Measurement–Based Admission Control: A Large Deviations Approach for Bufferless Multiplexers

Martin Reisslein
GMD FOKUS
Kaiserin–Augusta–Allee 31
10589 Berlin, Germany
phone: +49-30-3463-7282
fax: +49-30-3463-8000
reisslein@fokus.gmd.de
http://www.fokus.gmd.de/usr/reisslein

Abstract

In order to provide Quality of Service (QoS) assurances networks perform call admission control before accepting a new connection. Rather than relying on *a priori* traffic descriptors (such as leaky buckets), which often poorly characterize the actual traffic, measurement–based admission control bases admission decisions on measurements of the actual traffic. In this paper we first develop a novel Large Deviations (LD) approach to measurement–based admission control for bufferless multiplexers. We then conduct simulation studies with traces of MPEG 1 encoded movies to compare the performance of the admission rules in the literature with that of the Large Deviations approach. We demonstrate that for bufferless multiplexing the LD approach achieves both higher link utilizations and smaller loss probabilities. 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.

Keywords

Bufferless Multiplexing, Large Deviations, Measurement–Based Admission Control, Statistical QoS.

I. Introduction

NETWORKS perform call admission control to ensure that the connections' Quality of Service (QoS) requirements are met. A 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 [26]. 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 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 [8], [21], [25], [24].

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.

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 [27]. We study the measurement–based admission rules within the smoothing/bufferless multiplexing framework [25], [24]. The key aspects of the smoothing/bufferless multiplexing framework are to (1) pass each connection's traffic through a *buffered* smoother (peak rate limiter) at the connection's input to the network, and (2) 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 for bufferless multiplexers. In this LD approach aggregate 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 $\epsilon = 10^{-6}$, and rejected otherwise.

Secondly, we compare the performance of the LD approach with that of some measurement–based admission rules in the literature through simulations with traces of MPEG 1 encoded movies. We consider (1) the time scale decomposition approach [14], which relies on estimates of the aggregate arrival stream's mean and variance, and (2) the measured sum approach [16],

which relies on estimates of the arrival stream's mean. We find that the studied approaches differ in their performance at a bufferless multiplexer. Our simulation results indicate that the LD approach achieves both higher link utilization and smaller loss probabilities than the time–scale decomposition approach. The time scale decomposition approach in turn performs better than the measured sum approach. These results are in contrast to a recent comparative study by Breslau *et al.* [3]; 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.

A. 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.* [15], [16] and Casetti *et al.* [5] study the so–called measured sum approach, which bases admission decisions on an estimate of the mean aggregate arrival rate. Gibbens *et al.* [12], [11] 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 [10] as well as Brichet and Simonian [4] 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 [13], [14] 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.

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 [7], [18] 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. Walsh *et al.* [30], [1] study an admission rule based on the shape function [2]. 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. Grossglauser

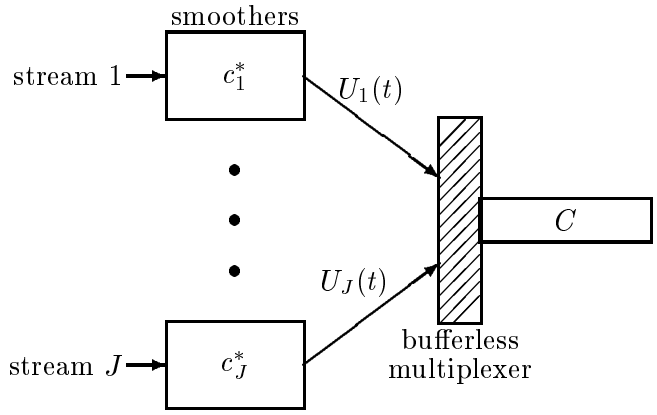


Fig. 1. The traffic of connection j is passed through a smoother that limits the peak rate to c_j^*. The smoothed traffic is then multiplexed onto a bufferless link with capacity C.

and Tse [29] 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).

We conclude this literature review by noting that Knightly and Qiu [19] study an admission rule for buffered multiplexers that does not involve estimates of moments of the arrival process. Their approach is to estimate maximal rates over different interval lengths (i.e., the maximal rate envelope [20]) from traffic measurements.

II. A Large Deviations Approach to Measurement–Based Admission Control

In this section we develop our Large Deviations (LD) approach to measurement–based admission control for bufferless multiplexers. We develop 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 for packet–scale queueing is needed.) 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 $U_j(t) \leq c_j^*$ $\forall t \geq 0$. Now regard the jth arrival process as a stochastic process. Let $(U_j(t),\ t \geq 0)$ 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 $(X(t),\ t \geq 0)$ denote the aggregate arrival process. The expected long–run fraction of traffic lost due to link overflow is

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

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

In practical systems 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.,

$$X_n = \int_{nT}^{(n+1)T} X(t)dt.$$

For small T we can reasonably approximate:

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

This approximation is particularly good when the fluctuation of the aggregate arrival process $(X(t),\ t \geq 0)$ 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 [6].)

Throughout this paper we shall assume that the approximation (2) is exact. Substituting (2) into (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 can not 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 behavior. We propose to base admissions decisions on the measured aggregate arrivals in the past M slots, i.e., $M \geq 1$ 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 $P_{\text{loss}}^{\text{est}} \leq \epsilon$ 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 $k - M, \ldots, k - 1$:

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

Furthermore, let $\mu_X(s)$ denote the estimate of $\ln E[e^{sX_k}]$, the logarithmic moment generating function of X_k. Again, we compute $\mu_X(s)$ by averaging over the M latest measurements:

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

The LD approximation of (4) is given by [26]

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

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

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

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 $P_{\text{loss}} \leq \epsilon$. First, we estimate the logarithmic moment generating function of the aggregate arrival stream based on the measurements in the last M slots using (5). We then estimate $P_{\text{loss}}^{\text{est}}$ using the LD approximation (6) and admit connection k if $P_{\text{loss}}^{\text{est}} \leq \epsilon$, otherwise connection k is rejected.

We evaluate the measurement–based admission rule using traces from MPEG 1 encoded movies. Because of page limitations we give here only a brief outline of our simulation approach and refer the interested reader to [22] for details.

The numerical results reported in this paper were obtained with the *"Silence of the Lambs"* (lambs) trace [28]. The lambs trace has 40,000 frames, corresponding to about 28 minutes. The

TABLE I

EVALUATION OF BASIC MEASUREMENT–BASED ADMISSION RULE.

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

lambs trace has an average frame size of 8,048 bit, which corresponds to an average rate of 193.2 kbit/sec. The trace has a peak–to–mean ratio of 18.4 and is therefore considered very bursty. We convert the discrete frame size trace to a fluid flow. 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 [25], [24]. Each video stream is passed through a smoother before it enters the bufferless multiplexer. The smoother 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 [25] for details). We initially set the delay limit for all connections to 10 frame periods, i.e., 10/24 seconds. The corresponding smoother rate for the lambs trace is 731.6 kbit/sec. Throughout this paper we set the rate of the bufferless multiplexer to $C = 45$ Mbit/sec.

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 (=1/24 seconds). In the simulation calls arrive according to a Poisson process with rate 1 call/10 frame period. For each call we draw a random life time from an exponential distribution with a mean of 6,000 frame periods (= 250 seconds). With these parameters the link operates in constant overload. We estimate J_{avg} and P_{loss} using the method of batch means [9]. We run each simulation until the width of the 90 % confidence interval of the loss probability is less than 20 % 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.

In the first set of simulation experiments we set the QoS parameter to $\epsilon = 10^{-6}$ and run simulations for different values of M, the length of the measurement window. The results are reported in Table I; 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 $\epsilon = 10^{-6}$. We provide a more detailed analysis of the basic admission rule, which we can not include here because of page limitations, in [22]. In summary, we find that weighing

TABLE II

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}	$4.5 \cdot 10^{-5}$	$4.4 \cdot 10^{-5}$	$4.0 \cdot 10^{-5}$	$3.1 \cdot 10^{-5}$	$6.2 \cdot 10^{-5}$	$8.1 \cdot 10^{-5}$	$1.6 \cdot 10^{-4}$	$2.1 \cdot 10^{-4}$
M_{eff}	6,000	6,000	6,000	6,000	6,000	4,506	1,912	998

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.

A. 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 $0 \leq p_i \leq 1$ and $\sum_{i=1}^{M} p_i = 1$. Throughout this paper we use exponentially decaying weights:

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

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

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

As before, these estimates are used to compute $P_{\text{loss}}^{\text{est}}$ (4) and the connection requesting establishment is accepted if $P_{\text{loss}}^{\text{est}} \leq \epsilon$ 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 use set the length of the measurement window to $M = 6{,}000$. In order to avoid unnecessary computation we ignore measurements that are assigned weights less than 10^{-9}. We denote M_{eff} for the number 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 $M_{\text{eff}} \cdot 0.13$ msec to perform one admission test on a SUN ULTRA 10 workstation. (The computation may be sped up further by applying the techniques of [23].)

Table II 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 reader to [22] for a detailed analysis.

B. 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 $\mu_X(s)$. Especially when the measurement window is long and older measurements are assigned relatively large weights, the few samples including stream u have very little influence on $\mu_X(s)$. 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 $y_i,\ i = 1, \ldots, M$, denote the peak rates of the admitted streams in slots $k - i,\ i = 1, \ldots, M$. y_i is set to zero if no new stream was admitted in slot $k - i$. Let $q_i,\ i = 1, \ldots, M$, denote weights with $0 \leq q_i \leq 1$. Throughout this paper we use exponentially decaying weights

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

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

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

TABLE III

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

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 $P_{\text{loss}}^{\text{est}} \leq \epsilon$ 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 III. 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 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 $\epsilon = 10^{-6}$. Unless stated otherwise these tuning parameters are used for all numerical experiments in the remainder of this paper.

Figure 2 (see p. 19) 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 $1.875 \cdot 10^6$ bit/slot (= 45 Mbps $\cdot$ 1/24 sec), while incurring actual losses only once around slot time 63,500.

III. Comparison with other Measurement–based Admission Rules

In this section we compare the performance of the Large Deviations based admission rule with that of two admission rules in the literature. First, we consider the measured sum approach of Jamin *et al.* [16].

Measured Sum Approach

Jamin *et al.* develop a measurement–based admission rule for a network consisting of buffered multiplexers. 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 [25], [24]. First, we briefly review the measured sum approach; see [15], [16] for more details. 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. Jamin *et al.* base admission decisions on the estimate of the average 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. They suggest to set the averaging period to 0.5 seconds. 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 with measurement window W (in multiples of the averaging period T). Finally, let v denote a prespecified utilization target. Jamin *et al.* suggest to set $v = 0.9$. A new stream with leaky bucket rate r is accepted if $\hat{x} + rT \leq vCT$, 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. 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. The curves are obtained by varying the utilization target v. We observe 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 $W = T$. The plots in Figure 3 give also 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 $\epsilon = 10^{-6}$ and $\epsilon = 10^{-4}$. As in Section II we run the simulations for the LD admission rule until the width of the 90 % confidence interval of the loss probability is less than 20 % 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 % confidence interval of the loss probability is less than 10 % of the point estimate. The 90 % 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 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 $\epsilon = 10^{-6}$ our admission rule admits on average 178 connections and the loss probability is $1.1 \cdot 10^{-6}$. For the same average link utilization, i.e., for 178 connections, the admission rule of Jamin *et al.* gives a loss probability of roughly $6 \cdot 10^{-6}$. 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.

Time–Scale Decomposition Approach

Next we consider the time–scale decomposition approach developed by Grossglauser and Tse [13],

[14]. 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. 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. Furthermore, let σ_U^2 denote the estimate of the variance of the arrivals per connection in a slot. For details on how these estimates are obtained we refer the reader to [14]. A new connection with peak rate c_k^* is accepted if

$$Q\left(\frac{(C-c_k^*)T - Jm_U}{\sqrt{J\sigma_U^2}}\right) \leq \epsilon,$$

where $Q(\cdot)$ denotes the complementary cumulative distribution function of a standard normal random variable and J denotes the current number of ongoing connections.

The load–loss points ($J_{\text{avg}},\ P_{\text{loss}}$) (with 90 % 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 measured sum approach, which takes only the first moment into consideration. By taking the entire moment generating function into consideration the Large Deviation approach can accommodate even more connections (on average) while maintaining a given QoS requirement.

At this juncture we note an important study by Breslau *et al.* [3]. 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 [16] and the Large Deviation approach for a buffered multiplexer [18]. 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 % 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 [3] and were therefore not observed in that study.

IV. 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.

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. Suppose that the connection–j traffic at the smoother output is characterized by the traffic constraint function $\mathcal{E}_j(t) = \min(c_j^* t,\ \sigma_j + \rho_j t)$, that is, the output of smoother j is constrained by the smoother rate (peak rate) c_j^* and a single leaky bucket $(\sigma_j,\ \rho_j)$, 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 [8], [25].

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 $\mathcal{E}_{\text{lambs}}(t)$ by following the procedure described in [25]. With a maximum delay in the smoother, of 10 frame periods ($= 10/24$ seconds), we obtain the smoother rate $c^*_{\text{lambs}} = 731.6$ kbit/sec. For the 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 [8], [25]. The results are plotted as the solid line (labeled "adversarial") in Figure 4.

Histogram–based Admission Control

We next consider an admission rule that is specifically designed for prerecorded sources [23]. 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. For the numerical evaluation we assume that all multiplexed traffic streams are smoothed lambs video streams. The dashed line (labeled "histogram") in Figure 4 is the load–loss curve of this admission rule.

We also verify the accuracy of the histogram rule through simulation. For this purpose we use the simulation program used to evaluate the measurement–based admission rules in Sections II and III. 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 % confidence interval of the loss probability is less than 10 % 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 $\epsilon = 10^{-4}$ and $\epsilon = 10^{-6}$. The parameters of the admission rule are set to $\tau_p = 120$ and $\tau_q = 1,250$ for these simulations. The lifetime of each stream is fixed at $N = 40,000$ frame periods to ensure a fair comparison with the other admission rules.

Several points are noteworthy about Figure 4. First, 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 [25] 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 $\epsilon = 10^{-6}$, 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, 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.

V. Conclusion

In this paper we have studied measurement–based admission control for unbuffered multiplexers. We have developed a Large Deviations approach to measurement–based admission control that relies on aggregate measurements. We found in our simulations with MPEG–1 encoded videos that the LD admission rule compares favorably 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.

Acknowledgment: I thank Nick Duffield for pointing me to the work of Dublin's Applied Probability Group on measurement–based admission control. I am grateful to Matthias Grossglauser for clarifying the filtering mechanisms used in the time–scale decomposition approach [14].

References

[1] B.McGurk and C.Walsh. Investigations of the performance of a measurement–based connection admission control algorithm. In *Proceedings of 5th IFIP Workshop on Performance Modelling and Evaluation of ATM Networks*, Ilkley, UK, July 1997.

[2] D. D. Bovitch and N. G. Duffield. Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. *Queueing Systems*, 20:293–320, 1995.

[3] L. Breslau, S. Jamin, and S. Shenker. Measurement–based admission control: An empirical performance comparison. In *preprint*, February 1999.

[4] F. Brichet and A. Simonian. Conservative gaussian models applied to measurement–based admission control. In *Proceedings of IWQoS*, Napa, CA, May 1998. extended version available as Technical report of Project COST 257 at `ftp://www-info3.informatik.uni-wuerzburg.de/pub/cost/cost257/september97/257/td97056.ps`.

[5] C. Casetti, J. Kurose, and D. Towsley. A new algorithm for measurement–based admission control in integrated services packet networks. In *Proceedings of the Protocols for High Speed Networks Workshop*, October 1996. Available at `ftp://gaia.cs.umass.edu/pub/Case96:MB-Adm.ps.gz`.

[6] N. G. Duffield. Asymptotic sampling properties of effective bandwidth estimation for admission control. In *Proceedings of IEEE Infocom '99*, pages 1532–1538, New York, NY, March 1999.

[7] N.G. Duffield, J.T. Lewis, N. O'Connell, R. Russell, and F. Toomey. Entropy of ATM traffic streams: a tool for estimating quality of service parameters. *IEEE Journal on Selected Areas in Communications*, 13(6):981–990, 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] G. S. Fishman. *Principles of Discrete Event Simulation*. Wiley, 1991.

[10] S. Floyd. Comments on measurement–based admission control for controlled–load services. *submitted*, July 1996. available at `http://www-nrg.ee.lbl.gov:80/floyd`.

[11] R. J. Gibbens and F. P. Kelly. Measurement–based admission control. In *Proceedings of 15th International Teletraffic Congress (ITC15)*, pages 879–888, Washington, D.C., June 1997.

[12] R. J. Gibbens, F. P. Kelly, and P. B. Key. A decision theoretic approach to call admission control in ATM networks. *IEEE Journal on Selected Areas in Communications*, 13(6):1101–1114, August 1995.

[13] M. Grossglauser and D. Tse. A framework for robust measurement–based admission control. In *Proceedings of ACM SIGCOMM*, pages 237–248, Cannes, France, September 1997.

[14] M. Grossglauser and D. Tse. A time–scale decomposition approach to measurement–based admission control. In *Proceedings of IEEE Infocom '99*, pages 1539–1547, New York, NY, March 1999.

[15] S. Jamin, P. B. Danzig, S. J. Shenker, and L. Zhang. A measurement–based admission control algorithm for integrated services packet switched networks. In *Proceedings of ACM SIGCOMM '95*, pages 2–13, 1995.

[16] S. Jamin, P. B. Danzig, S. J. Shenker, and L. Zhang. A measurement–based admission control algorithm for integrated services packet switched networks (extended version). *IEEE/ACM Transactions on Networking*, 5(1):56–70, February 1997.

[17] S. Jamin and S. Shenker. Measurement–based admission control algorithms for controlled–load service: A structural examination. Technical Report CSE–TR–333–97, University of Michigan, April 1997.

[18] J.T.Lewis, R.Russell, F.Toomey, B.McGurk, S.Crosby, and I. Leslie. Practical connection admission control for ATM networks based on on–line measurements. *Computer Communications*, 21(17):1585–1596, November 1998.

[19] E. W. Knightly and J. Qiu. Measurement–based admission control with aggregate traffic envelopes. In *Proceedings of 10th IEEE International Tyrrhenian Workshop on Digital Communications*, Ischa, Italy, September 1998.

[20] 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.

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

[22] M. Reisslein. Measurement–based admission control: A large deviations approach for bufferless multiplexers (extended version). Technical report, GMD FOKUS, Berlin, Germany, June 1999. available at `http://www.fokus.gmd.de/usr/reisslein`.

[23] 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.

[24] M. Reisslein, K. W. Ross, and S. Rajagopal. Guaranteeing statistical QoS to regulated traffic: The multiple node case. In *Proceedings of 37th IEEE Conference on Decision and Control (CDC)*, pages 531–538, Tampa, FL, December 1998. available at `http://www.fokus.gmd.de/usr/reisslein` or `http://www.eurecom.fr/~ross/`.

[25] M. Reisslein, K. W. Ross, and S. Rajagopal. Guaranteeing statistical QoS to regulated traffic: The single node case. In *Proceedings of IEEE Infocom '99*, pages 1060–1071, New York, NY, March 1999.

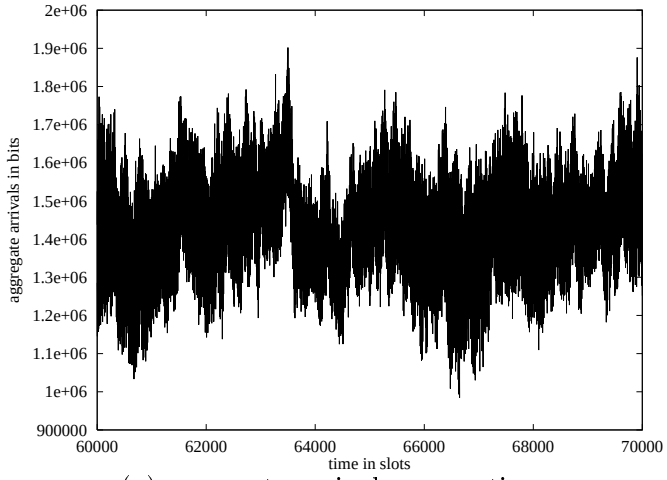
[26] J. Roberts, U. Mocci, and J. Virtamo (Eds.). *Broadband Network Traffic: Performance Evaluation and Design of Broadband Multiservice Networks, Final Report of Action COST 242, (Lecture Notes in Computer Science, Vol. 1155)*. Springer Verlag, 1996.

[27] J. W. Roberts. Realizing quality of service guarantees in multiservice networks. In T. Hasegawa, H. Takagi, and Y. Takahashi, editors, *Performance and Management of Complex Communication Networks — Proceedings of IFIP Conference PMCCN '97*, Tsukuba, Japan, November 1997. Chapmann Hall.

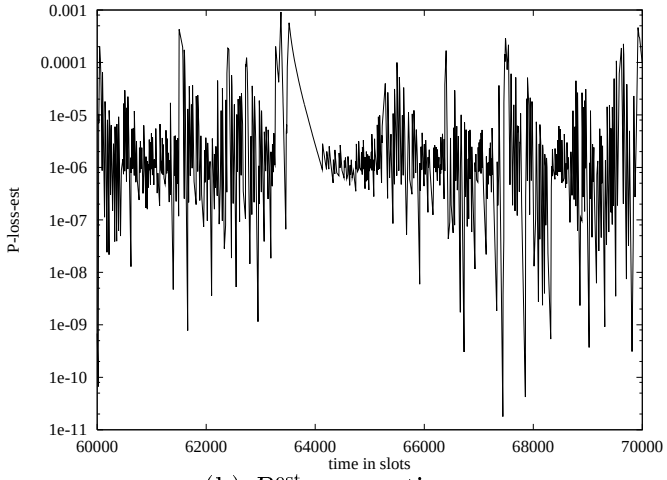
[28] O. Rose. Statistical properties of MPEG video traffic and their impact on traffic modelling in ATM systems. Technical Report 101, University of Wuerzburg, Insitute of Computer Science, Am Hubland, 97074 Wuerzburg, Germany, February 1995.
ftp address and directory of the used video traces:
ftp-info3.informatik.uni-wuerzburg.de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

[29] D. Tse and M. Grossglauser. Measurement–based admission control: Analysis and simulation. In *Proceedings of Infocom '97*, Kobe, Japan, April 1997.

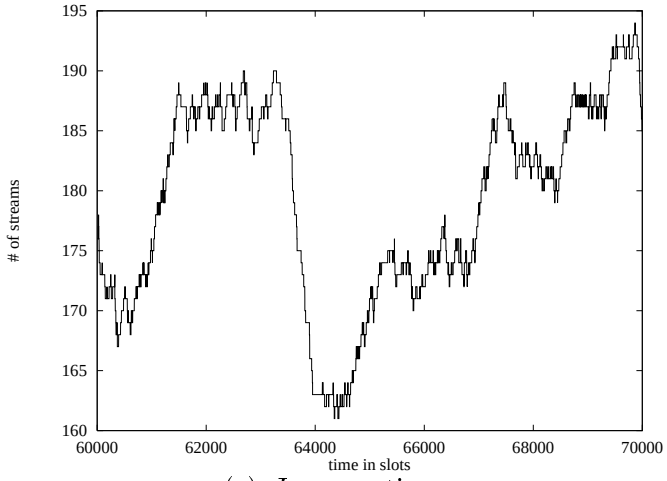
[30] C. Walsh and N.G. Duffield. Predicting QoS parameters for ATM traffic using shape–function estimation. In *Proceedings of 14th UK Teletraffic Symposium*, Manchester, UK, March 1997.



(a) aggregate arrivals versus time

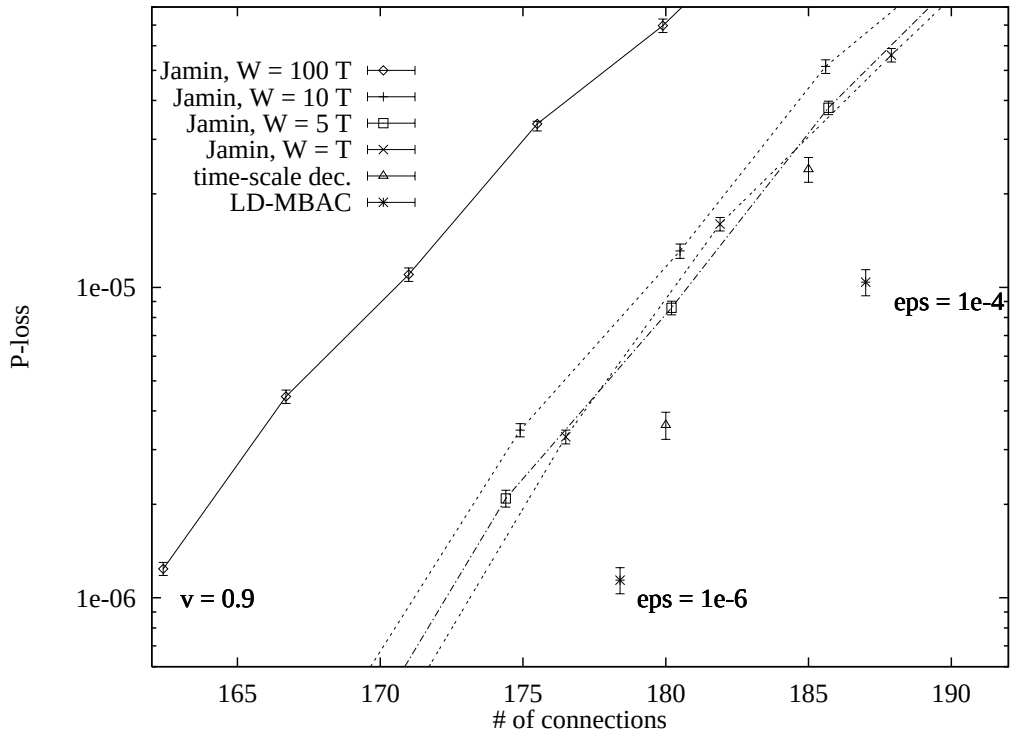


(b) $P_{\text{loss}}^{\text{est}}$ versus time

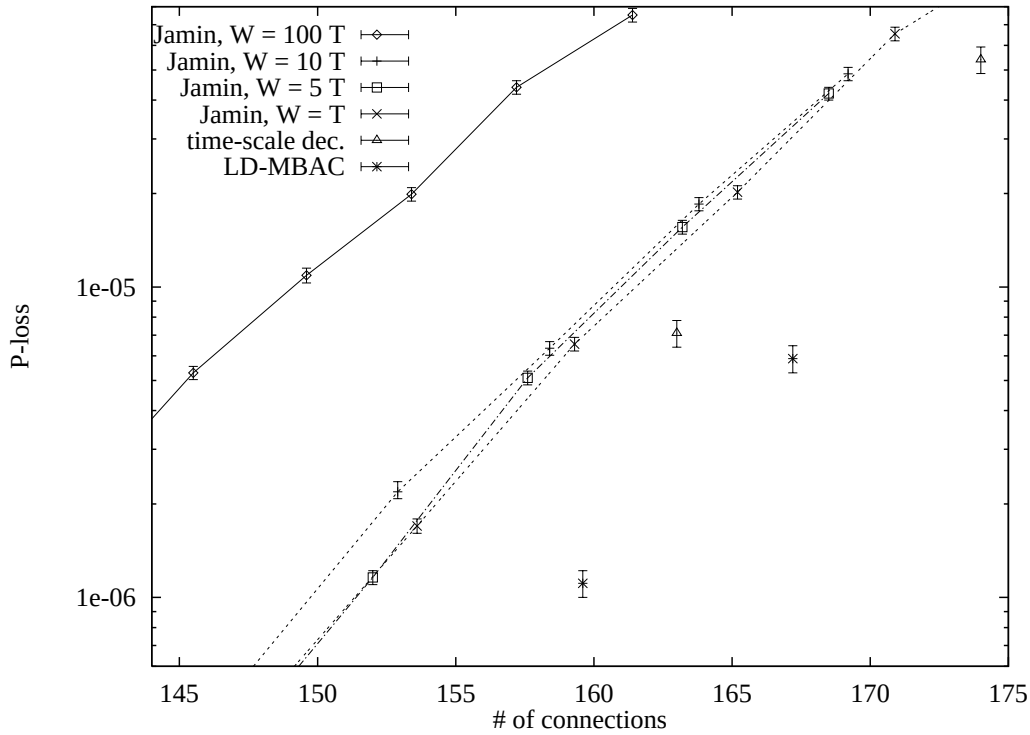


(c) J versus time

Fig. 2. Sample path plots from simulation of measurement–based admission rule with peak rate reservation for $\tau_p = 120$ and $\tau_q = 125$.



(a) maximum smoother delay = 10 frame periods



(b) no smoothing

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

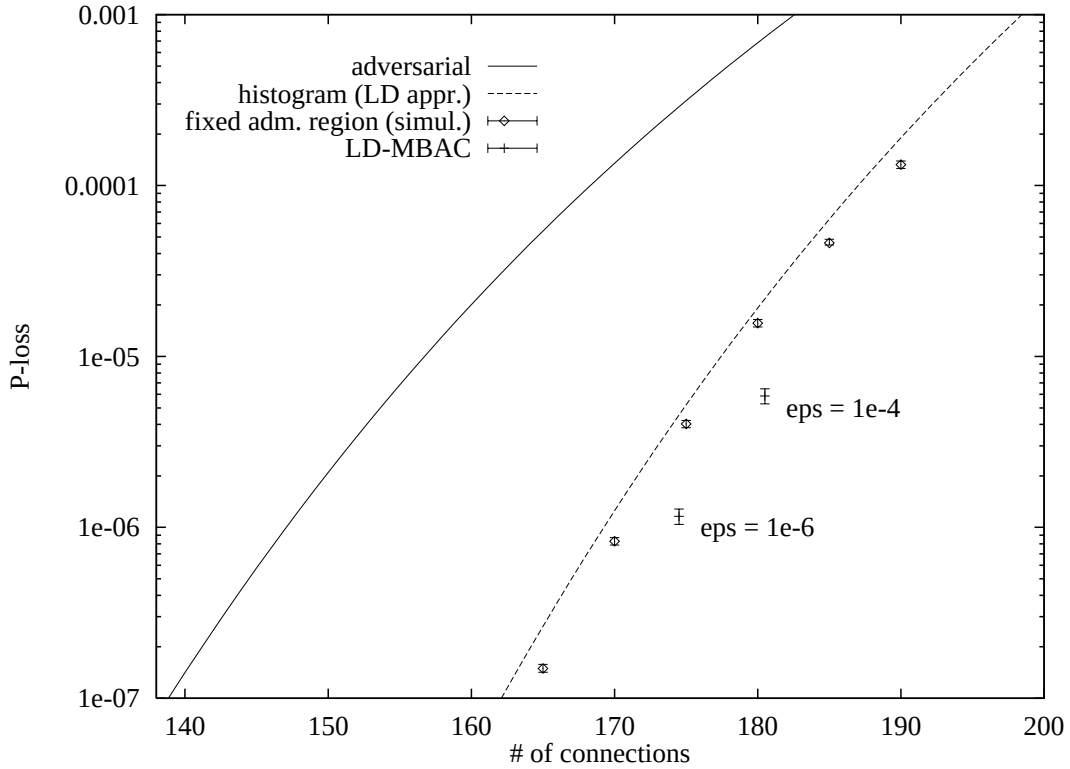


Fig. 4. Comparison of measurement–based admission control with traditional admission control.