for the admission test at each node traversed by the new connection. However, to simplify the discussion and highlight the measurement aspect of the admission rules we focus on a single bufferless node in this paper.

The contributions of this paper are threefold. First, we develop and evaluate a novel large deviations (LD) approach to measurement-based admission control. In this LD approach traffic measurements are used to estimate the logarithmic moment generating function of the aggregate arrival stream. From this estimate of the logarithmic moment generating function we compute an estimate of the loss probability at the node using the LD approximation. A new connection requesting establishment is accepted if the estimated loss probability is less than some miniscule QoS parameter ε, say $\varepsilon = 10^{-6}$, and rejected otherwise.

Secondly, we provide an extensive review of the existing literature on measurement-based admission control. We compare the performance of the measurement-based admission rules in the literature with that of our LD approach through simulations with traces of MPEG 1 encoded movies. Our simulation results indicate that the LD approach achieves both higher link utilizations and smaller loss probabilities than the time-scale decomposition approach [8]. The time-scale decomposition approach in turn performs better than the measured sum approach [9]. These results are in contrast to a recent comparative study by Breslau *et al.* [10]; they find that all measurement-based admission rules achieve that same performance at a buffered multiplexer.

Lastly, we compare the performance of measurement-based admission control with that of traditional admission control which relies exclusively on *a priori* traffic characterizations. We demonstrate that measurement-based admission control achieves significantly higher link utilizations than traditional admission control that relies on leaky bucket characterizations and assumes worst-case on–off traffic patterns.

1.1. Related work

There is a large body of literature on measurement-based admission control which is complementary to the issues addressed in this paper. Jamin *et al.* [11,9] and Casetti *et al.* [12] study the so-called measured sum approach, which bases admission decisions on an estimate of the mean aggregate arrival rate. Gibbens *et al.* [13] and Gibbens and Kelly [14] study Chernoff bound-based admission rules. They assume on–off traffic and consider a tangent on the effective bandwidth function in their admission rule. Floyd [15] as well as Brichet and Simonian [16] study measurement-based admission control based on the Hoeffding bound. They employ an exponential weighted moving average measurement mechanism. All these approaches have structural similarities, which are studied by Jamin and Shenker [17]. Roughly speaking, they all rely on the mean of the measured arrivals (higher moments are not considered).

The time-scale decomposition approach of Grossglauser and Tse [18,8] relies on estimates of the first and second moment of the arrival stream. They estimate both mean and variance of the arrivals from the measurements and estimate the loss probability at the bufferless multiplexer with the normal approximation. Lee and Zukerman [19] study the assumption of Gaussian aggregate traffic, which is used in the time-scale decomposition approach.

Large deviation approaches to measurement-based admission control differ from the previous approaches in that they take the entire moment generating function of the arrivals into consideration. Large-deviation-based admission rules for buffered multiplexers are studied by Dublin's Applied Probability Group; see [20,21] and references therein. They estimate the generating function of the arrival process from measurements and use the large buffer asymptotic to estimate the loss probability at a buffered multiplexer. Rácz [22] studies the robustness of these admission rules. Walsh and Duffield [23] and McGurk and Walsh [24] study an admission rule based on the shape function [25]. This approach is more flexible in that it can be employed for buffered as well as bufferless multiplexers. The drawback of the studied shape function approach is that it requires per-flow traffic measurements, which are difficult to conduct in practice. Tse and Grossglauser [26] study a large deviation admission rule for bufferless multiplexer in which the generating function of the arrivals is estimated from per-flow measurements. Our admission rule for *bufferless* multiplexers differs from the approaches in the literature in that the generating function of the arrivals is estimated from measurements of the *aggregate* traffic stream (per-flow measurements are not required).

Zukerman and Lee [27] and Lee *et al.* [28] propose and evaluate a comprehensive framework for measurement-based admission control for bufferless as well as buffered multiplexers. In their framework admission decisions are based on histograms of the measured aggregate arrivals over several time scales.

We conclude this literature review by noting that Knightly and Qiu [29] study an admission rule for buffered multiplexers that estimates maximal rates over different interval lengths (i.e. the maximal rate envelope [30]) from traffic measurements.

2. A LARGE DEVIATIONS APPROACH TO MEASUREMENT-BASED ADMISSION CONTROL

In this section[‡] we discuss our large deviations (LD) approach to measurement–based admission control. We discuss a basic admission rule first and study then some important refinements. 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 focus throughout this paper on a single node consisting of a bufferless multiplexer that feeds into a link of capacity C. (For packetized traffic a small buffer is needed for packet-scale queueing; we consider a bufferless multiplexer in the sense that there is no burst-scale queueing [1].) Consider a set of J connections. In the smoothing/bufferless multiplexing framework each connection j, $j = 1, \ldots, J$, is passed through a buffered smoother before it is multiplexed onto the bufferless link. The smoother limits the peak rate of connection-j traffic entering the bufferless multiplexer to c_j^* (see Figure 1). Let $U_j(t), j = 1, \ldots, J$, denote the rate at which connection-j traffic arrives to the bufferless multiplexer at time t. The smoother ensures that $U_j(t) \leqslant c_j^* \; \forall t \geqslant 0$. Now regard the jth arrival process as a stochastic process. Let $(U_j(t),\ t \geqslant 0)$ denote the jth arrival process. Let $X(t)$ denote the aggregate arrival rate at time t:

$$X(t) = \sum_{j=1}^{J} U_j(t)$$

and let $(X(t), t \geqslant 0)$ denote the aggregate arrival process. The expected long-run fraction of traffic lost due to link overflow is

$$P_{\text{loss}} = E\left[\lim_{\Delta \to \infty} \frac{\int_0^\Delta (X(t) - C)^+ \mathrm{d}t}{\int_0^\Delta X(t)\,\mathrm{d}t}\right] \tag{1}$$

where the expectation is over all arrival processes and $(x)^+ := \max(0, x)$.

Our goal is to develop a measurement-based call admission rule that ensures that P_{loss} is less than some minute ε, such as $\varepsilon = 10^{-4}$ or 10^{-6}. The call admission decisions are based on measurements of the aggregate arrival rate. In practical systems, however, it is impossible to measure the instantaneous arrival rate $X(t)$. For this reason, we divide time into slots of length T and measure the amount of traffic arriving in an interval of length T. Let X_n denote the amount of traffic arriving in the interval $[nT, (n+1)T]$, i.e.

$$X_n = \int_{nT}^{(n+1)T} X(t)\,\mathrm{d}t$$

For small T we can reasonably approximate

$$\int_{nT}^{(n+1)T} (X(t) - C)^+ \mathrm{d}t \approx (X_n - CT)^+ \tag{2}$$

[‡] Parts of this section have been presented at the *IEEE Symposium on Computers and Communications*, Antibes, France, July 2000.

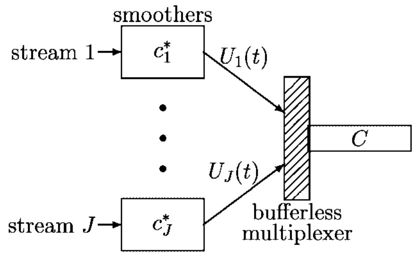

Figure 1. The traffic of connection j is passed through a smoother that limits the peak rate to c_j^*. The smoothed traffic is then multiplexed onto a bufferless link with capacity C.

This approximation is particularly good when the fluctuation of the aggregate arrival process $(X(t), t \geqslant 0)$ is on a time scale larger than the slot length T. The slot length should therefore be set to the smallest value that allows for meaningful traffic measurements. In practical systems we suggest to set T to a few packet transmission times. (For a detailed analysis of sampling for measurement-based admission control we refer the interested reader to Reference [31].)

Throughout this paper we shall assume that Approximation (2) is exact. Substituting (2) into Equation (1) we obtain

$$P_{\text{loss}} = E\left[\lim_{N\to\infty} \frac{\sum_{n=0}^{N}(X_n - CT)^+}{\sum_{n=0}^{N} X_n}\right] \tag{3}$$

A practical measurement-based call admission rule cannot rely on measurements over an infinite time horizon, but instead must base its decisions on some finite portion of the history of the aggregate streams behaviour. We propose to base admissions decisions on the measured aggregate arrivals in the past M slots, i.e. $M \geqslant 1$ is the measurement window. Before we describe our admission rule in detail we need to introduce some notation. Let k denote the slot in which a new stream with smoother rate c_k^* requests connection establishment. Our admission rule relies on the measured aggregate arrivals in slots $k-M,\ldots,k-1$. Let x_i, $i=1,\ldots,M$, denote the measured aggregate arrivals in slot $k-i$.

Now consider the random variable X_k denoting the (not yet measured) aggregate arrivals in slot k. Define the estimated loss probability $P_{\text{loss}}^{\text{est}}$ as follows:

$$P_{\text{loss}}^{\text{est}} = \frac{E[(X_k - (C - c_k^*)T)^+]}{E[X_k]} \tag{4}$$

$P_{\text{loss}}^{\text{est}}$ is the expected fraction of traffic lost by the already established connections at a bufferless link of capacity $C - c_k^*$ during slot k. Note that we are conservatively setting aside the peak rate c_k^* for the stream requesting establishment. Our strategy is to base admission decisions on $P_{\text{loss}}^{\text{est}}$. If $P_{\text{loss}}^{\text{est}} \leqslant \varepsilon$ connection k is admitted, otherwise it is rejected.

We evaluate $P_{\text{loss}}^{\text{est}}$ using the large deviations (LD) approximation. Toward this end, let m_X denote the estimate of $E[X_k]$, the mean of X_k. We compute the estimate m_X by averaging over the aggregate arrivals in slots $k-M,\ldots,k-1$:

$$m_X = \frac{1}{M}\sum_{i=1}^{M} x_i$$

Table I. Summary of notation.

c_j^*	Smoother rate (= peak rate) of connection j in bit/s
C	Service rate of multiplexer in bit/s
ε	QoS parameter (= target loss probability)
k	Index of slot in which a new connection with smoother rate c_k^* requests establishment
m_X	Estimate of average aggregate arrivals in one slot in bit
$P_{\text{loss}}^{\text{est}}$	Estimated loss probability; computed from m_X and $\mu_X(s)$ using Large Deviation approximation
P_{loss}	Actual loss probability; obtained through simulation
T	Slot length in seconds
x_i	Measured aggregate arrivals in slot $k-i$ in bit
$\mu_X(s)$	Estimate of logarithmic moment generating function of aggregate arrivals in one slot

Furthermore, let $\mu_X(s)$ denote the estimate of $\ln E[\mathrm{e}^{sX_k}]$, the logarithmic moment generating function of X_k. (The notation used in this paper is summarized in Table I.) Again, we compute $\mu_X(s)$ by averaging over the M latest measurements:

$$\mu_X(s) = \ln \frac{1}{M} \sum_{i=1}^{M} \mathrm{e}^{sx_i} \tag{5}$$

Note that Equation (5) generalizes the notion of the sample mean and allows for the estimation of arbitrary moments of the aggregate arrival stream. For instance, we obtain the logarithm of m_X by evaluating the derivative of $\mu_X(s)$ with respect to s at $s = 0$, i.e. $\mu_X'(0) = \ln m_X$. The LD approximation of Equation (4) is given by [1]

$$P_{\text{loss}}^{\text{est}} \approx \frac{1}{m_X s^{\star 2} \sqrt{2\pi \mu_X''(s^\star)}} \mathrm{e}^{-s^\star (C - c_k^*)T + \mu_X(s^\star)} \tag{6}$$

where $s^\star$ minimizes the (convex) exponent in (6), i.e. $s^\star$ is the unique solution to

$$\mu_X'(s^\star) = (C - c_k^*)T \tag{7}$$

(Approximation (6) is a slight variation of the result by Bahadur and Rao [32] in that it approximates the 'long-run fraction of information lost' while Bahadur and Rao's result approximates the 'long-run fraction of time during which loss occurs'; see Reference [1] for details.) In summary, our basic measurement-based admission rule works as follows: suppose that in slot k a connection with peak rate c_k^* requests establishment and the QoS requirement is $P_{\text{loss}} \leqslant \varepsilon$. First, we estimate the logarithmic moment generating function of the aggregate arrival stream based on the measurements in the last M slots using Equation (5). We then estimate $P_{\text{loss}}^{\text{est}}$ using the LD approximation (6) and admit connection k if $P_{\text{loss}}^{\text{est}} \leqslant \varepsilon$, otherwise connection k is rejected.

We now evaluate the basic measurement-based admission rule using traces from MPEG 1 encoded movies. We obtained the frame size traces, which give the number of bits in each video frame, from the public domain [33]. The movies were compressed with the Group of Pictures (GOP) pattern IBBPBBPBBPBB at a frame rate of $F = 24$ frames/s. Each of the movie traces available from Reference [33] has $N = 40\,000$ frames, corresponding to about 28 min. Let $f_n(j)$,

$n = 1, \ldots, N$, denote the size of the nth frame of video j in bits. We convert the discrete frame size trace to a fluid flow by transmitting the nth frame of video j at rate $f_n(j) \cdot F$ over the interval $[(n-1)/F, n/F]$. The numerical results reported in this paper were obtained with the '*Silence of the Lambs*' (lambs) trace. The lambs trace is the burstiest trace available from the library of traces [33]. Specifically, the lambs trace has an average frame size of 8048 bit, which corresponds to an average rate of 193.2 kbit/s. The trace has a peak-to-mean ratio of 18.4 and is therefore considered extremely bursty. Because of its burstiness the lambs trace poses a particular challenge for admission control. In the numerical experiments reported in this paper all video streams use the lambs trace but each stream has its own independent random phase.

We evaluate the measurement-based admission rule within the smoothing/bufferless multiplexing framework [5,6]. Each video stream is passed through a smoother before it enters the bufferless multiplexer. The smoother for connection j consists of a buffer which serves the traffic at rate c_j^*. When the smoother buffer is non-empty, 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. The smoother thus limits the peak rate of connection-j traffic entering the multiplexer to c_j^*. The smoother rate c_j^* is set to the smallest value that guarantees that the video traffic is delayed by no more than a connection-specific delay limit in the smoother (see Reference [5] for details). We initially set the delay limit for all connections to 10 frame periods, i.e. 10/24 s. The corresponding smoother rate for the lambs trace is 731.6 kbit/s. Throughout this paper we set the rate of the bufferless multiplexer to $C = 45$ Mbit/s. Even though we consider homogeneous streams in the numerical work in this paper, we emphasize that the proposed approach naturally accommodates heterogeneous streams. The streams may have vastly different traffic characteristics and delay limits (and thus vastly different smoother rates c_j^*). The smoother rate of a connection, i.e. the peak rate at which the connection's traffic enters the bufferless multiplexer is accounted for in $P_{\text{loss}}^{\text{est}}$ (4). We also note that throughout this paper we focus on admission policies that allow for the complete sharing of the link bandwidth among the ongoing streams.

We simulate the system consisting of smoothers and bufferless multiplexer on a per frame period basis. Throughout we set the slot length of the measurement algorithm to the length of one frame period, i.e. $T = 1/F$. In the simulation calls arrive according to a Poisson process. We fix the call arrival rate at 1 call/10 frame period; thus the time between call arrivals is exponentially distributed with a mean of 10 frame periods. For each accepted call we draw a random starting frame. The starting frames are independent and uniformly distributed over $[1, N]$. For each call we also draw a random life time. In our first set of experiments we draw the lifetimes from an exponential distribution with a mean of 6000 frame periods ($= 250$ s). With these parameters the link operates in constant overload. (To see this, note that the stability limit of the 45 Mbps link is 232 lambs streams, each with an independent exponentially distributed lifetime with a mean of 6000 frame periods. Thus, at the stability limit on average 38 per cent of the offered calls are accepted.) We allow the simulations to warm up for 60 000 frame periods. We determined with Schruben's test [34] that this warm-up is sufficient for the systems to reach steady state.

The goal of the simulations is to obtain estimates for the average number of admitted streams, denoted by J_{avg}, and the loss probability P_{loss} (3). We estimate J_{avg} and P_{loss} using the method of batch means [35]. We use a batch size of 6000 frame periods. In order to ensure the independence of the batches, we separate successive batches by 12 000 frame periods, twice the average lifetime of a stream. We run each simulation until the width of the 90 per cent confidence interval of the loss probability is less than 20 per cent of the corresponding point estimate. We observed that the estimate for the average number of admitted streams converges much faster than the estimate for the loss probability. By the time the confidence interval for the loss probability has converged to less than 20 per cent of the point estimate, the width of the 90 per cent confidence interval for the number of admitted streams is typically less than 1 per cent of the point estimate.

In the first set of simulation experiments we evaluate the basic measurement-based admission rule. We set the QoS parameter to $\varepsilon = 10^{-6}$ and run simulations for different values of M, the length of the measurement window. The results are reported in Table II; in order to avoid visual clutter only point estimates are reported. We observe from the table that the loss probabilities are one to two orders of magnitude larger than the target loss probability $\varepsilon = 10^{-6}$. For a more detailed analysis of the basic admission rule we refer the interested reader to Reference [36]. In summary, we find that weighing the measured aggregate arrivals in the measurement window uniformly results in periodic surges in the number of admitted streams, which periodically lead to losses. We next try to improve the measurement-based admission rule by weighing the more recent measurement more heavily when estimating the logarithmic moment generating function.

2.1. Non-uniform weight refinement

The basic idea of the non-uniform weight refinement is to give the recent measurements more weight when estimating the logarithmic moment generating function. Toward this end, let p_i, $i = 1, \ldots, M$, denote weights with $0 \leqslant p_i \leqslant 1$ and $\sum_{i=1}^{M} p_i = 1$. Throughout this paper we use exponentially decaying weights:

$$p_i = \frac{\mathrm{e}^{-i/\tau_p}}{\sum_{l=1}^{M} p_l}, \quad i = 1, \ldots, M$$

where τ_p is a tuning parameter. With the non-uniform weights the estimates m_X and $\mu_X(s)$ are computed as

$$m_X = \sum_{i=1}^{M} p_i x_i \quad \text{and} \quad \mu_X(s) = \ln \sum_{i=1}^{M} p_i \,\mathrm{e}^{s x_i} \tag{8}$$

As before, these estimates are used to compute $P_{\text{loss}}^{\text{est}}$ (4) and the connection requesting establishment is accepted if $P_{\text{loss}}^{\text{est}} \leqslant \varepsilon$ and rejected otherwise. We refer to this call admission rule as *measurement-based admission rule with non-uniform weights.*

For the evaluation of the measurement-based admission rule with non-uniform weights we set the length of the measurement window to $M = 6000$. In order to avoid unnecessary computation we ignore measurements that are assigned weights less than 10^{-9}. We denote M_{eff} for the number

Table II. Evaluation of basic measurement-based admission rule. Average number of admitted streams and loss probability for different measurement window lengths. The QoS parameter is set to $\varepsilon = 10^{-6}$.

M	50	100	200	500	1000
J_{avg}	204	201	198	192	183
P_{loss}	9.8×10^{-4}	6.0×10^{-4}	3.8×10^{-4}	1.9×10^{-4}	1.3×10^{-4}
M	2000	4000	6000	9000	12000
J_{avg}	171	147	131	107	93
P_{loss}	7.6×10^{-5}	5.3×10^{-5}	4.5×10^{-5}	2.8×10^{-5}	3.2×10^{-5}

Table III. Evaluation of measurement-based admission rule with non-uniform weights.

τ_p	∞	6000	3000	1200	600	300	120	60
J_{avg}	131	132	134	137	157	173	191	198
P_{loss}	4.5×10^{-5}	4.4×10^{-5}	4.0×10^{-5}	3.1×10^{-5}	6.2×10^{-5}	8.1×10^{-5}	1.6×10^{-4}	2.1×10^{-4}
M_{eff}	6000	6000	6000	6000	6000	4506	1912	998

of samples actually used in the estimation. We note that the computational complexity of the LD admission test is $O(M_{\text{eff}})$. We found in our numerical experiments that it takes typically M_{eff} 0.13 ms to perform one admission test on a 1997 workstation. (A speed-up of the admission test is proposed in Section 3.)

Table III gives the average number of admitted streams, the loss probability and M_{eff} for different values of τ_p. We see from the table that the average number of connections increases as τ_p decreases. It is interesting to note that for decreasing τ_p the loss probability first decreases slightly and then increases. However, the loss probability is generally over one order of magnitude larger than the imposed QoS requirement. We refer the interested reader to Reference [36] for a detailed analysis.

2.2. Peak rate reservation refinement

To motivate the peak rate reservation refinement consider a scenario where a stream, say stream u, is admitted and a few slots later another stream, say stream v, requests establishment. When conducting the admission test for stream v only a few aggregate arrival measurements that include stream-u traffic are available. These few measurements that include stream-u traffic have little impact on the estimated logarithmic moment generating function $\mu_X(s)$. This is especially the case when the measurement window is long and older measurements are assigned relatively large weights. The new stream u is therefore underrepresented in $\mu_X(s)$ and the aggregate bandwidth demand is underestimated. As a result the estimated loss probability $P_{\text{loss}}^{\text{est}}$ is too small and too many connections are admitted. In summary, the problem with the measurement-based admission rules studied so far is that they 'forget' the peak rates of recently admitted streams even though the new stream's traffic is not yet fully reflected in the measurements.

To fix this shortcoming we add a refinement to the measurement-based admission rule with non-uniform weights. This refinement works roughly as follows. We keep a record of peak rates of the recently admitted streams. When conducting an admission test this record is used to compute a reserved peak rate denoted by c^*. The reserved peak rate c^* is computed by assigning weights to the recorded peak rates. Peak rates of relatively new streams are assigned weights close to one, while peak rates of relatively old streams are assigned weights close to zero. Thus, streams that are relatively new are mostly accounted for by the reserved peak rate. On the other hand, stream that have been established for a while are mostly accounted for by the traffic measurements. The reserved peak rate c^* is then subtracted from the link capacity C when computing the estimated loss probability $P_{\text{loss}}^{\text{est}}$.

To make these ideas a little more precise, suppose that in slot k a stream with peak rate c_k^* requests establishment. Let y_i, $i = 1, \dots, M$, denote the peak rates of the admitted streams in slots $k - i$, $i = 1, \dots, M$. y_i is set to zero if no new stream was admitted in slot $k - i$. Let q_i, $i = 1, \dots, M$, denote weights with $0 \leqslant q_i \leqslant 1$. Throughout this paper we use exponentially decaying weights

$$q_i = \mathrm{e}^{-i/\tau_q}, \quad i = 1, \ldots, M$$

where $\tau_q \geqslant 0$ is a tuning parameter. The reserved peak rate is computed as

$$c^* = c_k^* + \sum_{i=1}^{M} q_i y_i$$

We now define the estimated loss probability $P_{\text{loss}}^{\text{est}}$ as the expected fraction of traffic lost by the established connections at a bufferless link of capacity $C - c^*$, formally

$$P_{\text{loss}}^{\text{est}} := \frac{E[(X_k - (C - c^*)T)^+]}{E[X_k]}$$

As before, the estimated loss probability is computed using the LD approximation; the expression for the LD approximation of $P_{\text{loss}}^{\text{est}}$ (6) is modified in the obvious way. The logarithmic moment generating function $\mu_X(s)$ is evaluated using the non-uniform weight refinement (8). Connection k is admitted if $P_{\text{loss}}^{\text{est}} \leqslant \varepsilon$ and rejected otherwise. We refer to this admission rule as the *measurement-based admission rule with peak rate reservation.*

The parameter τ_q is used to tune the peak rate reservation. For $\tau_q = 0$ all the weights are zero and the measurement-based admission rule with peak rate reservation reduces to the admission rule with non-uniform weights. For strictly positive τ_q the weights q_i decay exponentially. The larger τ_q, the larger the peak rate reservation.

We now evaluate the measurement-based admission rule with peak rate reservation through simulation. The results are reported in Table IV. The table gives the average number of streams, J_{avg}, and the loss probability, P_{loss}, for different combinations of the tuning parameters τ_p and τ_q. Several points are noteworthy here. First, consider the column $\tau_p = 600$. We see that as τ_q increases from zero (i.e. no peak rate reservation) to 100 the average number of streams increases while the loss probability decreases. Loosely speaking the admission rule makes 'smarter' admission decisions by reserving more peak rate; it achieves both higher link utilizations

Table IV. Evaluation of measurement-based admission rule with peak rate reservation. Each table entry gives the average number of streams, J_{avg}, and the loss probability, P_{loss}, for a specific combination of the tuning parameters τ_p and τ_q.

		τ_p						
		1200	600	300	120	60	40	30
	0	137 3.1×10^{-5}	157 6.2×10^{-5}	173 8.1×10^{-5}	191 1.6×10^{-4}	198 2.1×10^{-4}	197 3.2×10^{-4}	198 3.8×10^{-4}
	50	147 4.3×10^{-6}	161 5.0×10^{-6}	174 8.9×10^{-6}	185 1.5×10^{-5}	188 2.7×10^{-5}	190 3.3×10^{-5}	192 4.6×10^{-5}
τ_q	100	150 8.0×10^{-7}	165 9.0×10^{-7}	173 1.1×10^{-6}	180 2.9×10^{-6}	183 4.1×10^{-6}	185 6.6×10^{-6}	186 9.0×10^{-6}
	125	154 4.8×10^{-7}	163 4.9×10^{-7}	172 6.4×10^{-7}	178 1.1×10^{-6}	181 2.0×10^{-6}	183 2.9×10^{-6}	184 4.3×10^{-6}
	200	156 2.4×10^{-9}	161 6.9×10^{-9}	166 3.2×10^{-8}	172 7.1×10^{-8}	174 2.3×10^{-7}	175 1.3×10^{-7}	177 4.7×10^{-7}

and smaller losses. As τ_q increases further, however, both J_{avg} and P_{loss} drop. Reading along any row of the table we see that for fixed τ_q both J_{avg} and P_{loss} increase with decreasing τ_p.

The goal of this simulation experiment is to find the combination of tuning parameters that gives good on-target performance, i.e. a loss probability nearly equal to ε, as well as high link utilizations. We see from the table that the combination $\tau_p = 120$ and $\tau_q = 125$ gives the highest J_{avg} among the combinations with P_{loss} nearly equal to $\varepsilon = 10^{-6}$. Unless stated otherwise these tuning parameters are used for all numerical experiments in the remainder of this paper.

Figure 2 shows typical sample path plots from the simulation for $\tau_p = 120$ and $\tau_q = 125$. Notice that the admission rule achieves a consistently high utilization of the bufferless link of capacity 1.875×10^6 bit/slot ($= 45$ Mbps $\times$ 1/24 s), while incurring actual losses only once around slot time 63 500.

3. COMPARISON WITH OTHER MEASUREMENT-BASED ADMISSION RULES

In this section we review the measurement-based admission rules in the existing literature and compare the performance of our large deviations based admission rule with that of the admission rules in the literature.

3.1. Measured sum approach

Jamin *et al.* in their seminal work on measurement-based admission control [11,9] develop a measurement-based admission rule for integrated services networks. They assume a network consisting of buffered multiplexers. Their admission rule consists of a delay criterion and a rate criterion. The delay criterion is designed to keep the delay in the network below a prespecified delay bound. The delay criterion takes the measured delay in the network and a leaky bucket characterization (i.e. average rate r and bucket depth b) of the connection requesting establishment into consideration. The rate criterion strives to keep the link utilizations below prespecified utilization targets. The rate criterion relies on the measured link utilizations and the leaky bucket rate r of the new stream. The new stream is admitted if it passes both the delay criterion and the rate criterion. In order to compare the performance of the admission rule of Jamin *et al.* with that of our large-deviations-based admission rule, we apply the admission rule of Jamin *et al.* to the smoothing/bufferless multiplexing networking architecture [5,6]. In the smoothing/bufferless multiplexing networking architecture the traffic is delayed by no more than the delay bound in the buffered smoother at the network edge. The bufferless multiplexers inside the network add no further delay. Therefore, there is no need to check the delay criterion. This simplifies call admission tremendously since delay measurements, which are difficult to conduct in practice, are no longer required.

We now briefly review the rate criterion; see References [11,9] for more details. First, we review the necessary notation. As before, T denotes the slot length. Note however, that the slots lengths of our Large Deviations based admission rule and the admission rule of Jamin *et al.* are fundamentally different. We base admission decisions on the estimate of the moment generating function of the aggregate arrival stream. To ensure that the estimate of the moment generating function correctly reflects the variability of the aggregate arrival stream we use a slot length short enough to capture individual bursts. Recall that for the numerical work in this paper we use a slot length of $1/F \approx 0.042$ s. Jamin *et al.* base admission decisions on the estimate of the average

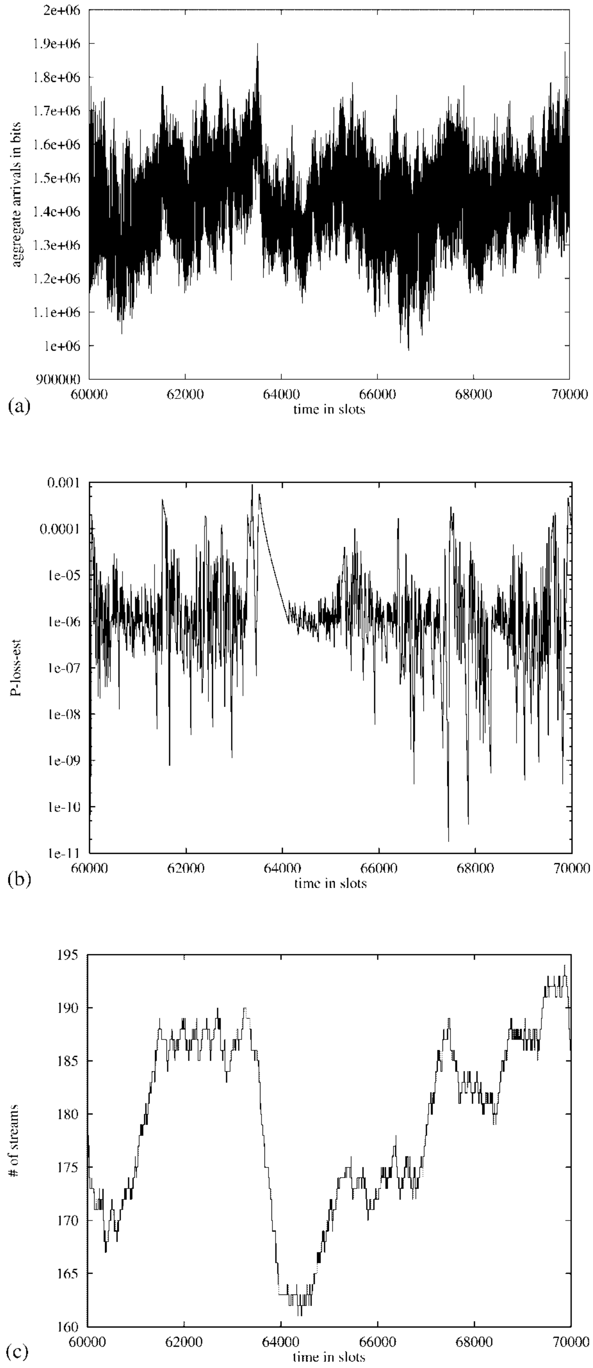

Figure 2. Sample path plots from simulation of measurement-based admission rule with peak rate reservation for $\tau_p = 120$ and $\tau_q = 125$. (a) aggregte arrivals versus time; (b) $P_{\text{loss}}^{\text{est}}$ versus time; (c) J versus time.

aggregate arrival rate. To obtain a good and stable estimate of the average aggregate arrival rate, they average the aggregate arrivals over a longer slot length and thus avoid capturing individual bursts. This is also reflected in the terminology as they refer to the slot length as an *averaging period*. They suggest to set the averaging period to 0.5 s. We therefore set the averaging period to $12/F$ for the numerical evaluation of the Jamin *et al.* approach in this paper.

Let $\hat{x}$ denote the estimate of the aggregate arrivals in one averaging period. To estimate $\hat{x}$ Jamin *et al.* employ a time-window measurement mechanism. Let W denote the length of the measurement window in multiples of the averaging period T. The time-window measurement mechanism works roughly as follows. The amount of traffic arriving in each averaging period is measured. At the end of the measurement window, that is, after obtaining W measurements, $\hat{x}$ is set to the largest of the W samples. When a new stream is admitted $\hat{x}$ is increased by the arrivals of the new stream (computed from the leaky bucket rate r), i.e. we set $\hat{x} \leftarrow \hat{x} + rT$. Furthermore, when a particular sample value is larger than $\hat{x}$ we do not wait until the end of the measurement window to update $\hat{x}$, but instead set $\hat{x}$ to this larger sample value right away. Finally, let v denote a prespecified utilization target. Jamin *et al.* suggest to set $v = 0.9$. With $v = 0.9$ the admission rule strives to keep the link utilization below 90 per cent . The rate criterion of Jamin *et al.* is to verify whether

$$\hat{x} + rT \leqslant vCT \tag{9}$$

The new stream with leaky bucket rate r is accepted if (9) holds, otherwise it is rejected.

We now compare the performance of the admission rule of Jamin *et al.* with that of our large-deviations-based admission rule. We use the load-loss curve [17] for the performance comparison. The load–loss curve is a plot of the loss probability P_{loss} versus the average number of admitted streams J_{avg}. Both P_{loss} and J_{avg} are obtained through simulation. We employ the simulation approach outlined in Section 2. For all simulations in this section we set the link capacity to $C = 45$ Mbps. All the traffic streams are '*Silence of the Lambs*' video streams, each with its own independent random phase. We consider two scenarios. Figure 3(a) shows the load–loss curves for the case where the video streams are passed through smoothers (see Figure 1) with a maximum smoothing delay of 10 frame periods before they enter the bufferless multiplexer. Figure 3(b) gives the load–loss curves for the case where the unsmoothed lambs video streams are multiplexed onto the bufferless link. The plots give the load–loss curves of the admission rule of Jamin *et al.* for different measurement windows W. Specifically, we show the load–loss curves for $W = 100T$, $10T$, $5T$ and $W = T$; recall that the averaging period is set to $T =$ 0.5 s as suggested by Jamin *et al.* [11,9]. The curves are obtained by varying the utilization target v. The curve for $W = 100T$, for instance, was obtained by running simulations for $v = 0.9$, 0.925, 0.95, 0.975 and 1.0. Two observations are in order. First, we observe that the actual link utilization differs significantly from the target utilization. For $W = 100T$ and $v = 0.9$ (see Figure 3(a)), for instance, the admission rule admits on average 162.4 unsmoothed lambs streams, which corresponds to an average link utilization of 70 per cent . The corresponding loss probability is 1.2×10^{-6}. Note that the admission rule of Jamin *et al.* is not designed to take a target loss probability as input. The second noteworthy observation is that for smaller measurement windows W the load–loss curves move towards the lower right corner of the plots. This means that for smaller W the admission control rule performs better; it achieves higher link utilizations and smaller loss probabilities. Jamin and Shenker [17] define the load–loss frontier as the load–loss curve that gives the smallest loss probabilities for the range of link utilizations. We see from the plots that the load–loss frontiers are composed of the load–loss curves for $W = 5T$ and T. The plots in Figure 3 give also

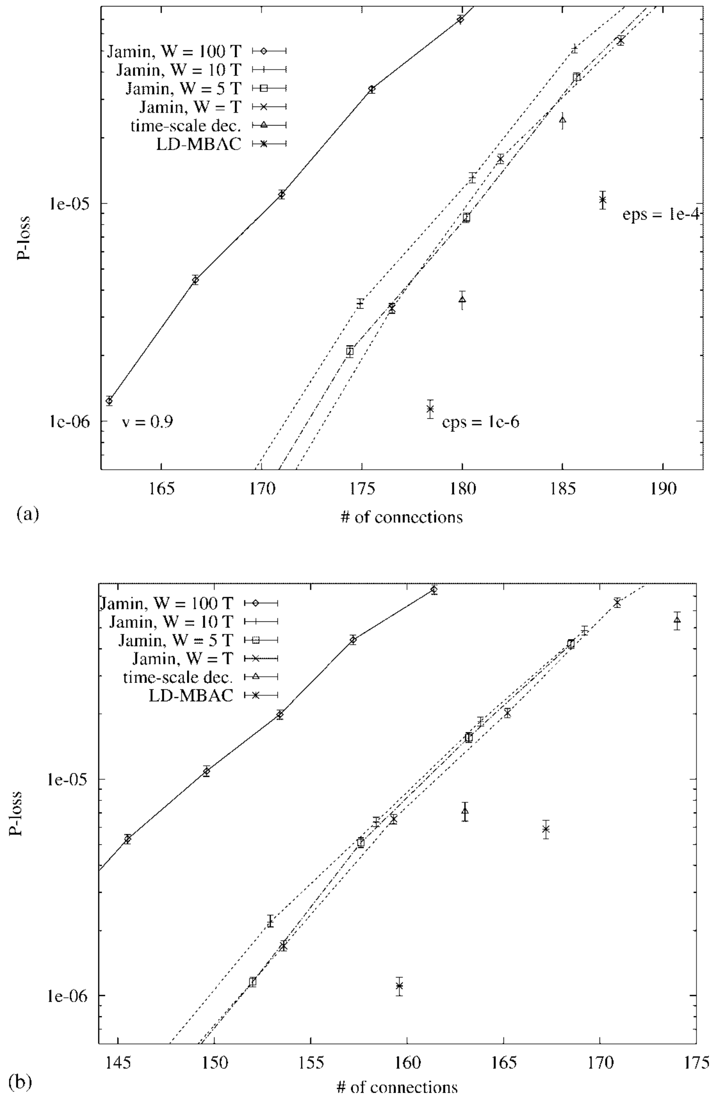

Figure 3. Comparison of admission rules in the literature with our LD based admission rule.

the load–loss points of our large-deviations-based admission rule (LD–MBAC). These points were obtained by setting $\tau_p = 120$ and $\tau_q = 125$ and running simulations for the target loss probabilities $\varepsilon = 10^{-6}$ and 10^{-4}. As in Section 2 we run the simulations for the LD admission rule until the width of the 90 per cent confidence interval of the loss probability is less than 20 per cent of the point estimate. The simulations for the admission rule of Jamin *et al.*, which is computationally less demanding than the LD admission rule, are terminated when the 90 per cent confidence interval of the loss probability is less than 10 per cent of the point estimate. The 90 per cent confidence intervals for the loss probability P_{loss} are plotted in Figure 3. We do not plot the confidence intervals for the average number of connections since these confidence intervals are much tighter. In fact, the widths of the 90 per cent confidence intervals for J_{avg} are less than 0.1 connection and do not show up on the plots. We observe that the load-loss points of our large-deviations-based admission rule are below the load–loss frontiers of the admission rule of Jamin *et al.* Considering Figure 3(a) we see that for $\varepsilon = 10^{-6}$ our admission rule admits on average 178 connections and the loss probability is 1.1×10^{-6}. For the same average link utilization, i.e. for 178 connections, the admission rule of Jamin *et al.* gives a loss probability of roughly 6×10^{-6}. Comparing the Plots 3(a) and 3(b) we observe that the gap in performance is wider when the burstier, unsmoothed video streams are multiplexed. We see from Figure 3(b) that for a given QoS requirement the LD admission rule admits on average 8 unsmoothed lambs video streams more. These numerical results indicate that by measuring individual bursts and capturing the variability of the arrival process in the moment generating function, the LD admission rule utilizes the available link capacity more efficiently.

We now compare the computational complexities of the measured sum approach of Jamin *et al.* and the LD admission rule. Both approaches rely on aggregate traffic measurements; there are no per-flow measurements required. The measured sum approach uses the very simple window measurement mechanism. It requires only one counter to determine the largest arrivals (per averaging period) in the measurement window. For the LD approach we need to maintain the measurements x_i, $i = 1, \ldots, M_{\mathrm{eff}}$, as well as the peak rates of the admitted streams y_i, $i = 1, \ldots, M$, in memory; the LD approach thus has a larger storage cost. Note that the storage cost is independent of the number of ongoing flows and the link capacity. An admission test in the measured sum approach involves checking whether (9) holds; a negligible computational effort. An admission test in the LD approach involves finding the s^* that satisfies (7) and then evaluating $P_{\mathrm{loss}}^{\mathrm{est}}$ (6). This takes roughly 250 ms on a 1997 workstation in the simulated scenario, where $M_{\mathrm{eff}} = 1912$ (for $\tau_p = 120$, see Table III). Most of the computational effort (roughly 245 ms) goes into calculating s^*, which we have done with Newton's method [37]. (Based on the observation that s^* typically changes only slightly from one admission test to the next, for each admission test we start Newton's method with the s^* computed in the previous admission test. We note that a detailed analytical study of the complexity of the LD approach is beyond the scope of this paper, therefore we provide actual timing measurements here.) A simple speed-up of the admission test works as follows. s^* is pre-computed (for a typical c_k^*) periodically, every second, say, and also after a new connection has been accepted or an established connection has terminated. When a connection requests establishment the admission decision is based on the $P_{\mathrm{loss}}^{\mathrm{est}}$ evaluated with the most recently pre-computed s^*; this admission test takes roughly 5 ms. Due to the convexity of the exponent in (6), which dominates $P_{\mathrm{loss}}^{\mathrm{est}}$, a suboptimal s^* gives a larger $P_{\mathrm{loss}}^{\mathrm{est}}$, thus leading to conservative admission decisions [38]. However, we observed in our simulations that the speed-up works very well. It reaches an incorrect decision (by rejecting a connection that with a current s^* would have been accepted) in less that 1 out of 1000 admission tests.

We conclude the review of the measurement-based admission rule of Jamin *et al.* by briefly discussing an enhancement of this admission rule that is due to Casetti *et al.* [12]. As we have seen in Figure 3 the performance of the admission rule of Jamin *et al.* depends greatly on the tuning of its parameters—the measurement window W and the target utilization v. The appropriate tuning of these parameters is still an area of ongoing research. Casetti *et al.* propose a feedback control mechanism that automatically tunes the measurement window W. Instead of the target utilization v their adaptive measurement-based admission rule takes a target loss probability as input. The measurement window W is dynamically adjusted based on measurements of the aggregate arrivals and the losses at the multiplexer. Roughly speaking, the adjustment works as follows: see Reference [12] for more details. If the aggregate arrivals are above a trigger value, which is computed internally by the feedback algorithm, and the measured loss probability exceeds the target loss probability the measurement window W is increased. This makes the admission rule more conservative and results in the acceptance of fewer new connections. Conversely, if the aggregate arrivals are below the trigger value and the measured loss probability is smaller than the target loss probability the measurement window W is decreased. This results in a less conservative admission rule. Casetti *et al.* demonstrate that their adaptive measurement-based admission rule works reasonably well for large loss probabilities on the order of 10^{-3} or larger. However, their adaptive admission rule fails for smaller loss probabilities as it is difficult to measure smaller loss probabilities with sufficient statistical significance.

3.2. Chernoff bound approach

We next review the work on measurement-based admission control by Gibbens *et al.* [13] and Gibbens and Kelly [14]. Gibbens *et al.* study different variations of Chernoff bound-based admission rules. They assume that the multiplexed traffic streams are on–off streams. Recall from Section 2 that $U_j(t)$, $j = 1, \ldots, J$, denote the rate of connection-j traffic entering the multiplexer at time t. Let U_j, $j = 1, \ldots, J$, denote the associated steady-state random variables. Recall that c_j^* denotes the peak rate of connection-j traffic entering the multiplexer, and let r_j denote the average rate of connection j (which is measured). Gibbens *et al.* assume that connection j transmits at rate c_j^* with probability r_j/c_j^* and is silent with probability $1 - r_j/c_j^*$, i.e. $P(U_j = c_j^*) = r_j/c_j^*$ and $P(U_j = 0) = 1 - r_j/c_j^*$. Let $\mu_{U_j}(s)$ denote the logarithmic moment generating function of U_j defined as $\mu_{U_j}(s) := \ln E[\mathrm{e}^{sU_j}]$. Clearly

$$\mu_{U_j}(s) = \ln\left[1 - \frac{r_j}{c_j^*} + \frac{r_j}{c_j^*}\mathrm{e}^{sc_j^*}\right] \tag{10}$$

The Chernoff bound [39] gives an upper bound on the probability that the sum of the independent random variables U_j exceeds the link capacity C:

$$P\left(\sum_{j=1}^{J} U_j > C\right) \leqslant \mathrm{e}^{s(\sum_{j=1}^{J}(1/s)\mu_{U_j}(s) - C)} \tag{11}$$

Notice from the exponent of the Chernoff bound that $(1/s)\mu_{U_j}(s)$ can be ascribed the meaning of bandwidth; in fact the term $(1/s)\mu_{U_j}(s)$ is commonly referred to as effective bandwidth of connection j [40]. Gibbens *et al.* view the effective bandwidth as function of the average rate r_j and bound this concave function of r_j by a tangent of the function. Depending on the point at which the tangent of the effective bandwidth function is constructed different admission rules are derived. Constructing a tangent at $r_j = 0$, for instance, gives the bound

$$\frac{1}{s}\,\mu_{U_j}(s) \leqslant \frac{\mathrm{e}^{sc_j^*} - 1}{sc_j^*}\,r_j$$

which is easily verified by calculating the derivative of μ_{U_j} (10) with respect to r_j at $r_j = 0$. Substituting this bound into the exponent of the Chernoff bound (11) and requiring that the Chernoff bound be less than some miniscule ε gives the admission control condition

$$\sum_{j=1}^{J} r_j \mathrm{e}^{sc_j^*} \leqslant C$$

see Appendix A of [14] for details. Interestingly, the QoS parameter ε does not appear in the admission control condition. The condition is tuned with the space parameter s of the moment generating function. The appropriate setting of the tuning parameter s is still the subject of ongoing research. Also, note that this admission control condition requires measurements of the average arrival rates r_j, $j = 1, \ldots, J$, of each individual connection j. In practice, however, it is very difficult to measure per-connection rates accurately and cost efficiently. Gibbens *et al.* therefore propose to assign individual traffic streams based on their peak rate c_j^* to a small number of traffic classes and measure the aggregate arrivals for each class. Gibbens *et al.* employ a simple point sample measurement mechanism. They measure the aggregate arrivals per class over an averaging period T and base admission decisions on the most recent measurement only; older measurements are not considered. (Note that this is equivalent to setting $W = T$ in the measurement window scheme of Jamin *et al.*) Thus, in the simplest case of only one class the admission control condition is

$$(rT + x_1)e^{sc^*T} \leqslant CT$$

where r is the leaky bucket rate of the connection requesting establishment, x_1 denotes the measured aggregate arrivals in the most recent averaging period of length T, and c^* is the peak rate of the class. Gibbens *et al.* study this admission rule, which is derived from the tangent of the effective bandwidth function at $r_j = 0$, in the context of a decision theoretic framework for admission control in Reference [13]. In Reference [14] Gibbens and Kelly derive a number of variations of this admission rule by considering tangents at different points of the effective bandwidth function.

Jamin and Shenker [17] compare the performance of their approach to measurement-based admission control with the different variations of the approach of Gibbens *et al.* They find in their extensive simulations that the load–loss frontiers of both approaches coincide. Hence both approaches have the same performance. This surprising result is due to structural similarities of the two approaches, which are studied in detail by Jamin and Shenker [17]. Roughly speaking, both approaches rely on the mean of the measured arrivals (higher moments are not considered). A constant, which represents the new connection, is added to the mean and the new connection is accepted if this sum is less than the link capacity multiplied by a constant.

Our large-deviations-based approach to measurement-based admission control is fundamentally different from the approach of Gibbens *et al.* in that we do not assume on-off traffic. Moreover, we do not bound the effective bandwidth function, which is derived from the logarithmic moment generating function, but use the actual logarithmic moment generating function in our admission rule. The LD admission rule thus takes the mean as well as the higher moments of the measured aggregate arrivals into account.

3.3. Hoeffding bound approach

Floyd [15] studies measurement-based admission control based on the Hoeffding bound [41, 42]. The Hoeffding bound is a Chernoff-style bound for sums of bounded, independent random variables. Recall that U_j, $j = 1, \ldots, J$, are steady-state random variables denoting the rate at which connection-j traffic enters the multiplexer. Furthermore, recall that c_j^* and r_j denote the peak rate and average rate of connection j. The Hoeffding bound gives an upper bound on the probability that the sum of the random variables U_j exceeds the sum of the average rates r_j by some positive constant δ:

$$P\left(\sum_{j=1}^{J} U_j > \sum_{j=1}^{J} r_j + \delta\right) \leqslant \mathrm{e}^{-2\delta^2/\sum_{j=1}^{J} c_j^{*2}}$$

Floyd derives an admission control condition from the Hoeffding bound by requiring that the bound be less than some miniscule QoS parameter ε. Clearly

$$\mathrm{e}^{-2\delta^2/\sum_{j=1}^{J} c_j^{*2}} \leqslant \varepsilon$$

for

$$\delta \leqslant \sqrt{\frac{\ln(1/\varepsilon)}{2} \sum_{j=1}^{J} c_j^{*2}}$$

Noting furthermore that

$$P\left(\sum_{j=1}^{J} U_j > C\right) \leqslant P\left(\sum_{j=1}^{J} U_j > \sum_{j=1}^{J} r_j + \delta\right)$$

for $\sum_{j=1}^{J} r_j + \delta \leqslant C$, Floyd arrives at the admission control condition

$$\sum_{j=1}^{J} r_j + \sqrt{\frac{\ln(1/\varepsilon)}{2} \sum_{j=1}^{J} c_j^{*2}} \leqslant C \tag{12}$$

We note that Gibbens and Kelly derive this condition as a special case of their Chernoff bound based admission control conditions. They obtain (12) by bounding the effective bandwidth of an on–off stream $(1/s)\mu_{U_j}(s)$ (where $\mu_{U_j}(s)$ is given by (10)) by a tangent of slope one at $r_j = 1/s - c_j^*/(\mathrm{e}^{sc_j^*} - 1)$; see Appendix A3 of Reference [14] for details. Measurement-based admission control based on the Hoeffding bound is also studied by Brichet and Simonian [16]. They derive a tighter bound on the effective bandwidth of an on-off stream $(1/s)\mu_{U_j}(s)$ by considering a series expansion of $\mu_{U_j}(s)$ for small s.

Floyd [15] as well as Brichet and Simonian [16] employ an exponential weighted moving average measurement mechanism. Let $\hat{x}$ denote the estimate of the aggregate arrivals in an averaging period of length T, i.e. $\hat{x}$ denotes the estimate of $T\sum_{j=1}^{J} r_j$. The estimate $\hat{x}$ is updated using the recursion

$$\hat{x} = (1 - \omega)\hat{x} + \omega x_i$$

where ω is the weight used to tune the measurement mechanism and x_i denotes the aggregate arrivals in the just expired averaging period of length T. Now suppose that a new connection $J + 1$ with peak rate c_{J+1}^* requests establishment. The new connection is accepted if

$$\hat{x} + T\sqrt{\frac{\ln(1/\varepsilon)}{2} \sum_{j=1}^{J} c_j^{*2}} + c_{J+1}^* T \leqslant CT$$

and rejected otherwise.

Jamin and Shenker [17] consider the measurement-based admission control approach employing the Hoeffding bound and the exponential weighted moving average measurement mechanism in their simulation studies. They find that the load–loss frontier of the Hoeffding bound approach coincides with the load–loss frontiers of the Jamin *et al.* [9] and Gibbens *et al.*

[13,14] approaches. Hence, these three approaches have the same performance; this is due to their structural similarities [17]. Jamin and Shenker also note that the Hoeffding bound admission rule performs far off target; setting the QoS parameter to $\varepsilon = 0.99$, for instance, resulted in an actual loss probability of 10^{-5} in their experiments.

3.4. Time-scale decomposition approach

Finally, we provide a brief review of the work on measurement-based admission control by Grossglauser and Tse [18,8]. Roughly speaking, their approach is to estimate mean and variance of the arrivals from the measurements and estimate the loss probability at the node using the normal approximation [39]. In Reference [18] they conduct an extensive analysis of this normal approximation approach for the case when per-flow measurements are available. In Reference [8] they extend the analysis to the more practicable case when only aggregate traffic measurements are available. We focus on this latter case, which is also referred to as time-scale decomposition approach, in this review. In their analysis Grossglauser and Tse identify a critical time-scale $\tilde{T}_{\mathrm{h}} = T_{\mathrm{h}}/\sqrt{J_{\mathrm{cap}}}$, where T_{h} denotes the average holding time (lifetime) of a connection and J_{cap} denotes the capacity of the multiplexer (which is the multiplexer rate divided by the average rate of a typical connection). The key idea of the time-scale decomposition approach is to decompose the aggregate arrival process (i.e. the measurements x_i over time) into a low-frequency component and a high-frequency component. The low-frequency component is obtained by passing the measurements through a low-pass filter with a cutoff frequency of $1/\tilde{T}_{\mathrm{h}}$, while the high-frequency component is obtained through a high-pass filter with a cutoff frequency of $1/\tilde{T}_{\mathrm{h}}$. The low-frequency component, which tracks the slow time-scale fluctuations of the arrival process, is used to estimate the mean of the arrivals, while the high-frequency component, which tracks the fast time-scale fluctuations, is used to estimate the variance of the arrivals. Formally, let g_i, $i \geqslant 1$, denote the impulse response of the low-pass filter. Grossglauser and Tse propose to use a geometrically decaying impulse response

$$g_i = \frac{1}{\tilde{T}_{\mathrm{h}}}\left(1 - \frac{1}{\tilde{T}_{\mathrm{h}}}\right)^{i-1}$$

Suppose that in slot k a new connection with smoother rate (peak rate) c_k^* requests establishment. Recall that x_i, $i = 1, \ldots, M_{\mathrm{eff}}$, denote the measured aggregate arrivals in slot $k - i$. Let J_i, $i = 1, \ldots, M_{\mathrm{eff}}$, denote the number of ongoing streams in slot $k - i$. Let m_U denote the estimate of the average arrivals per connection in a slot. This estimate is obtained by averaging the measured aggregate arrivals over the ongoing connections and convolving the measurements with the impulse response of the low-pass filter:

$$m_U = \sum_{i=1}^{M_{\mathrm{eff}}} \frac{x_i}{J_i} g_i$$

Furthermore, let σ_U^2 denote the estimate of the variance of the arrivals per connection in a slot. For details on how this estimate is obtained we refer the reader to Reference [8]. The new connection with peak rate c_k^* is accepted if

$$Q\left(\frac{(C - c_k^*)T - J_1 m_U}{\sqrt{J_1 \sigma_U^2}}\right) \leqslant \varepsilon$$

where $Q(\cdot)$ denotes the complementary cumulative distribution function of a standard normal random variable, which is given by

$$Q(a) = \frac{1}{\sqrt{2\pi}} \int_a^{\infty} \mathrm{e}^{-y^2/2}\, \mathrm{d}y$$

Grossglauser and Tse define the loss probability as the long-run fraction of time during which loss occurs, that is, the long-run fraction of slots with aggregate arrivals exceeding CT. We denote this loss measure by $P_{\text{loss}}^{\text{time}}$ to distinguish it from the loss measure P_{loss} (3), which is the long-run fraction of information (bits) lost. Formally, $P_{\text{loss}}^{\text{time}}$ is defined as

$$P_{\text{loss}}^{\text{time}} = E\left[\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} 1(X_n > CT)\right]$$

The time-scale decomposition approach is designed to ensure that $P_{\text{loss}}^{\text{time}} \leqslant \varepsilon$. We evaluate the time-scale decomposition approach using the simulation set-up of Section 2. We run simulations for different ε and record both loss measures, $P_{\text{loss}}^{\text{time}}$ and P_{loss}. The results are reported in Table V. Part (a) of the table gives the results for the case where the video streams are passed through smoothers with a maximum smoothing delay of 10 frame periods before being multiplexed. Part (b) gives the results for the case where the unsmoothed video streams are multiplexed. We observe from the table that the time-scale decomposition approach gives good on-target performance. We also observe that P_{loss} is typically almost two orders of magnitude smaller than $P_{\text{loss}}^{\text{time}}$. This is because only the fraction $(X_n - CT)/CT$ of bits is lost in slots with aggregate arrivals exceeding the link capacity CT. The load–loss points (J_{avg}, P_{loss}) (with 90 per cent confidence intervals for P_{loss}) are plotted in Figure 3. We observe from the plots that the load–loss frontier of the time-scale decomposition approach lies between the load–loss frontiers of the approach of Jamin *et al.* and our LD approach. This indicates that by taking the first two moments of the arrival process into consideration the time-scale decomposition approach can accommodate more connections (on average) than the other reviewed approaches, which take only the first moment into consideration.

We conclude the discussion of the time-scale decomposition approach by noting that the time-scale decomposition approach has no explicit tuning parameters. However, it requires knowledge of the average lifetime of the connections, T_{h}, and the average rate of a typical connection (to

Table V. Evaluation of time-scale decomposition approach of Grossglauser and Tse.

ε	10^{-1}	10^{-2}	10^{-3}	10^{-4}
(*a*) *Maximum smoother delay = 10 frame periods*				
J_{avg}	211	197	185	180
$P_{\text{loss}}^{\text{time}}$	1.0×10^{-1}	1.2×10^{-2}	1.5×10^{-3}	1.6×10^{-4}
P_{loss}	3.8×10^{-3}	3.2×10^{-4}	2.8×10^{-5}	3.6×10^{-6}
(*b*) *No smoothing*				
J_{avg}	205	187	174	163
$P_{\text{loss}}^{\text{time}}$	1.0×10^{-1}	1.3×10^{-2}	1.7×10^{-3}	2.7×10^{-4}
P_{loss}	5.2×10^{-3}	4.8×10^{-4}	5.4×10^{-5}	7.1×10^{-6}

determine the capacity J_{cap}). The time-scale decomposition approach is therefore an attractive admission control approach if these two parameters are fairly well known. If these parameters are unknown or change over time, however, the time-scale decomposition approach faces tuning problems similar to the other approaches. As for the computational complexity of the time-scale decomposition approach, we note that the calculation of the variance σ_U^2 requires time consuming convolutions; see Reference [8] for details. The admission test could be sped up by pre-computing σ_U^2 periodically or by bypassing the time consuming convolutions with fast Fourier transform techniques.

At this juncture we note an important study by Breslau *et al.* [10]. They compare the performance of a number of measurement-based admission rules at a *buffered* multiplexer. Among other approaches they consider the measured sum approach [9] and the large deviation approach for a buffered multiplexer [21]. They find in their simulations that the load–loss frontiers of all approaches coincide. This means that all approaches—when tuned optimally—achieve the same average link utilization for a given loss probability requirement (or incur the same loss probability for a given average link utilization) at a buffered multiplexer.

Our results indicate that this is not the case at a *bufferless* multiplexer. We find that there are differences in the performance that measurement-based admission control rules can achieve at a bufferless multiplexer. However, as we see from Figure 3, these differences are not very large; generally less than half an order of magnitude in loss probability or less than 5 per cent in average link utilization. We conjecture that these inherent differences of the measurement-based admission control rules are 'smoothed' out by the buffer used in the simulations in Reference [10] and were therefore not observed in that study.

4. COMPARISON WITH TRADITIONAL ADMISSION CONTROL

In this section we compare measurement-based admission control with traditional admission control that bases admission decisions on *a priori* traffic characterizations.

4.1. Adversarial admission control

First, we consider an admission rule that takes leaky bucket traffic characterizations as input and assumes that the connections are adversarial to the extent permitted by the leaky bucket characterizations, i.e., transmit worst-case on–off traffic. Suppose that the connection-j traffic at the smoother output is characterized by the traffic constraint function $\mathscr{E}_j(t)$, that is, the amount of traffic leaving smoother j over an interval of length t is less than $\mathscr{E}_j(t)$. Specifically, suppose that $\mathscr{E}_j(t) = \min(c_j^* t, \sigma_j + \rho_j t)$, that is, the output of smoother j is constrained by the smoother rate (peak rate) c_j^* and a single leaky bucket (σ_j, ρ_j), where σ_j is the maximum burst size and ρ_j bounds the long-term average rate of connection j. The adversarial admission rule assumes that each connection transmits worst-case on–off traffic [2,5], that is, connection j transmits at rate c_j^* with probability ρ_j/c_j^* and is silent with probability $1 - \rho_j/c_j^*$. Recall that $U_j, j = 1, \ldots, J$, are steady-state random variables denoting the rate at which connection-j traffic enters the multiplexer. The adversarial assumption is that $P(U_j = c_j^*) = \rho_j/c_j^*$ and $P(U_j = 0) = 1 - \rho_j/c_j^*$. The logarithmic moment generating function of U_j, defined as $\mu_{U_j}(s) := \ln E[\mathrm{e}^{sU_j}]$, is clearly

$$\mu_{U_j}(s) = \ln\left[1 - \frac{\rho_j}{c_j^*} + \frac{\rho_j}{c_j^*}\mathrm{e}^{sc_j^*}\right]$$

Let X denote the sum of the random variables $U_j, j = 1, \ldots, J$, i.e., $X = \sum_{j=1}^{J} U_j$. Furthermore, let $\mu_X(s)$ and m_X denote the logarithmic moment generating function and mean of X. Assuming that the connections generate traffic independently, i.e. $U_j, j = 1, \ldots, J$, are mutually independent, we obtain $\mu_X(s) = \sum_{j=1}^{J} \mu_{U_j}(s)$.[§] The expected fraction of traffic lost at the multiplexer is

$$P_{\text{loss}} = \frac{E[(X - C)^+]}{E[X]} \tag{13}$$

The Large Deviation approximation of (13) is given by Reference [1]

$$P_{\text{loss}} \approx \frac{1}{m_X s^{\star 2} \sqrt{2\pi \mu_X''(s^\star)}} \mathrm{e}^{-s^\star C + \mu_X(s^\star)} \tag{14}$$

where s^* satisfies

$$\mu_X'(s^*) = C$$

A set of J connections is permissible on a link with capacity C if the loss probability P_{loss} (14) is less than some miniscule QoS parameter ε.

For the numerical work in this section we use again the '*Silence of the Lambs*' (lambs) trace. We obtain the traffic constraint function of the lambs trace $\mathscr{E}_{\text{lambs}}(t)$ by following the procedure described in Reference [5]. We give here only a brief outline of this procedure and refer the reader to Reference [5] for details. The first step is to compute the empirical envelope, which is the tightest traffic constraint function of the video sequence. The empirical envelope is however not necessarily concave. It is therefore bounded by a piecewise linear function so that the traffic constraint function can be represented by a cascade of leaky buckets. Given this leaky bucket characterization and a maximum delay in the smoother, which we set again to 10 frame periods (= 10/24 s), we obtain the smoother rate $c^*_{\text{lambs}} = 731.6$ kbit/s by applying the theory developed in Reference [5]. The lambs traffic at the smoother output is characterized by this smoother rate (peak rate) and the leaky bucket $(\sigma_{\text{lambs}}, \rho_{\text{lambs}}) = (3.16$ Mbyte, 193.2 kbit/s). For the following numerical evaluation we set the multiplexer rate to $C = 45$ Mbps and assume that all traffic streams are independent lambs video streams. We vary the number of connections J and compute the loss probability for each J using the LD approximation (14). The results are plotted as the solid line (labeled 'adversarial') in Figure 4. We defer the discussion of this result to the end of this section.

4.2. Histogram-based admission control

We next consider an admission rule that is specifically designed for prerecorded sources [38]. This admission rule bases admission decisions on the marginal distribution of the sources' traffic. For video traffic the histogram of the frame sizes is used to compute the logarithmic moment generating function of the video stream. Recall from Section 2 that $f_n(j)$, $n = 1, \ldots, N$, denotes the frame size of the nth frame of video j in bits. Also recall from Section 2 that the transmission of a frame is spread out over one frame period of length $1/F$, i.e. the frame $f_n(j)$ is transmitted at rate $f_n(j) \cdot F$ over the interval $[(n-1)/F, n/F]$. Let $u_n(j)$, $n = 1, \ldots, N$, denote the smoothed frame size

[§] Note that we have redefined $\mu_X(s)$ in this section. In Section (2) $\mu_X(s)$ is defined as the estimate of the logarithmic moment generating function of X_k, the *amount* of traffic (in bit) arriving in slot k. In this section $\mu_X(s)$ is defined as the logarithmic moment generating function of X, the *rate* of arriving traffic (in bit/s).

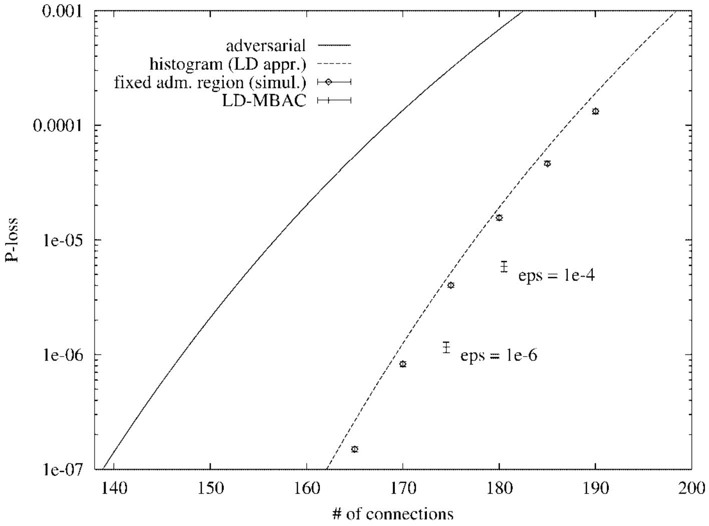

Figure 4. Comparison of measurement-based admission control with traditional admission control.

trace obtained by simulating the transmission of the original frame size trace $f_n(j)$, $n = 1, \ldots, N$, through the buffered smoother with peak rate c_j^* (see Figure 1). The logarithmic moment generating function of the smoothed video stream j is calculated directly from the smoothed frame size trace:

$$\mu_{U_j}(s) = \ln \frac{1}{N} \sum_{n=1}^{N} e^{sFu_n(j)} \tag{15}$$

A set of J connections is admitted if the loss probability (14) is less than some prespecified QoS parameter ε. For the numerical evaluation we assume that all multiplexed traffic streams are smoothed lambs video streams. The dashed line (labelled 'histogram') in Figure 4 is the load–loss curve of this admission rule, which is obtained by computing the loss probability (14) for a range of link utilizations.

We also verify the accuracy of the LD approximation through simulation. For this purpose we use the simulation program used to evaluate the measurement-based admission rules in Sections 2 and 3. Instead of employing any of the studied measurement-based admission rules, we fix a maximum number of admissible streams, $J_{\max}$. A connection requesting establishment is accepted if there are currently less than $J_{\max}$ connections in progress, and rejected otherwise. As before, all of the streams are generated from the lambs trace. Each stream has its own independent random phase, which is uniformly distributed over [1, N]. The lifetime of the streams is fixed at $N = 40\,000$, and the lambs trace is wrapped around to generate the streams. We set the connection inter arrival time to zero, thus there are always $J_{\max}$ connections in progress. We run the simulation until the width of the 90 per cent confidence interval of the loss probability is less than 10 per cent of the point estimate. The results are plotted in Figure 4 (labeled 'fixed adm. region (simul.)').

The figure also shows the load-loss points of our LD approach to measurement-based admission control (LD-MBAC). These points are obtained by running simulations for $\varepsilon = 10^{-4}$

and 10^{-6}. The parameters of the admission rule are set to $\tau_p = 120$ and $\tau_q = 1250$ for these simulations. The time between call arrivals is exponentially distributed with a mean of 67 frame periods. Furthermore, the lifetime of each stream is fixed at $N = 40\,000$ frame periods to ensure a fair comparison with the other admission rules. (The traffic constraint function obtained through the procedure described in Reference [5] gives the tightest leaky bucket characterization of the full-length video segment; for shorter segments, however, the characterization may be loose.)

Several points are noteworthy about Figure 4. Firstly, note that the $J_{\max}$-simulation ('fixed adm. region (simul.)') verifies the accuracy of the histogram admission rule. We observe that the histogram admission rule is a little too conservative, but generally very accurate. Secondly, we observe that the adversarial admission rule, which assumes worst-case on–off traffic, results in low link utilizations. Note that by following the procedure described in Reference [5] we have obtained the tightest leaky bucket characterization of the prerecorded lambs trace. The difference in link utilization (horizontal distance) between the 'adversarial' and 'histogram' curves in Figure 4 therefore gives an indication of the conservatism of the assumption of adversarial on–off traffic. With a QoS parameter of $\varepsilon = 10^{-6}$, for instance, the adversarial admission rule admits 147 lambs video streams while the histogram rule admits 169 streams and the measurement-based admission rule admits on average 174.5 streams. In the case of live video transmission where one has to resort to loose leaky bucket characterizations the link utilization with adversarial admission control is even lower, while measurement-based admission control still achieves high link utilizations.

The third noteworthy observation is that measurement-based admission control outperforms histogram-based admission control, which has perfect knowledge of the marginal distribution of the streams' traffic. This can be intuitively explained as follows. The histogram-based admission rule bases admission decisions on the connections' logarithmic moment generating functions $\mu_{U_j}(s)$ (15), which characterize the connections' traffic over their entire lifetime. A new connection is accepted if the long-run fraction of traffic lost (due to excursions of the aggregate arrival process X above the link capacity CT) is less than ε. Most of the time, however, the aggregate arrival process X is below the threshold CT and the slack capacity $CT - X$ is wasted. Measurement-based admission control bases admission decisions on measurements of the aggregate arrival process X. It admits new connections when slack capacity is available. Conversely, measurement-based admission control stops the acceptance of new connections when the aggregate arrivals are close to the link capacity or even exceed the link capacity. It does not accept any new streams until departing connections have created slack capacity. Measurement-based admission control thus utilizes the link capacity efficiently by taking advantage of the connection arrival and departure dynamics. Note however, that measurement-based admission control is bound to fail when the connection arrival and departure times collude, that is, when the connections arrive roughly at the same time and have identical lifetimes. In the worst-case scenario when all connections arrive in the same time slot the LD-MBAC rule bases admission decisions on the connections' peak rate specification, i.e. it admits 61 smoothed lambs streams ($= C/c^{*}_{\text{lambs}}$). Traditional admission control, on the other hand, achieves the link utilizations shown in Figure 4 irrespective of the connection arrival and departure dynamics.

5. CONCLUSION

In this paper we have studied measurement-based admission control for unbuffered multiplexers. We have discussed a large deviations approach to measurement-based admission control. We have provided an extensive review of the existing literature on measurement-based admission control. We found in our simulations with MPEG-1 encoded videos that the LD admission rule compares favourably with the admission rules in the existing literature. Finally, we compared measurement-based admission control with traditional admission control, which relies on *a priori* traffic characterizations. Our numerical work indicates that measurement-based admission control can achieve significantly higher link utilizations.

In our current research we are addressing the parameter tuning problem. We are investigating the use of feedback control to automatically tune the parameters of the LD admission rule. Our numerical work suggests that the LD admission rule performs on-target and achieves high link utilizations when the estimated loss probability $P_{\text{loss}}^{\text{est}}$ oscillates with small amplitudes around the target loss probability ε (see for instance Figure 2(b)). We are therefore studying feedback control policies that tune the parameters of the LD admission rule based on $P_{\text{loss}}^{\text{est}}$; in particular integral control policies that strive to minimize the area under the $P_{\text{loss}}^{\text{est}}$ curve. This approach avoids the difficult problem of measuring the actual, miniscule losses at the multiplexer.

Another avenue for future research is to study the combination of measurement-based admission control with traditional admission control. The goal is to develop a hybrid admission rule that strives to achieve the high link utilizations of measurement-based admission control and at the same time guards against gross mistakes by the measurement-based admission rule due to incorrect parameter settings or colluding connection arrivals/departures.

ACKNOWLEDGEMENTS

I thank Nick Duffield for pointing me to the work of Dublin's Applied Probability Group on measurement-based admission control. I am also grateful to Cormac Walsh for clarifying this work. I am grateful to Matthias Grossglauser for clarifying the filtering mechanisms used in the time-scale decomposition approach [8].

REFERENCES

1. Roberts J, Mocci U, Virtamo J (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: Berlin, 1996.
2. Elwalid A, Mitra D, Wentworth RH. A new approach for allocating buffers and bandwidth to heterogeneous regulated traffic in an ATM node. *IEEE Journal on Selected Areas in Communications* 1995; **13**(6):1115–1127.
3. LoPresti F, Zhang Z, Towsley D, Kurose J. Source time scale and optimal buffer/bandwidth trade-off for regulated traffic in a network node. *IEEE/ACM Transactions on Networking* 1999; **7**(4):490–501. A shorter version has appeared in *Proceedings of IEEE Infocom*, Kobe, Japan, April 1997.
4. Rajagopal S, Reisslein M, Ross KW. Packet multiplexers with adversarial regulated traffic. *Proceedings of IEEE Infocom '98*, San Francisco, CA, April 1998; 347–355.
5. Reisslein M, Ross KW, Rajagopal S. Guaranteeing statistical QoS to regulated traffic: the single node case. *Proceedings of IEEE Infocom '99*, New York, NY, March 1999; 1060–1071.
6. Reisslein M, Ross KW, Rajagopal S. Guaranteeing statistical QoS to regulated traffic: the multiple node case. *Proceedings of 37th IEEE Conference on Decision and Control (CDC)*, Tampa, FL, December 1998; 531–538.
7. Roberts JW. Realizing quality of service guarantees in multiservice networks. In *Performance and Management of Complex Communication Networks—Proceedings of IFIP Conference PMCCN '97*, Tsukuba, Japan, Hasegawa T, Takagi H, Takahashi Y. (eds). Chapman & Hall: London, November 1997.
8. Grossglauser M, Tse D. A time-scale decomposition approach to measurement-based admission control. *Proceedings of IEEE Infocom '99*, New York, NY, March 1999; 1539–1547.
9. Jamin S, Danzig PB, Shenker SJ, Zhang L. A measurement-based admission control algorithm for integrated services packet switched networks (extended version). *IEEE/ACM Transactions on Networking* 1997; **5**(1):56–70.

10. Breslau L, Jamin S, Shenker S. Comments on the performance of measurement-based admission control algorithms. *Proceedings of IEEE Infocom*, Tel Aviv, Israel, March 2000.
11. Jamin S, Danzig PB, Shenker SJ, Zhang L. A measurement-based admission control algorithm for integrated services packet switched networks. *Proceedings of ACM SIGCOMM '95*, 1995; 2–13.
12. Casetti C, Kurose J, Towsley D. A new algorithm for measurement-based admission control in integrated services packet networks. *Proceedings of the Protocols for High Speed Networks Workshop*, October 1996.
13. Gibbens RJ, Kelly FP, Key PB. A decision theoretic approach to call admission control in ATM networks. *IEEE Journal on Selected Areas in Communications* 1995; **13**(6):1101–1114.
14. Gibbens RJ, Kelly FP. Measurement-based admission control. *Proceedings of 15th International Teletraffic Congress (ITC15)*, Washington, DC, June 1997; 879–888.
15. Floyd S. Comments on measurement-based admission control for controlled-load services. Technical Report ACIRI, July 1996.
16. Brichet F, Simonian A. Conservative gaussian models applied to measurement-based admission control. *Proceedings of IWQoS*, Napa, CA, May 1998.
17. Jamin S, Shenker S. Measurement-based admission control algorithms for controlled-load service: a structural examination. *Technical Report CSE-TR*-333–97, University of Michigan, April 1997.
18. Grossglauser M, Tse D. A framework for robust measurement-based admission control. *Proceedings of ACM SIGCOMM*, Cannes, France, September 1997; 237–248.
19. Lee TK, Zukerman M. Efficiency comparisons between different model-based and measurement-based connection admission control under heavy load. *Proceedings of IEEE Globecom*, Rio de Janeiro, Brazil, December 1999.
20. Duffield NG, Lewis JT, O'Connell N, Russell, R, Toomey F. Entropy of ATM traffic streams: a tool for estimating quality of service parameters. *IEEE Journal on Selected Areas in Communications* 1995; **13**(6):981–990.
21. Lewis JT, Russell R, Toomey F, McGurk B, Crosby S, Leslie I. Practical connection admission control for ATM networks based on on-line measurements. *Computer Communications* 1998; **21**(17):1585–1596.
22. Rácz A. How to build robust call admission control based on on-line measurements. *Proceedings of IEEE Globecom*, Rio de Janeiro, Brazil, December 1999.
23. Walsh C, Duffield NG. Predicting QoS parameters for ATM traffic using shape-function estimation. *Proceedings of 14th UK Teletraffic Symposium*, Manchester, U.K., March 1997.
24. McGurk B, Walsh C. Investigations of the performance of a measurement-based connection admission control algorithm. *Proceedings of 5th IFIP Workshop on Performance Modelling and Evaluation of ATM Networks*, Ilkley, U.K., July 1997.
25. Bovitch DD, Duffield NG. Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. *Queueing Systems* 1995; **20**:293–320.
26. Tse D, Grossglauser M. Measurement-based admission control: analysis and simulation. *Proceedings of Infocom '97*, Kobe, Japan, April 1997.
27. Zukerman M, Lee TK. A framework for real-time measurement-based connection admission control in multi-service networks. *Proceedings of IEEE Globecom '98*, Sydney, Australia, November 1998; 2983–2988.
28. Lee TK, Zukerman M, Cameron F. Utilization comparisons between several admission control schemes under realistic traffic conditions. *Proceedings of IFIP ATM '99*, Antwerpen, Netherlands, 1999.
29. Knightly EW, Qiu J. Measurement-based admission control with aggregate traffic envelopes. *Proceedings of 10th IEEE International Tyrrhenian Workshop on Digital Communications*, Ischa, Italy, September 1998.
30. Knightly EW, Zhang H. D-BIND: an accurate traffic model for providing QoS guarantees to VBR traffic. *IEEE/ACM Transactions on Networking* 1997; **5**(2):219–231.
31. Duffield NG. Asymptotic sampling properties of effective bandwidth estimation for admission control. *Proceedings of IEEE Infocom '99*, New York, NY, March 1999; 1532–1538.
32. Bahadur RR, Rao RR. On deviations of the sample mean. *Annals of Mathematical Statistics* 1960; **31**: 1015–1027.
33. Rose O. 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/.
34. Schruben LW. Detecting initialization bias in simulation output. *Operations Research* 1982; **30**:569–590.
35. Fishman GS. *Principles of Discrete Event Simulation*. Wiley: New York, 1991.
36. Reisslein M. Measurement-based admission control: a large deviations approach for bufferless multiplexers (extended version). *Technical Report*, GMD FOKUS, Berlin, Germany. Available at http://www.fokus.gmd.de/usr/reisslein, a shorter version has appeared in *Proceedings of the IEEE Symposium on Computers and Communications (ISCC)*, Antibes, France, July 2000.
37. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. *Numerical Recipes in C, The Art of Scientific Computing* (2nd edn). Cambridge University Press: Cambridge, MA, 1992.
38. Reisslein M, Ross KW. Call admission for prerecorded sources with packet loss. *IEEE Journal on Selected Areas in Communications* 1997; **15**(6):1167–1180.

39. Billingsley P. *Probability and Measure* (3rd edn). Wiley: New York, 1995.
40. Kelly FP. Notes on effective bandwidths. In *Stochastic Networks: Theory and Applications*, *Royal Statistical Society*, Kelly FP, Zachary S, Ziedins IB (eds). Lectures Note Series, vol. 4. Oxford University Press: Oxford, 1996; 141–168.
41. Hoeffding W. Probability inequalities for sums of bounded random variables. *American Statistical Association Journal* 1993; **58**:13–30.
42. Hofri M. *Probabilistic Analysis of Algorithms*. Springer: Berlin, 1987.

AUTHOR'S BIOGRAPHY

Martin Reisslein is an Assistant Professor in the Department of Electrical Engineering at Arizona State University, Tempe. He received the Dipl-Ing (FH) degree from the Fachhochschule Dieburg, Germany, in 1994, and the MSE degree from the University of Pennsylvania, Philadelphia, in 1996, both in Electrical Engineering. He received his PhD in Systems Engineering from the University of Pennsylvania in 1998. During the academic year 1994–1995 he visited the University of Pennsylvania as a Fulbright scholar. From July 1998 through October 2000 he was a scientist with the German National Research Centre for Information Technology (GMD FOKUS), Berlin. While in Berlin he was teaching courses on performance evaluation and networking at the Technical University Berlin. His research interests are in the areas of Internet Quality of Service, Wireless Networking, and Optical Networking. He is particularly interested in traffic management for multimedia services with statistical Quality of Service in the Internet and Wireless communication systems.