Robust Measurement-Based Admission Control Using Markov's Theory of Canonical Distributions*

C. Pandit and S. Meyn†

January 13, 2006

Abstract

This paper presents models, algorithms and analysis for measurement-based admission control in network applications in which there is high uncertainty concerning source statistics. In the process it extends and unifies several recent approaches to admission control.

A new class of algorithms is introduced based on results concerning Markov's *canonical distributions*. In addition, a new model is developed for the evolution of the number of flows in the admission control system. Performance evaluation is done through both analysis and simulation. Results show that the proposed algorithms minimize buffer-overflow probability among the class of all *moment-consistent* algorithms.

Keywords: measurement-based admission control, robust estimation, worst-case source models, canonical distributions.

*This paper is based upon work supported by the National Science Foundation under Award Nos. ECS 02 17836 and ITR 00-85929. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation. Portions of the results presented here previously appeared in *Extremal distributions, entropy minimization and their application to admission control*, Proceedings of the International Symposium on Information Theory (ISIT), 2003.

†Coordinated Science Laboratory and the University of Illinois, 1308 W. Main Street, Urbana, IL 61801, URL `http://decision.csl.uiuc.edu:80/~meyn` (meyn@uiuc.edu).

1 Introduction

In computer networks and future wireless networks it is necessary to implement an admission control algorithm to determine which flow requests will be honored at a given server, typically a router in the Internet. Typical admission control algorithms assume some a priori knowledge such as the declared parameters of the flow (as in ATM), and/or a statistical model for the bandwidth-requests of flows. In measurement-based admission control (MBAC), statistical measurements are made on the aggregate packet arrival process in the recent past, and these measurements are used to determine which flow-requests will be honored.

Following the seminal paper [38], interest in MBAC grew through the mid-nineties, and has been sustained ever since. Although initial research concentrated on deterministic techniques based on token bucket characterizations of packet traffic [29], later authors have favored probabilistic approaches [8, 11, 15, 17, 16, 35, 36, 14, 5, 21, 22, 20, 19, 33].

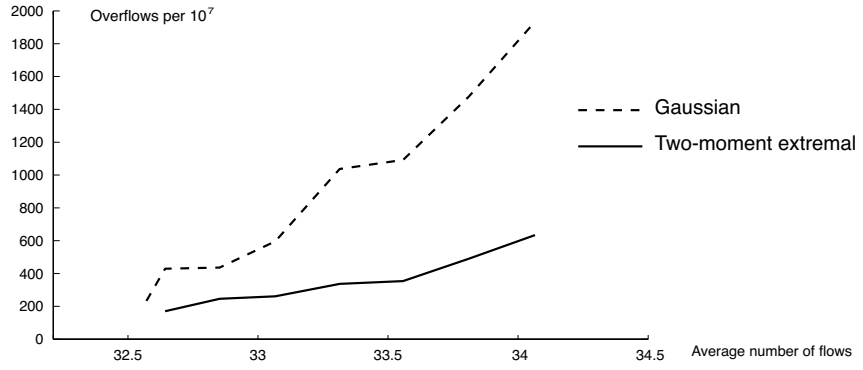

Figure 1: *Tradeoff between utilization and buffer overflow.* The two plots compare the performance of the algorithm based on the two-moment canonical distribution, and the algorithm of [22, 11] that is based on a Gaussian-source hypothesis. The true source distribution was i.i.d. uniform in this experiment.

A common performance metric used to evaluate an admission control algorithm is the *probability of overflow*. In order for the network to provide quality of service (QoS) guarantees, this probability is required to be below a pre-specified value, in the range 10^{-5} to 10^{-9}. Given these small values, it is argued that the theory of Large Deviations is justified in the analysis of performance of admission control algorithms [11, 16, 8, 5, 33, 36].

The *measurments* used in MBAC algorithms depend upon the particular algorithm and assumptions imposed. These range from packet delay [29], to virtual buffer overflows [8], to moment statistics [11, 16, 22, 20, 15, 5], to empirical distributions [36, 11].

In particular, the algorithms considered in [16, 5] are based on first-moment measurements only. In [11, 22, 20, 5] it is assumed that first and second moments are estimated, and based on this the MBAC algorithm is constructed based on a Gaussian approximation to estimate overflow probabilities.

This looks like first moments to me: [15, 5] use variations of the Hoeffding inequality [26] as a basis for admission control.

This paper adopts a *worst-case* approach to source modelling, similar in spirit to that of [31, 23]. The idea is to make moment measurements on the packet traffic, and subject to these measurements, choose a source model that is worst-case in the sense that it maximizes the estimate of the overflow probability. This estimate is based upon a large deviations approximation, as in [11]. These approximations lead naturally to a robust algorithm for MBAC.

The following is a summary of the main contributions of this paper:

(i) A novel class of algorithms for MBAC is introduced, based on a particular class of source models whose marginal distribution is *canonical*. Canonical distributions, first introduced by A. A. Markov [32], arise as unique solutions to a collection of (infinite-dimensional) linear programs involving moment constraints [33, 34, 32].

include Gabor in survey?

The source model with canonical marginal is worst-case in the sense that the associated large deviation asymptotics yield the highest probability of overflow in a standard queueing model. This leads naturally to associated MBAC algorithms based on moment estimates.

(ii) A flow model that takes into account the effect of a MBAC algorithm has been lacking in the literature. Indeed, most authors have preferred to study the effect of MBAC algorithms in isolation, neglecting their effect on the arrival and departure of flows. This gap is filled in the present paper through the introduction of a Markov model for the evolution of flows in the admission control system, based on the physically natural assumption of *time-scale separation* (see e.g. [17, 16, 39]). This model is considered in our analysis of both bufferless as well as buffered models.

(iii) The performance of the MBAC algorithm is investigated through both analysis and simulations. The analysis shows that the MBAC algorithm presented in this paper is optimal among a wide class of algorithms, in the sense that it minimizes the associated buffer-overflow probability.

(iv) Buffer-overflow probability has been emphasized as a performance metric, but of course one is also interested in maximizing utilization. However, by regulating traffic flow carefully one may achieve both high utilization and low buffer-overflow. Figure 1 illustrates the performance of the algorithms introduced here with respect to both of these metrics. The vertical axis shows the number of buffer-overflows, and the horizontal axis the average number of flows in the system. The two trade-off curves shown correspond to an instance of one of the algorithms introduced here, and the "Gaussian algorithm" of [22, 11] for comparison. It is seen in Figure 1 that the new algorithm outperforms the Gaussian algorithm with respect to each performance metric in this experiment. That is, for a fixed value of the buffer-overflow probability, the new algorithm has higher utilization, and for fixed utilization, the algorithm has a lower buffer-overflow probability. Experiments described in Section 5 using different models give consistent results.

The remainder of the paper is organized as follows: Section 2 begins with a description of the bufferless admission control model in Section 2.1. The MBAC algorithms introduced in this paper are described in Section 2.3 for the bufferless model. This section also contains a key result concerning canonical distributions. A performance analysis of these algorithms is contained in Section 3. The models, algorithms, and analysis are extended to the buffered model in Section 4.

Simulation studies are presented in Section 5, and conclusions and open problems are discussed in Section 6.

In the following two section we restrict to a bufferless server. This simplifies the model and its analysis to a large extent, and provides valuable insight into the behavior of the buffered model as well. Extensions to the buffered case are described in detail in Section 4.

2 Measurement Based Admission Control

In this section we present algorithms for MBAC that have been proposed in literature, and discuss their properties and drawbacks. We then recall Markov's theory of canonical distributions [32], and propose a new algorithm based on a key result from this literature.

We begin with a description of the server model.

2.1 Admission control model

The admission control models considered in this paper are designed to reflect the behavior of a high-capacity Internet router, which is accessed by a correspondingly large number of flows. This is captured by an integer scaling parameter $N \gg 1$. We assume that the server has capacity $\bar{C} = NC$, and the conditions imposed below imply that the mean number of flows in the system also scales linearly with N.

reviewer wants more explanation on scaling, explanation of tradeoff curves.

Each flow accessing the server sends a stream of packets as illustrated in Figure 2. If the server has a buffer, the packets are queued for service, and are dropped if the queue length $Q(t)$ exceeds the buffer size. If the server is bufferless, packets are not queued and are dropped if the total packet arrivals in the preceding time-slot exceeds the server capacity. When a new flow request arrives at the server, an admission control decision is made to either accept or reject the new flow. This decision is based on measurements of past observations at the server.

Figure 2: The admission control system for bufferless and buffered servers. The server has capacity $\bar{C} = NC$, and the buffer-capacity is $\bar{B} = NB$ in the buffered model, where $N \gg 1$ is a scaling-parameter.

We impose some simplifying assumptions to simplify discussion in the remainder of this section: For each $N \geq 1$, the number of flows is assumed *fixed*, of the the form Nm for some constant $m > 0$, and we restrict to the bufferless server. A dynamic flow model is introduced in Section 3.1, and buffered models are introduced in Section 4.

The packet arrival process from the ith flow at time t is denoted $X_i(t)$. It is assumed that the random variables $\{X_i(t) : i \geq 1,\ t \geq 0\}$ are mutually independent and identically distributed. The common marginal distribution Γ° is supported on $\{0, 1, \ldots, R\}$ for a known constant $R > 0$. The independence assumption across time is relaxed in the buffered model.

The total number of packets generated by the mN flows in the tth time-slot is denoted,

$$A(t) := \sum_{i=1}^{Nm} X_i(t)\,, \qquad t \geq 0. \tag{1}$$

The probability of overflow at time t is given by,

$$\mathsf{P}\{A(t) > NC\} = \mathsf{P}\Big\{\frac{1}{Nm}\sum_{i=1}^{Nm} X_i(t) > C/m\Big\} \tag{2}$$

For large N the overflow probability is close to 1 when $m > C/\xi_1$ with ξ_1 equal to the common mean of $X_i(t)$, by the weak law of large numbers. The overflow probability may be bounded using Cramér's Theorem [10, Thm 2.2.3] for $m \in (C/\xi_1, C/R)$. Recall that the rate-function I_Γ for a given distribution Γ on $\mathbb{R}$ is expressed as the convex dual,

$$I_\Gamma(r) := \sup_{\theta\in\mathbb{R}} \{\theta r - \Lambda_\Gamma(\theta)\}, \qquad r \in \mathbb{R}\,, \tag{3}$$

where the log moment-generating function is given by

$$\Lambda_\Gamma(\theta) := \log \sum e^{\theta x}\, \Gamma(x), \qquad \theta \in \mathbb{R}. \tag{4}$$

Since $\{X_i(t)\}$ are assumed i.i.d. with common marginal distribution Γ°, the expression (2) combined with Cramér's Theorem gives,

$$\begin{aligned} \log\big(\mathsf{P}\{\text{overflow at time } t\}\big) &\le -Nm\bar{I}_{\Gamma^\circ}(C/m)\,, \qquad N \ge 1\,; \\ \lim_{N\to\infty} \frac{1}{N} \log\big(\mathsf{P}\{\text{overflow at time } t\}\big) &= -m\bar{I}_{\Gamma^\circ}(C/m)\,, \end{aligned} \tag{5}$$

where for any distribution Γ,

$$\bar{I}_\Gamma(r) := \begin{cases} 0, & r < \xi_1 \\ I_\Gamma(r), & \xi_1 \le r < R \\ \infty, & r \ge R \end{cases} \tag{6}$$

The bound (5) uses $\bar{I}_\Gamma$ instead of the standard rate function because of the strict inequality in the definition of the probability of overflow in (2).

2.2 Algorithms based on the empirical distribution

The objective of admission control is to regulate the overflow probability to be less than some number $\eta > 0$, which is typically in the range $10^{-5} - 10^{-9}$. The limit (5) suggests that this target value should scale with the parameter N: For some constant $I_\eta > 0$,

$$\eta(N) = e^{-I_\eta N}, \qquad N \ge 1. \tag{7}$$

The following stationary policy for the bufferless model is suggested by (5):

> *Given that there are Nm flows accessing the server in time-slot t, a new flow request at time t will be accepted if and only if $\bar{I}_{\Gamma^\circ}(C/m) \ge m^{-1} I_\eta$.* (8)

What information is required to enforce the decision rule (8)? In addition to the target probability η and the number of flows m, a critical piece of information that is typically not known in advance is the rate-function $I_{\Gamma^\circ}(\cdot)$. Consequently, implementation of this decision rule requires some form of *rate-function estimation* from statistical measurements made on the packet process.

Since the packet arrivals $X_i(t), i \ge 1$ are assumed to be i.i.d. across i, the statistics of $X_i(t)$ may be estimated by considering only the arrivals in the current time slot t. In several

recent papers, rate-function estimates for MBAC are formulated using the associated *empirical distribution*, given as follows:

$$\widehat{\Gamma}_t(k) = \frac{1}{Nm} \sum_{i=1}^{Nm} \mathbb{I}\{X_i(t) = k\}\,, \qquad k \in \mathbb{Z}_+. \tag{9}$$

We let $c_i(x) = x^i$ for $i \geq 1$, and define the *empirical ith moment* by,

$$\widehat{\xi}_i(t) = \widehat{\Gamma}_t(c_i) = \frac{1}{Nm} \sum_{j=1}^{Nm} (X_j(t))^i. \tag{10}$$

The following three approaches to MBAC have received significant attention in the recent literature:

(i) Assume that the peak rate R is known. Based on the empirical mean $\widehat{\xi}_1 = \widehat{\xi}_1(t)$ at time t, consider the distribution Γ supported on 0 and R, with $\Gamma(R) = 1 - \Gamma(0) = \widehat{\xi}_1/R$. The *Bernoulli algorithm* considered in [16] is the decision rule (8) with Γ° replaced by Γ.

(ii) Measure the empirical mean $\widehat{\xi}_1$, and also the empirical variance,

$$\widehat{\sigma}^2 := \widehat{\xi}_2 - \widehat{\xi}_1^2 = \frac{1}{Nm-1} \sum_{i=1}^{Nm} (X_i(t) - \widehat{\xi}_1)^2.$$

The distribution Γ is defined to be Gaussian $N(\widehat{\xi}_1, \widehat{\sigma}^2)$, whose associated rate-function is given by $I_\Gamma(r) = \bar{I}_\Gamma(r) = (r - \widehat{\xi}_1)^2/(2\widehat{\sigma}^2)$, $r \geq \widehat{\xi}_1$. The resulting decision rule using Γ instead of Γ° in (8) was considered previously in [22, 20]. We will refer to this as the *Gaussian algorithm*.

(iii) The *certainty-equivalent* approach is to apply (8) using the empirical distribution $\widehat{\Gamma}_t$. This approach is considered in several papers, e.g., [36, 11].

Each of these approaches presents drawbacks:

(a) The Bernoulli algorithm can be overly conservative, and hence reject too many flows, since it uses so little information. This is especially true when the peak rate R is large.

simulations show otherwise!

(b) The Gaussian algorithm is based on a Central Limit Theorem approximation and is consequently appropriate for moderately large values of the target η (see discussion in [22] and simulations in Section 5.)

(c) Implementation of the third approach requires computation of the empirical log moment-generating function $\widehat{\Lambda}(\theta) := \log(\widehat{\Gamma}_{N,t}(e^{\theta c_1}))$ for each $\theta > 0$, along with its convex dual. In addition to this computational burden, the empirical distributions will inevitably bring large variances and hence significant estimation error.

In the next section we introduce a new class of models that allows a compromise between the complex certainty-equivalent approach, and the simple Bernoulli algorithm.

2.3 Moment-consistent algorithms

The algorithms presented here are generalizations of the Bernoulli algorithm described in Section 2.2. We fix an integer $M \geq 1$ corresponding to the number of moments to be measured, and define the vector function $c\colon \mathbb{R} \to \mathbb{R}^M$ via $c(x) = (x, x^2, \ldots, x^M)^{\tau}$. A prescribed bound on the overflow probability of the form (7) is given, and the following data is assumed to be available at each time t: (i) The current number of flows $Nm(t)$, and (ii) The empirical mean $\widehat{\xi}(t) \in \mathbb{R}^m$ defined in (10).

Let $\mathcal{M}$ denote the space of probability distributions on $[0, R]$, and $\Delta \subset \mathbb{R}^M$ the set of feasible moment vectors,

$$\Delta := \{\Gamma(c) : \Gamma \in \mathcal{M}\} \subset \mathbb{R}^M. \tag{11}$$

For $\xi \in \Delta$ we let $\mathcal{M}_\xi$ denote the set of consistent marginal distributions,

$$\mathcal{M}_\xi = \{\Gamma \in \mathcal{M} : \Gamma(c) = \xi\}. \tag{12}$$

A *moment-consistent map* is any measurable function $\Gamma_{(\,\cdot\,)}$ from Δ to $\mathcal{M}$ that satisfies $\Gamma_\xi \in \mathcal{M}_\xi$ for each $\xi \in \Delta$. That is,

$$\Gamma_\xi(c) = \xi, \qquad \xi \in \Delta.$$

Given a moment-consistent map $\Gamma_{(\,\cdot\,)}$ and target propbability $\eta(N) = e^{-I_\eta N}$, we define an associated MBAC algorithm as follows.

Moment-consistent algorithm $\mathcal{A}$: Measure the vector of empirical moments $\widehat{\xi} = \widehat{\xi}(t)$, and identify the corresponding distribution $\Gamma_{\widehat{\xi}} \in \mathcal{M}_\xi$. If a flow arrives at time t, it is accepted if, and only if

$$\bar{I}_{\Gamma_{\widehat{\xi}}}(C/m) \geq m^{-1} I_\eta\,,$$

where mN is equal to the number of flows accessing the server. Equivalently, a flow is admitted at time t if and only if $\widehat{\xi}(t)$ belongs to the *acceptance region*,

$$\Delta_m := \left\{\xi \in \mathbb{R}^M : \bar{I}_{\Gamma_\xi}(C/m) \geq m^{-1} I_\eta\right\}. \tag{13}$$

■

There are of course many ways to define a moment consistent mapping. Theorem 2.1 below implies that there exists a moment-consistent mapping that is *optimal* in the sense that it minimizes the associated large deviations rate-function.

Subject to a finite number of moment constraints ξ, a distribution $\Gamma^* \in \mathcal{M}_\xi$ is called *canonical* if it minimizes the associated large deviations rate-function point-wise, in the sense that $I_{\Gamma^*}(r) \leq I_\Gamma(r)$ for all $r \geq \xi_1$, and all distributions $\Gamma \in \mathcal{M}_\xi$. A remarkable theory due to Markov establishes the existence of a canonical distribution satisfying this global lower bound [32]. The distribution Γ^* enjoys many attractive properties that lend themselves to the construction of simple, effective algorithms.

Theorem 2.1 and several generalizations are proved in [34], following Markov's original result [32].

Theorem 2.1 *For each $\xi \in \Delta$, there exists a unique probability distribution $\Gamma^*_\xi \in \mathcal{M}_\xi$ called the* canonical distribution *with the following properties:*

(i) $I_{\Gamma^*_\xi}(r) \le I_\Gamma(r)$ *for any other distribution* $\Gamma \in \mathcal{M}_\xi$*, and for any* $r \in \mathbb{R}$*.*

(ii) *If the moment vector* ξ *lies in the interior of* Δ*, then the canonical distribution is discrete, with at most* $\lceil n/2 \rceil + 1$ *points of support, and the end-point* R *always lies in its support. If* M *is odd, then both end-points,* 0 *and* R*, lie in the support of* Γ^*_ξ*.* note!

(iii) *When* $M = 1$ *the canonical distribution is the unique binary distribution* $\Gamma^* \in \mathcal{M}_\xi$ *with support on the two points* $\{0, R\}$*. When* $M = 2$ *the canonical distribution is again binary,*

$$\Gamma^* = p^* \, \delta_{x^*} + (1 - p^*) \, \delta_R \,, \tag{14}$$

where $x^* = [R - \xi_1]^{-1}(\xi_1 R - \xi_2)$ *and* $p^* = [R^2 + \xi_2 - 2\xi_1 R]^{-1}(R - \xi_1)^2$*.* □

Our main interest in Theorem 2.1 is its evident application to the construction of large deviations bounds. The corollary below follows directly from Cramér's Theorem applied to an i.i.d. process with canonical marginal distribution.

Proposition 2.2 *Consider an i.i.d. process* $\{X_i\}$ *with marginal distribution* $\Gamma^\circ \in \mathcal{M}_\xi$*. Then we have the uniform bound,*

$$\mathsf{P}\Big\{\sum_{i=1}^{N} X_i(t) > NC\Big\} \le \exp(-N I_{\Gamma^*_\xi}(C)) \,, \qquad \xi_1 \le C \le R, \ N \ge 1 \,,$$

where $\Gamma^* \in \mathcal{M}_\xi$ *denotes the canonical distribution, and* $I_{\Gamma^*_\xi}$ *the associated rate function.* □

Proposition 2.2 extends and unifies well-known approaches to the construction of exponential bounds on error probabilities. On specializing to $M = 1$ this is a version of Hoeffding's inequality [26], and the special case $M = 2$ is identical to Bennett's Lemma [1]. See [12, 13] for recent generalizations of Hoeffding's inequality to Markov models.

These results provide ample motivation for the following *extremal moment-consistent algorithm*:

Algorithm $\mathcal{A}^*$: If a flow arrives at time t, it is accepted if, and only if $\widehat{\xi} \in \Delta^*_m$, where $\widehat{\xi} = \widehat{\xi}(t)$ is the vector of empirical moments, and

$$\Delta^*_m := \Big\{\xi \in \mathbb{R}^M : m\bar{I}_{\Gamma^*_\xi}(C/m) \ge I_\eta\Big\} \qquad \xi \in \Delta \,. \tag{15}$$

■

The algorithm $\mathcal{A}^*$ reduces to the Bernoulli algorithm of [16] when $M = 1$. For large M it is approximated by the certainty equivalence approach described in Section 2.2, in the sense that the corresponding acceptance regions Δ^*_m converge (see Theorem 3.1).

3 Performance Analysis

We now several results that illustrate the dynamics and performance of the MBAC algorithm $\mathcal{A}^*$. The following conclusions are obtained under mild conditions. Recall that Γ° is the true marginal distribution of $X_i(t)$, and $\xi^\circ = \Gamma^\circ(c)$.

(i) Theorem 3.1 shows that the algorithm $\mathcal{A}^*$ is optimal within the class of moment-consistent algorithms, in the sense that the overflow probability is minimized for each finite value of N.

(ii) It follows from Theorem 3.2 that the algorithm $\mathcal{A}^*$ is approximated by a threshold policy when N is large. The threshold is of the form m^*N, where

$$m^* := \sup\{m : \xi \in \Delta_m^*\} = \sup\left\{m : m\bar{I}_{\Gamma^*_{\xi^\circ}}(C/m) \geq I_\eta\right\}. \tag{16}$$

(iii) As one corollary, Proposition 3.3 establishes the existence of a constant $m^\bullet \leq m^*$ such that the number of flows in the system in steady state is very likely to be near $\lfloor Nm^\bullet \rfloor$ for large N.

(iv) Theorem 3.4 establishes an LDP for the steady-state probability of overflow.

The following technical assumptions are imposed whenever LDP bounds are invoked for the M-dimensional process $\{c(X_j(t)) : j \geq 1\}$

(A1) The vector mean $\xi^\circ := \Gamma^\circ(c)$ lies in the interior of Δ (defined in (11).)

(A2) $C < Rm^*$, so that $\bar{I}_{\Gamma^\circ}(C/m^*) = I_{\Gamma^\circ}(C/m^*) < \infty$.

Assumption A2 is made primarily for ease of analysis: it implies that the probability of overflow with $\lfloor Nm^* \rfloor$ flows in the system decays exponentially (and not super-exponentially). Neither assumption is essential: The results presented below will hold, with some minor modifications, even if A1 and A2 are relaxed.

We now present a model of flow dynamics.

3.1 Flow model

While there is a vast literature on queueing models, the development of stochastic flow models is still a topic of active research (see e.g. [30, 6, 37]). In fact, most references on MBAC the number of flows in the system is assumed to be fixed. In particular, the analysis of [11] is restricted to a model with an infinite backlog of flows.

We introduce here a *Markov Decision Process* model to provide a framework for performance (or QoS) evaluation. Recall that $N \gg 1$ denotes a scaling parameter, and $\bar{C} = NC$ denotes the server capacity. We let NW denote a maximum value for the number of flows in the system: the admission control algorithm will reject any new flow request if the current number of flows is NW. This model is meant to approximate a continuous-time model in which flow requests arrive according to a Poisson process with rate λ, and the holding time of each flow has an exponential distribution with parameter denoted $N^{-1}\mu$. Through a time-scaling we may assume without loss of generality that $\lambda + W\mu = 1$.

The state process $\boldsymbol{\Phi}$ represents the number of flows accessing the router, with state space equal to $\mathbb{Z}_+$. The control process $\boldsymbol{U}$ taking values in $\{0, 1\}$ represents the decision process of acceptance or rejection of new flow requests. If $U(t) = 1$ then a new flow request will be honored at time t. The controlled transition probabilities $P_a(i, j) = \mathsf{P}\{\Phi(n + 1) = j \mid \Phi(n) = i,\ U(n) = a\}$, $i, j \in \mathbb{Z}_+$, $a \in \{0, 1\}$ are defined for $i \in \{0 \dots NW - 1\}$ by,

$$P_a(i + 1, i) = N^{-1}\mu i, \qquad P_a(i, i + 1) = \lambda a, \tag{17}$$

and $P_a(i,i) = 1 - P_a(i,i-1) - P_a(i,i+1)$.

there is no reason to mention cts time!

Control policies can depend on measurements of the flow process $\mathbf{\Phi}$ and the packet arrival processes $\{X_j(t)\}$. It is assumed that for each $n \geq 1$ the future of the packet process $\{X_j(n+\ell) : j, \ell \geq 1\}$ is independent of past and previous observations $\{\Phi(i), U(i), X_j(i) : i \leq n,\ j \geq 1\}$. The statistics of the aggregate packet-arrivals at the server are defined for any time $n \geq 0$ and any $k \in \mathbb{Z}_+$, through the conditional distributions,

$$\mathsf{P}\Big\{A(n+1) = k \mid A(i), \Phi(i), U(i), X_j(i),\ i \leq n,\ j \geq 1\Big\} = \sum_{y=0}^{\infty} P_a(x,y)\mathsf{P}\Big\{\sum_{j=1}^{y} X_j(n+1) = k\Big\}, \tag{18}$$

where $U(n) = a \in \{0,1\},\ \Phi(n) = x \in \mathbb{Z}_+$.

3.2 Optimality of the extremal algorithm

Performance bounds are obtained here for the extremal algorithm based on the flow model described in Section 3.1, maintaining the assumptions on the packet processes described earlier.

We consider for comparison an arbitrary moment-consistent algorithm $\mathcal{A}$, with associated moment-consistent map $\Gamma_{(\,\cdot\,)}$ and acceptance region Δ_m as defined in Section 2.3. The controlled process $\mathbf{\Phi}$ is a Markov chain with transition probabilities,

$$P(i,i+1) = \lambda \mathsf{P}\{\widehat{\xi} \in \Delta_{i/N}\}, \qquad P(i+1,i) = i\mu/N, \qquad i \in \{0 \dots NW-1\}\,. \tag{19}$$

Although $\widehat{\xi} = \widehat{\xi}(t)$ depends on t, the transition probabilities are independent of time in the bufferless model.

We let π denote the unique steady-state distribution, and we henceforth restrict attention to the steady-state overflow probability given by

$$\eta = \eta(N) = \sum_{i=1}^{\infty} \pi(i)\mathsf{P}[\,\text{overflow} \mid j]\,.$$

The following result establishes minimality of the overflow probability for the algorithm $\mathcal{A}^*$. The proof of Theorem 3.1 and all of the results that follow are collected together in the appendix.

Theorem 3.1 *Let $\eta^* = \eta^*_M$ denote the steady-state overflow probability under $\mathcal{A}^*$ with $M \geq 1$ moment constraints; η_M the corresponding quantity for an arbitrary moment-consistent algorithm $\mathcal{A}$; and η^*_∞ the overflow probability for the certainty-equivalent algorithm described in Section 2.2. Then, $\eta^*_M \leq \min(\eta^*_\infty, \eta_M)$ for each finite N and M. Moreover, for each fixed $N \geq 1$,*

$$\lim_{M\to\infty} \eta^*_M = \eta^*_\infty.$$

□

The next result concerns the asymptotic behavior of the algorithm for large N. The M-dimensional rate function for the i.i.d. sequence $\{c(X_j(t)) : j \geq 1\}$ is denoted,

$$I_{M,\Gamma^\circ}(v) := \sup_{\theta\in\mathbb{R}^M} \{\theta^T v - \Lambda(\theta)\}\,, \quad v \in \mathbb{R}^M. \tag{20}$$

where $\Lambda_{M,\Gamma^\circ}(\theta) := \log \Gamma^\circ(e^{\theta^T c})$, $\theta \in \mathbb{R}^M$. We define $K\colon \mathbb{R}_+ \to \mathbb{R}_+$ by,

$$K(m) = \inf\{mI_{M,\Gamma^\circ}(\xi) : \xi \in \mathrm{cl}(\Delta_m^*)\},$$

where $\mathrm{cl}(D)$ denotes the closure of a set $D \subset \mathbb{R}^M$.

For large N, the algorithm $\mathcal{A}^*$ is approximated by a threshold policy:

Theorem 3.2 *For the algorithm $\mathcal{A}^*$, $\mathsf{P}[\widehat{\xi}(t) \in \Delta_m^*] \to 1$ as $N \to \infty$ when $m < m^*$, $\limsup_{N\to\infty} \mathsf{P}[\widehat{\xi}(t) \in \Delta_m^*] < 1$ when $m = m^*$, and*

$$\limsup_{N\to\infty} N^{-1} \log\, \mathsf{P}[\widehat{\xi}(t) \in \Delta_m^*] \le -K(m) < 0, \qquad m > m^*.$$

□

Note that when combined with (19), the above proposition gives a corresponding LDP for the transition probabilities $P(Nm, Nm+1)$.

The following corollary shows that the invariant distribution π is essentially supported on the interval $[0, \lfloor Nm^* \rfloor]$: The probability that there are more than $\lfloor Nm^* \rfloor$ flows in the system decays to zero super-exponentially as $N \to \infty$, and thus the Markov chain $\mathbf{\Phi}$ behaves like an $M/M/d/d$ queue with $d = Nm^*$, arrival rate λ, and departure rate μ/N.

Let $\rho = \lambda/\mu$, set $m^\bullet := \min(m^*, \rho)$, and define

$$L(m) = \begin{cases} (m^\bullet - m)(1 + \log(\rho)) + m\log(m) - m^\bullet \log(m^\bullet); & 0 \le m \le m^*, \\ \infty; & m > m^*. \end{cases} \tag{21}$$

Proposition 3.3 *The invariant distribution π^* of the Markov chain under algorithm $\mathcal{A}^*$ satisfies the following:*

(i) *The probability mass is concentrated around $m^\bullet N$ for large N, i.e.,*

$$\lim_{N\to\infty} \sum_{j=\lfloor N(m^\bullet-\epsilon)\rfloor}^{\lfloor N(m^\bullet+\epsilon)\rfloor} \pi^*(j) = 1 \qquad \text{for every } \epsilon > 0. \tag{22}$$

(ii) *For each $m \ge 0$, $\lim_{N\to\infty} N^{-1}\log(\pi^*(\lfloor Nm \rfloor)) = -L(m)$.* □

An LDP for the overflow probability η is given next. We denote the conditional probability of overflow given j flows in the system by,

$$q_j = q_j(N) := \mathsf{P}\{\sum_{i=1}^{j} X_i(t) > NC\}. \tag{23}$$

Cramér's Theorem gives the limit,

$$\lim_{N\to\infty} N^{-1}\log(q_{\lfloor Nm \rfloor}) = -m\bar{I}_{\Gamma^\circ}(C/m)\,, \qquad m\xi_1 \le C \le mR.$$

Combining this with Proposition 3.3, then gives,

$$\lim_{N\to\infty} N^{-1}\log(\pi^*_{\lfloor Nm \rfloor} q_{\lfloor Nm \rfloor}) = -(L(m) + m\bar{I}_{\Gamma^\circ}(C/m)).$$

Since the steady state overflow probability η^* is a sum of terms of the form $\pi^*_{\lfloor Nm \rfloor} q_{\lfloor Nm \rfloor}$, the following result is an affirmation of the usual large deviations principle that the term with the slowest rate of decay dominates the rate of decay of the sum.

Theorem 3.4 *The steady state probability of overflow probability η^* for the algorithm $\mathcal{A}^*$ satisfies the large deviations principle,*

$$\lim_{N\to\infty} N^{-1}\log(\eta^*) = -\inf_{m>0}\{L(m) + m\bar{I}_{\Gamma^\circ}(C/m)\}. \tag{24}$$

Moreover, the infimum above is achieved at $m = m^*$ if $m^* \le \rho$. □

4 Consideration Of Buffers

The bufferless model is convenient in analysis, but unfortunately does not reflect the behavior of many physical systems. In this section we extend our models, algorithms and analysis to the buffered server illustrated at right in Figure 2.

4.1 The buffered model

We maintain a discrete-time model for packet arrivals, but we now consider more general statistics. For each $0 < s < t < \infty$ we let $X_i(s,t)$ denote the packet arrivals due to flow i in the time-interval $(s,t]$. It is assumed that, for any given $s \in \mathbb{Z}_+$, $\{X_i(s,t) : t \ge s\}$ has stationary and ergodic increments with peak packet arrival rate R (i.e. $X_i(s,t) \le R(t-s)$ for all s,t.) We again denote by $X_i(t) = X_i(t-1,t)$ the packet arrivals in a single time slot t, and the total number of packets generated by the flows at time t is again defined in (1) and denoted $A(t)$.

The server contains a queue as shown at right in Figure 2. The queue process $\boldsymbol{Q}$ satisfies the usual Lindley recursion,

$$Q(t+1) = [Q(t) + A(t+1) - \bar{C}]_+, \qquad t \ge 0,$$

where $\bar{C} = NC$ is again the server capacity. The buffer-capacity is $\bar{B} = NB$, where again $N \gg 1$ is a scaling-parameter. An overflow occurs at time t if $Q(t) > NB$.

For a model with buffers, an approximation for the buffer-overflow probability generalizing (5) is given by the *many-sources asymptotic* of [9, 4]: Let $I_{\Gamma^\circ,t}$ denote the large deviations rate-function of $X_i(0,t)$ (and therefore of $X_i(s,s+t)$, since $X_i(\cdot)$ has stationary increments), and define $\bar{I}_{\Gamma^\circ,t}$ based on $I_{\Gamma^\circ,t}$ as in (6). We then have,

$$\lim_{N\to\infty} N^{-1}\log(\mathsf{P}[\text{overflow}]) = -m\inf_{t\ge 0}\bar{I}_{\Gamma^\circ,t}((Ct+B)/m), \tag{25}$$

The expression (25) suggests the following admission control policy for the buffered model: accept a flow at time s if and only if $\bar{I}_{\Gamma^\circ,T^*(m)}((CT^*(m)+B)/m)) \ge m^{-1}I_\eta$, where $T^*(m)$ is the value of t achieving the minimum in (25). Unfortunately, to implement this decision rule one requires both the value of $T^*(m)$ and the value of $\bar{I}_{\Gamma^\circ,T^*(m)}$, neither of which is known *a priori*. To circumvent this difficulty, we take a fixed value T and estimate $I_{\Gamma^\circ,T}$ based on past measurements.

check: Where was N in original paper?

The constant T is also used as the window-length over which the statistics of the arrivals are estimated: For each time t we denote the M empirical moments by,

$$\widehat{\xi}(t,T) = \frac{1}{Nm}\sum_{j=1}^{Nm} c(X_j(t-T,t)). \tag{26}$$

This can be expressed $\widehat{\xi}(t,T) = \widehat{\Gamma}_{t,T}(c)$, where the empirical distributions are now defined by,

$$\widehat{\Gamma}_{t,T}(k) = \frac{1}{Nm}\sum_{j=1}^{Nm} \mathbb{I}\{X_j(t-T,t) = k\}\,, \qquad k \in \mathbb{Z}_+. \tag{27}$$

The set of marginals $\mathcal{M}$ is redefined as probability distributions supported on $[0, RT]$, and we maintain the definition of consistent marginals $\mathcal{M}_\xi$ given in (12) for $\xi \in \Delta$. Moment-consistent algorithms in the buffered model are defined exactly as in Section 2, based on a moment consistent map $\Gamma_{(\,\cdot\,)}$. Given the empirical estimate $\widehat{\xi} = \widehat{\xi}(t,T) \in \mathbb{R}^M$, a new flow request is admitted if and only if

$$\bar{I}_{\Gamma_{\widehat{\xi}},T}((CT+B)/m) \geq m^{-1} I_\eta.$$

Equivalently, a flow is admitted if and only if $\widehat{\xi}$ belongs to the acceptance region,

$$\Delta_m := \left\{\xi \in \mathbb{R}^M : \bar{I}_{\Gamma_{\widehat{\xi}},T}((CT+B)/m) \geq m^{-1} I_\eta\right\}. \tag{28}$$

As in the bufferless case, Theorem 2.1 motivates the following algorithm based upon the canonical distributions obtained from a given set of moment estimates.

Algorithm $\mathcal{A}^{**}$: If a flow arrives at time t, it is accepted if, and only if $\widehat{\xi} \in \Delta_m^*$, where $\widehat{\xi} = \widehat{\xi}(t)$ is the vector of empirical moments, Δ_m^* is defined by,

$$\Delta_m^* := \left\{\xi \in \mathbb{R}^M : \bar{I}_{\Gamma_\xi^*,T}(C/m) \geq m^{-1} I_\eta\right\}, \tag{29}$$

and mN is equal to the number of flows accessing the server at time t. ∎

An important issue in implementing the algorithm $\mathcal{A}^{**}$ is the choice of a measurement window length T. Ideally one would like to adapt T so that $T = T^*(m)$, where $T^*(m)$ is the most likely burst period before overflow when there are Nm flows in the system. When knowledge of this burst time-scale is not immediately available one can obtain approximations [8, 14], or bounds by making use of the declared parameters of the flows (such as maximum burst length in ATM.) However, simulation results presented in Section 5 suggest that the overflow probability is relatively insensitive to the exact value of T used in the implementation of the algorithm $\mathcal{A}^{**}$.

4.2 Flow model

To complete the description of the buffered model we specify the flow dynamics as above. The controlled process $\mathbf{\Phi}$ representing the number of flows in the system is again modeled as a Markov chain with transition probabilities,

$$P(i,i+1) = \lambda \mathsf{P}[\widehat{\xi}(t,T) \in \Delta_{i/N}], \quad P(i+1,i) = i\mu/N, \qquad i \in \{0 \ldots NW-1\}\,, \tag{30}$$

where $\widehat{\xi}(t,T)$ is defined in (26) with $m = i/N$.

In the asymptotic regime $N \to \infty$, we may employ a time-scale separation heuristic to reasonably approximate the overflow probability using (25). Given that there are Nm flows in the

only conditionally iid before: , as the packet arrival process $\boldsymbol{A}$ is not i.i.d.. In fact, the transition probabilities are no longer time-homogeneous in the buffered model.

system, the conditional probability of overflow may be approximated as $\mathsf{P}[\text{overflow} \mid Nm] \approx q_{Nm}$, where

$$\lim_{N\to\infty} N^{-1} \log q_{Nm} = -m \inf_{t\geq 0} \bar{I}_{\Gamma,t}((Ct+B)/m).$$

The steady probability of overflow η may then be approximated as,

$$\eta = \sum_{j=1}^{\infty} \pi_j q_j,$$

where π is the invariant distribution for the Markov chain $\mathbf{\Phi}$.

4.3 Performance analysis

The performance analysis for a moment-consistent algorithm in the buffered model is similar to that for the bufferless model since the transition probabilities of the corresponding Markov chains are of the same form given in (19). The results for the invariant distribution π and the probability of overflow η are analogous, and are described below.

In the statement of our results below we reuse symbols earlier used for quantities in the bufferless case, for the corresponding quantities in the buffered case. In particular, the common marginal distribution of $\{X_i(0,T)\}$ is denoted Γ°.

The first result is a statement of the optimality of algorithm $\mathcal{A}^{**}$:

Theorem 4.1 *Let $\eta^* = \eta^*_M$ denote the steady-state overflow probability under algorithm $\mathcal{A}^{**}$; η_M the corresponding quantity for an arbitrary moment-consistent algorithm $\mathcal{A}$; and let η^*_∞ denote the overflow probability for certainty-equivalent algorithm of Section 2.2. Then for each fixed $N \geq 1$ we have $\eta^*_M \leq \min(\eta^*_\infty, \eta_M)$, and moreover,*

$$\lim_{M\to\infty} \eta^*_M = \eta^*_\infty.$$

□

The following result is analogous to Theorem 3.2. First we define m^* as,

$$m^* := \sup\{m : \xi \in \Delta^*_m\} = \sup\left\{m : m\bar{I}_{\Gamma^*_{\xi^\circ},T}((CT+B)/m) \geq I_\eta\right\}, \tag{31}$$

where $\xi^\circ := \Gamma^\circ(c)$. We impose the following technical assumptions:

(A1) The vector mean $\xi^\circ := \Gamma^\circ(c)$ lies in the interior of Δ.

(A2) $C + B < Rm^*$, so that $\bar{I}_{\Gamma,1}((C+B)/m^*) = I_{\Gamma,1}((C+B)/m^*) < \infty$.

(A3) The infimization in (25) can be restricted to $0 \leq t \leq T_{\max}$ for a fixed constant $T_{\max} > 0$:

$$\inf_{t\geq 0} \bar{I}_{\Gamma,t}((Ct+B)/m) \;=\; \inf_{0\leq t\leq T_{\max}} \bar{I}_{\Gamma,t}((Ct+B)/m) \qquad \text{for } 0 < m \leq W$$

The third assumption puts a bound on the value of the critical time-scale $T^*(m)$ for any m. This is reasonable since in any practical system, one would not expect overflows to take place over an unbounded period. Moreover, a bound on the measurement window T is required in any practical implementation, and such a bound is provided by $T_{\max}$.

The following result, analogous to Theorem 3.2, asserts that the algorithm $\mathcal{A}^{**}$ behaves like a threshold policy for large N, with threshold m^*N. Define $K\colon \mathbb{R}_+ \to \mathbb{R}_+$ by,

$$K(m) = \inf\{mI_{M,\Gamma^\circ,T}(\xi) : \xi \in \mathrm{cl}(\Delta_m^*)\}$$

where $I_{M,\Gamma^\circ,T}(\cdot)$ denotes the M-dimensional large deviations rate-function for Γ°, defined as in (20).

really should introduce Λ_T

Proposition 4.2 *For the algorithm $\mathcal{A}^{**}$,*

$$\begin{aligned} \lim_{N\to\infty} \mathsf{P}[\widehat{\xi}(t,T) \in \Delta_m^*] &= 1, & m < m^* \\ \liminf_{N\to\infty} \mathsf{P}[\widehat{\xi}(t,T) \in \Delta_m^*] &> 0, & m = m^* \\ \limsup_{N\to\infty} \frac{1}{N} \log \mathsf{P}[\widehat{\xi}(t,T) \in \Delta_m^*] &\le -K(m), & m > m^*. \end{aligned} \tag{32}$$

□

The next result, similar to Proposition 3.3, demonstrates that for large N, the invariant distribution π is again similar to that of an $M/M/d/d$ queue with $d = Nm^*$. Recall that $L\colon \mathbb{R}_+ \to \mathbb{R}_+ \cup \infty$ is defined in (21).

Proposition 4.3 *The invariant distribution π of the Markov chain under algorithm $\mathcal{A}^{**}$ satisfies the following:*

(i) *The probability mass is concentrated around $m^\bullet N$, i.e.,*

$$\lim_{N\to\infty} \sum_{j=\lfloor N(m^\bullet-\epsilon)\rfloor}^{\lfloor N(m^\bullet+\epsilon)\rfloor} \pi(j) = 1 \qquad \text{for every } \epsilon > 0 \tag{33}$$

(ii) *For each $m \ge 0$, $\lim_{N\to\infty} N^{-1}\log(\pi(\lfloor Nm\rfloor)) = -L(m)$.* □

An asymptotic expression for the probability of overflow η^* in the buffered model follows as in Section 3,

Theorem 4.4 *The steady state probability of overflow probability η^* for algorithm $\mathcal{A}^{**}$ satisfies the large deviations principle,*

$$\lim_{N\to\infty} N^{-1}\log(\eta^*(N)) = -\inf_{m>0}\left\{L(m) + \inf_t (m\bar{I}_{\Gamma,t}((Ct+B)/m))\right\}. \tag{34}$$

Moreover, the infimum is achieved at $m = m^*$ if $m^* \le \rho$. □

5 Simulations

We have seen that the algorithms $\mathcal{A}^*$ and $\mathcal{A}^{**}$ minimize the buffer-overflow probability among all moment-consistent algorithms. We now examine the performance of these algorithms through simulation.

In particular, we seek answers to the following questions: (i) How do the algorithms $\mathcal{A}^*$, $\mathcal{A}^{**}$ compare to the Gaussian algorithm of [22, 11]? (ii) How robust are the algorithms $\mathcal{A}^*$, $\mathcal{A}^{**}$ to variations in the arrival rate λ and the nominal departure rate μ? (iii) How sensitive is the algorithm $\mathcal{A}^{**}$ to the measurement window length T?

Note that Theorem 3.1 tells us nothing about the relative performance of the Gaussian algorithm with respect to any extremal moment-consistent algorithm. The Gaussian algorithm is *not* a moment-consistent algorithm since a non-trivial Gaussian distribution is not supported on $[0, R]$.

We find that in both the bufferless and the buffered model, algorithms $\mathcal{A}^*$ and $\mathcal{A}^{**}$ have a lower buffer-overflow probability than the Gaussian algorithm. In addition, these algorithms have a better *trade-off* curve than the Gaussian algorithm, in terms of bandwidth utilization versus probability of overflow. Furthermore, the results indicate that the algorithms proposed here are less sensitive to changes in λ, μ and T than the Gaussian algorithm.

In order to compare the algorithms $\mathcal{A}^*$, $\mathcal{A}^{**}$ to the Gaussian algorithm we have set $M = 2$, and estimate first and second moments exactly as in the Gaussian algorithm. Recall that the algorithms are denoted $\mathcal{A}^*(2)$, $\mathcal{A}^{**}(2)$, respectively, in this case. The canonical distribution Γ^* for two moment measurements is given in (14).

We first concentrate on the bufferless case in which packet arrivals are i.i.d., with parameters identified in Table 1. The source distribution is either uniform or discrete. Further details are given below.

Parameter	Description	Values
N	Scale parameter	40
η	Target overflow probability	10^{-5}
λ	Flow arrival rate	$0.05 - 0.5$
μ	Nominal departure rate	$0.03 - 0.09$
C	Nominal capacity	5.0
B	Nominal buffer size	0.0/2.0
R	Peak rate	8.0

Table 1: List of parameter values

Performance comparison In the first simulation we use an i.i.d. source whose marginal distribution is uniform over $[0, 8]$. The plot at left in Figure 3 provides a comparison of $\mathcal{A}^*(2)$ and the Gaussian algorithm for a range of arrival rates from 0.05 to 0.5. From the figure we see that the algorithm $\mathcal{A}^*(2)$ achieves the target overflow probability for a large range of arrival rates, and violates the target marginally for high arrival rates. For high arrival rates, the effect of a small error in measuring moments results in the erroneous admission of a significant number of

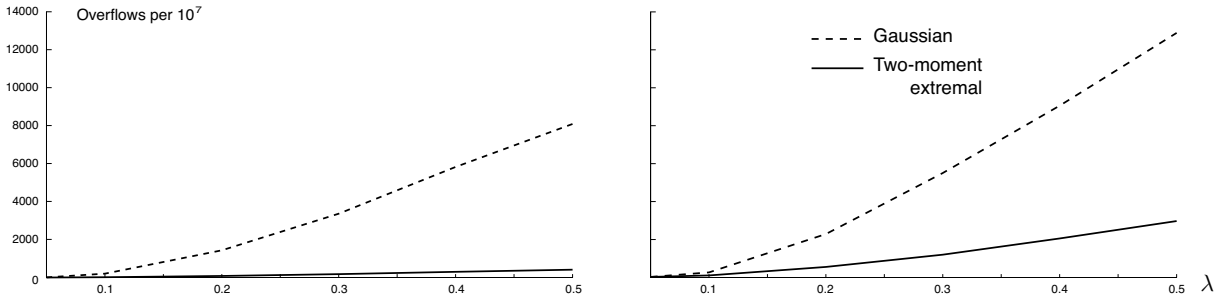

Figure 3: *Performance in the bufferless model.* A comparison of algorithm $\mathcal{A}^*(2)$ with the Gaussian algorithm for a range of arrival rates. The figure at left shows results from experiments with a uniform source distribution, and at right the source distribution was discrete.

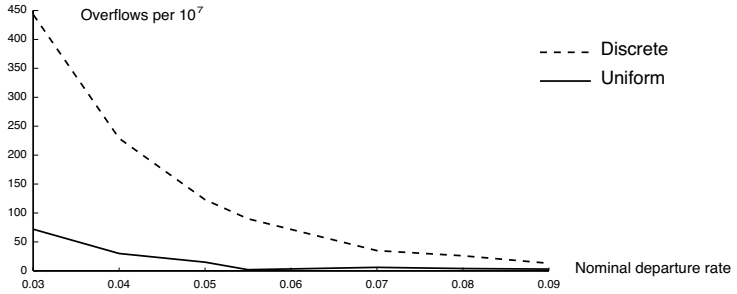

Figure 4: *Sensitivity to the departure rate μ.* Performance of algorithm $\mathcal{A}^*(2)$ over a range of departure rates. The two plots were obtained using a discrete source, and a uniform source, respectively.

flows, thus increasing the overflow probability. For low arrival rates the impact of measurement error is small.

This observation also holds true for the Gaussian algorithm, with a clear degradation in performance with increasing arrival rate. In comparison with $\mathcal{A}^*(2)$, however, the Gaussian algorithm performs much worse, violating the target probability by as much as a factor of 80 for an arrival rate of 0.5.

At right in Figure 3 we see the results from an analogous set of experiments in which the source distribution is discrete, rather than uniform. The marginal distribution is given by $0.33\,\delta_{0.5} + 0.33\,\delta_{3.5} + 0.34\,\delta_{8.0}$. Both algorithms perform relatively worse in this case since the source distribution is more bursty. However, algorithm $\mathcal{A}^*(2)$ still performs better than the Gaussian algorithm. Its overflow probability is approximately four times lower than the Gaussian algorithm for the highest arrival rates.

The trade-off curves for each algorithm are illustrated in Figure 1 for the uniform source. Each plot shows the probability of overflow versus the average number of flows in the system (i.e., bandwidth utilization.) The trade-off curve for the algorithm $\mathcal{A}^*$ lies below the corresponding curve for the Gaussian algorithm, meaning that for the same level of bandwidth utilization, algorithm $\mathcal{A}^*$ achieves a lower probability of overflow.

Figure 4 shows the sensitivity of the algorithm $\mathcal{A}^*$ to the nominal departure rate μ for both uniform and discrete sources. The performance of $\mathcal{A}^*$ is not very sensitive to changes in the value of the departure rate for the uniform source; The sensitivity is moderate for the more

bursty discrete source.

Buffered case Simulations of a buffered server were carried out using an on/off source which sends packets at a rate of 8.0 when on, and transitions between on and off states at rate given by $100 * \mu/N$. Thus, on average, the on/off source switches state 100 times during the lifetime of a flow. This is consistent with our implicit assumption of time-scale separation.

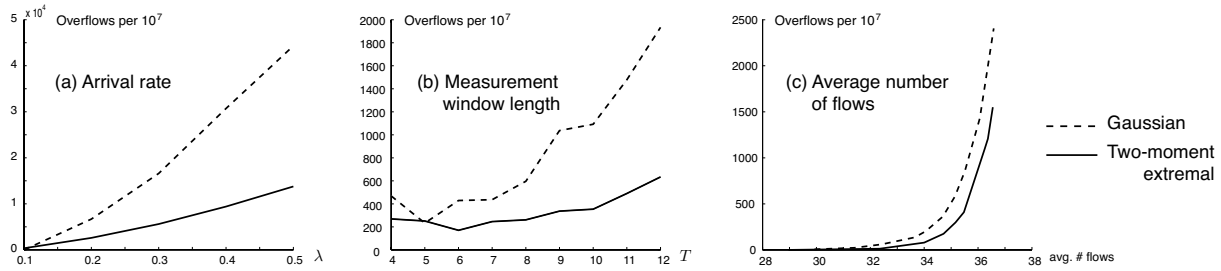

Figure 5: *Performance in the buffered model.* The algorithm $\mathcal{A}^{**}(2)$ and the Gaussian algorithm were compared in several experiments using an on/off source. In each plot, the vertical axis indicates the overflow probability. The two plots at far left compare the performance of the two algorithms over a range of arrival rates. Both algorithms rely on an estimate of the critical time-scale. The plot at center shows that sensitivity to the parameter T is very low in the algorithm $\mathcal{A}^{**}(2)$. Finally, the plot at right shows how performance degrades in each model as the average number of flows in the system increases.

Shown in Figure 5 are several plots comparing the performance of the algorithm $\mathcal{A}^{**}(2)$ and the Gaussian algorithm in the buffered model. In each plot, the vertical axis indicates the empirical overflow probability.

Note that the performance is poor when compared to the results shown in bufferless case, especially when the arrival rate is high. This is mainly because the on/off source used in these simulations is more bursty than the i.i.d. sources used in the bufferless case. Also, for finite N, the error in the formula given by the *many sources asymptotic* (25) may be significant.

Qualitatively, the results are analogous to those obtained in the bufferless model: In Figure 5 (a) we see that the algorithm $\mathcal{A}^{**}(2)$ is relatively insensitive to the arrival rate λ when compared to the Gaussian algorithm. Figure 5 (c) shows the trade-off curves for each of the two algorithms. The curve for $\mathcal{A}^{**}(2)$ lies below the corresponding curve for the Gaussian algorithm, meaning that for the same level of bandwidth utilization, $\mathcal{A}^{**}(2)$ achieves a lower overflow probability.

Finally, we note that the critical time-scale T^* is not known exactly. In practice, either algorithm will be implemented using an approximate value T. Figure 5 (b) illustrates the performance as a function of the parameter T for each algorithm based on a single source model. The performance of the the algorithm $\mathcal{A}^{**}(2)$ does not vary significantly with a change in the value of T. Sensitivity is higher in the Gaussian algorithm.

6 Conclusion

In this paper we introduced a new class of algorithms for measurement-based admission control, together with a portrait of closed-loop behavior through both analysis and simulation.

The main idea in the construction of these algorithms is to avoid explicit rate-function estimation, but instead search for useful rate-function *bounds*. These bounds are all based upon the central result Theorem 2.1, which is itself based upon the theory of canonical distributions. By exploiting the simple structure of canonical distributions we obtain simple, effective algorithms for admission control.

Many questions and open problems remain:

(i) Alternative data sets may be used to improve performance: Instead of estimating moments, one can instead estimate $\mathsf{E}[f_i(X)]$, $i = 1, \ldots, n$ for suitably chosen functions f_i based on recent analytical results in [2, 3, 34].

(ii) Another issue to be considered is numerical computation. In particular, we do not know how to efficiently compute the canonical distribution for large values of M [3, 34].

(iii) The ideas developed in this paper may also be used in conjunction with other methods of admission control, such as virtual buffers [8]. Moreover, in systems such as ATM, it may be possible to include extra information about the flows being served. These could include mean rate specifications and/or maximum burst length specifications.

It is surprising that the ideas of Markov described in [32] have not had greater impact in systems theory and statistical modeling. The viewpoint and the results of [32] have had substantial impact on our own recent research. In particular,

(i) Motivated in part by canonical distributions we have constructed in [25, 34] simple discrete queueing models for the purposes of simulation and control. This has provided a setting to extend the variance reduction techniques introduced in [24] to a very general class of network models.

(ii) In [28] the theory of canonical distributions has motivated a new class of algorithms for the computation of efficient channel codes based on optimal discrete input distributions.

(iii) We are presently considering the application of canonical distributions to robust hypothesis testing, and to source coding. Some preliminary results are contained in [34, 27]. The paper [34] includes results from numerical experiments that illustrate how a worst-case rate function depends on the dimension M of the constraint vector.

We are convinced that the theory of canonical distributions will have significant impact in many other areas that involve statistical modeling and prediction.

A Appendix: Lemmas and Proofs

We begin with some useful lemmas. We omit the proof of the following result, which follows from standard large deviations theory (see in particular [10, Section 2.2].)

Lemma A.1 *For each $\Gamma \in \mathcal{M}_\xi$, the function $g_\Gamma(m) := m\bar{I}_\Gamma(C/m), m > 0$, has the following properties:*

(i) *$g_\Gamma(m)$ is monotone decreasing.*

(ii) *$g_\Gamma(m)$ is finite, continuous and strictly decreasing on $(C/R, C/\xi_1]$.*

(iii) $g_\Gamma(m) = \infty$ *for* $m \le C/R$, *and* $g_\Gamma(m) = 0$ *for* $m > C/\xi_1$. □

The worst-case rate function is convex as a function of the vector of moment constraints:

Lemma A.2 *For any* $r \in \mathbb{R}_+$, *the function* $h : \Delta \to \mathbb{R}$ *given by* $h(\xi) = \bar{I}_{\Gamma^*_\xi}(r)$ *is convex.*

Proof. Suppose that ξ^1, ξ^2 are two vectors in the set Δ, and let Γ^1, Γ^2 be two probability distributions satisfying $\Gamma^i \in \mathcal{M}_{\xi^i}$ for each i. For each $\alpha \in [0,1]$ define $\Gamma^\alpha = \alpha\Gamma^1 + (1-\alpha)\Gamma^2$. We evidentally have $\Gamma^\alpha \in \mathcal{M}_\xi$ with $\xi = \alpha\,\xi^1 + (1-\alpha)\,\xi^2$, and also $\bar{I}_{\Gamma^\alpha}(r) \le \alpha\bar{I}_{\Gamma^1}(r) + (1-\alpha)\bar{I}_{\Gamma^2}(r)$ from the convexity of $\bar{I}_\Gamma(r)$. Consequently,

$$\bar{I}_{\Gamma^*_\xi}(r) \;\le\; \bar{I}_\Gamma(r) \;\le\; \alpha\bar{I}_{\Gamma^1}(r) + (1-\alpha)\bar{I}_{\Gamma^2}(r) \tag{35}$$

Infimizing the expression on the right hand side of (35) over all $\Gamma^i, i = 1, 2$ satisfying the moment constraints establishes the desired convexity. □

The following is required in the proof of Theorem 3.1. The proof follows from the fact that $\widehat{\Gamma}_t$ has at most NW points of support, and is hence determined by its first $2NW$ moments.

Lemma A.3 *The empirical distributions and extremal distributions coincide,* $\Gamma^*_{\widehat{\xi}} = \widehat{\Gamma}_t$, *for all* $M \ge 2NW$. □

Proof of Theorem 3.1. We first prove that $\eta^*_M \le \min(\eta_M, \eta^*_\infty)$ using an application of *likelihood ratio ordering* [7, Section 1.3]. Let π^* be the invariant distribution of the Markov chain under algorithm $\mathcal{A}^*$, π the corresponding invariant distribution under a general moment-consistent algorithm $\mathcal{A}$, and π^*_∞ be the corresponding distribution under the certainty-equivalent algorithm. Since η^* is a sum of the form $\sum_{j=0}^{NW} \pi^*_j q_j$ and q_j is clearly an increasing sequence, in order to use [7, Lemma 1.14], we have to verify that $\pi^*(j)/\pi^*(j+1) \ge \max(\pi(j)/\pi(j+1), \pi^*_\infty(j)/\pi^*_\infty(j+1))$ for $0 \le j \le NW$. This is equivalent to showing that

(i) $\mathsf{P}[\widehat{\xi}(t) \in \Delta^*_{j/N}] \le \mathsf{P}[\widehat{\xi}(t) \in \Delta_{j/N}], j = 0, \dots, NW$ for every acceptance region Δ_m obtained from a moment-consistent algorithm, and

(ii) $\mathsf{P}[\widehat{\xi}(t) \in \Delta^*_{j/N}] \le \mathsf{P}[j\bar{I}_{\widehat{\Gamma}_t}(NC/j) \ge I_\eta]$.

From Theorem 2.1, we know that $j\bar{I}_{\Gamma^*_\xi}(NC/j) \le j\bar{I}_{\Gamma_\xi}(NC/j)$ for every moment-consistent map Γ and for every vector $\xi \in \Delta$. Thus $\Delta^*_{j/N} \subset \Delta_{j/N}$, and the first condition is verified. To verify the second condition, note that $\widehat{\Gamma}_t$ satisfies $\widehat{\Gamma}_t(c) = \widehat{\xi}(t)$. Together with Theorem 2.1 this implies $j\bar{I}_{\Gamma^*_{\widehat{\xi}}}(NC/j) \le j\bar{I}_{\widehat{\Gamma}_t}(NC/j)$ when $\widehat{\xi} = \widehat{\xi}(t)$, and (ii) then follows from the definition of Δ^*.

Finally, the claim that that the algorithm $\mathcal{A}^*$ approaches the certainty-equivalent algorithm when $M \to \infty$ follows from Lemma A.3. □

Proof of Theorem 3.2. For $m > m^*$, the LDP result follows from Cramér's Theorem for $\mathbb{R}^M$-valued random variables. The rate-function I_{M,Γ° is known to be lower semi-continuous [10, Section 2.2]. Consequently, since $\mathrm{cl}(\Delta_m^*)$ is closed, one can find $c \in \mathrm{cl}(\Delta_m^*)$ satisfying $K(m) = mI_{M,\Gamma^\circ}(c)$. Finally, $K(m) = mI_{M,\Gamma^\circ}(c) > 0$ since $\xi^\circ \notin \mathrm{cl}(\Delta_m^*)$ for $m > m^*$.

For $m = m^*$ we apply Lemma A.2, which implies that the set $\{\Delta_{m^*}^*\}^c$ is convex. Consequently, there exists an M-dimensional half-space B satisfying $\xi \in B \subset \Delta_{m^*}^*$. This implies the bound,

$$\mathsf{P}[\widehat{\xi}(t) \in \Delta_m^*] = \mathsf{P}[\sqrt{N}(\widehat{\xi}(t) - \xi) \in \sqrt{N}(\Delta_m - \xi)] \geq \mathsf{P}[\sqrt{N}(\widehat{\xi}(t) - \xi) \in \sqrt{N}(B - \xi)], \tag{36}$$

and hence by the Central Limit Theorem,

$$\liminf_{N\to\infty} \mathsf{P}[\sqrt{N}(\widehat{\xi}(t) - \xi) \in (B - \xi)] = \liminf_{N\to\infty} \mathsf{P}[\sqrt{N}(\widehat{\xi}(t) - \xi) \in B] > 0. \tag{37}$$

Putting (36) and (37) together, we obtain the desired result, $\liminf_{N\to\infty} \mathsf{P}[\widehat{\xi}(t) \in \Delta_{m^*}^*] > 0$.

The result for $m < m^*$ follows similarly, based on the weak law of large numbers. ☐

The next two results will be applied repeatedly in the proofs that follow.

Lemma A.4 follows directly from the definition of Δ^* and property (i) in Lemma A.1.

Lemma A.4 *The sets defined in (29) are monotone decreasing on* $(0, \infty)$*:* $\Delta_{m_2}^* \subset \Delta_{m_1}^*$ *for all* $m_1 \geq m_2 > 0$. ☐

Lemma A.5

$$\frac{\pi(\lfloor Nv \rfloor)}{\pi(\lfloor Nu \rfloor)} = \prod_{i=\lfloor Nu \rfloor}^{\lfloor Nv \rfloor - 1} \left(\frac{N\lambda}{i\mu}\right) \mathsf{P}\{\widehat{\xi}(t) \in \Delta_{i/N}^*\}$$

$$= \left(\frac{(N\rho)^{\lfloor Nv \rfloor - \lfloor Nv \rfloor}(\lfloor Nu \rfloor - 1)!}{(\lfloor Nv \rfloor - 1)!}\right) \prod_{i=\lfloor Nu \rfloor}^{\lfloor Nv \rfloor - 1} \mathsf{P}\{\widehat{\xi}(t) \in \Delta_{i/N}^*\}$$

Proof. This follows from the detailed-balance equations,

$$\pi(j)P(j, j+1) = \pi(j+1)P(j+1, j)\,, \qquad \sum_{j\geq 0} \pi(j) = 1\,, \tag{38}$$

which is a consequence of the skip-free property for $\mathbf{\Phi}$. ☐

Proposition A.6 is based on Lemma A.5 combined with Stirling's formula.

Proposition A.6 *The following hold for the algorithm* $\mathcal{A}^*$*:*

(i) *If* $0 \leq u \leq v \leq m^*$ *then*

$$\lim_{N\to\infty} N^{-1} \log\left(\frac{\pi(\lfloor Nv \rfloor)}{\pi(\lfloor Nu \rfloor)}\right) = -(L(v) - L(u))$$

(ii) $\lim_{N\to\infty} N^{-1} \log(\pi(\lfloor Nm^\bullet \rfloor)) = 0 = -L(m^\bullet)$*, with* $m^\bullet := \min(m^*, \rho)$.

(iii) $\lim_{N\to\infty} N^{-1} \log(\pi([\lfloor Nv \rfloor, \infty))) = -\infty = -L(m)$ *for* $m > m^*$.

Proof. We first establish (iii). Defining $m' := (m + m^*)/2$ for $m > m^*$, we have from Lemma A.5,

$$\pi(\lfloor Nm \rfloor) \leq \pi(\lfloor Nm' \rfloor) \left(\frac{\rho}{m}\right)^{\lfloor Nm \rfloor - \lfloor Nm' \rfloor} \prod_{i=\lfloor Nm' \rfloor}^{\lfloor Nm \rfloor - 1} \mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{i/N}\}$$

The right hand side above can be further bounded by using the inequalities, (i) $\rho/m \leq \rho/m^*$, (ii) $\pi(\lfloor Nm' \rfloor) \leq 1$, and (iii) $\mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{i/N}\} \leq \mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{\lfloor Nm' \rfloor/N}\}$ for $\lfloor Nm' \rfloor \leq i \leq \lfloor Nm \rfloor$ (from Lemma A.4), to obtain:

$$\pi(\lfloor Nm \rfloor) \leq \left(\frac{\rho}{m^*} \mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{\lfloor Nm' \rfloor/N}\}\right)^{\lfloor Nm \rfloor - \lfloor Nm' \rfloor}$$

Now from Theorem 3.2, we have

$$\limsup_{N\to\infty} N^{-1} \log\left(\mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{\lfloor Nm' \rfloor/N}\}\right) \leq -K(m') < 0,$$

and combining this with the previous inequality we conclude that

$$\lim_{N\to\infty} N^{-1} \log(\pi(\lfloor Nm \rfloor)) = -\infty, \qquad \text{for } m > m^*,$$

which establishes (iii).

We now turn to (i). Stirling's formula gives,

$$(\lfloor Nu \rfloor - 1)! \approx \sqrt{2\pi(\lfloor Nu \rfloor - 1)} \left(\frac{\lfloor Nu \rfloor - 1}{e}\right)^{\lfloor Nu \rfloor - 1}$$

And substituting this into the formula given in Lemma A.5 we obtain,

$$\begin{aligned} \lim_{N\to\infty} N^{-1} \log\left((N\rho)^{\lfloor Nv \rfloor - \lfloor Nu \rfloor} \frac{(\lfloor Nu \rfloor - 1)!}{(\lfloor Nv \rfloor - 1)!}\right) &= (v-u)\log\rho + u\log(u) - v\log(v) + u - v \\ &= -(L(v) - L(u)) \end{aligned} \tag{39}$$

Thus in order to prove (i) it is sufficient to show that

$$\lim_{N\to\infty} \frac{1}{N} \log\left(\prod_{i=\lfloor Nu \rfloor}^{\lfloor Nv \rfloor - 1} \mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{i/N}\}\right) = 0. \tag{40}$$

If $v < m^*$ then from Theorem 3.2,

$$\lim_{N\to\infty} \log \mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{\lfloor Nv \rfloor/N}\} = 0,$$

which implies (40).

If on the other hand we have $v = m^*$ and $u < v$, then the proof is more complex. For $\epsilon > 0$, we define $m^- = m^* - \epsilon$. Then from the above arguments we know that

$$\lim_{N\to\infty} \frac{1}{N} \log\left(\prod_{i=\lfloor Nu \rfloor}^{\lfloor Nm^- \rfloor - 1} \mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{i/N}\}\right) = 0$$

Using Lemma A.4 again, we obtain the bound,

$$\mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{\lfloor Nm^* \rfloor/N}\}^{N\epsilon+1} \;\leq\; \prod_{i=\lfloor Nm^- \rfloor}^{\lfloor Nm^* \rfloor - 1} \mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{i/N}\}$$

From Theorem 3.2, we know that $z := \liminf_{N\to\infty} \mathsf{P}[\widehat{\xi}(t) \in \Delta^*_{m^*}] > 0$. Combining this bound with the above inequalities we obtain,

$$\liminf_{N\to\infty} \frac{1}{N} \log \left(\prod_{i=\lfloor Nu \rfloor}^{\lfloor Nm^* \rfloor - 1} \mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{i/N}\} \right) = \liminf_{N\to\infty} \frac{1}{N} \log \left(\prod_{i=\lfloor Nm^- \rfloor}^{\lfloor Nm^* \rfloor - 1} \mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{i/N}\} \right) \geq \epsilon \log z.$$

Since $\epsilon > 0$ was arbitrary, we again obtain (40), which completes the proof of (i).

In the proof of (ii) we distinguish between two cases: $m^* \leq \rho$ and $m^* > \rho$. Consider first the case when $m^* \leq \rho$, so that $m^\bullet = m^*$. Fix an $\epsilon > 0$, and define $m^- := m^* - \epsilon$ and $m^+ := m^* + \epsilon$. The proof consists of bounding the sum $\sum_{j=0}^{\infty} \pi(j)$ by separating it into three smaller sums, between the limits $\{0, \lfloor Nm^- \rfloor\}, \{\lfloor Nm^- \rfloor, \lfloor Nm^+ \rfloor\}$, and $\{\lfloor Nm^+ \rfloor, \infty\}$.

Using Lemma A.5 we have for $j \leq k \leq \lfloor Nm^* \rfloor$,

$$\left(\frac{\pi(k)}{\pi(j)} \right) \geq \prod_{i=j}^{k-1} \mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{i/N}\} \geq \mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{k/N}\}^{k-j} \geq \mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{k/N}\}^{k}, \tag{41}$$

where the first inequality follows from $N\lambda/j\mu \geq 1$ for $j \leq \lfloor Nm^* \rfloor \leq \lfloor N\rho \rfloor$, and the second inequality uses Lemma A.4. Using the above inequalities, we obtain,

$$\sum_{j=0}^{\lfloor Nm^- \rfloor} \pi(j) \;\;\leq\;\; \pi(\lfloor Nm^- \rfloor) \left(\lfloor Nm^- \rfloor \mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{m^-}\}^{-\lfloor Nm^- \rfloor} \right) \tag{42}$$

Now for $\lfloor Nm^- \rfloor \leq j \leq \lfloor Nm^+ \rfloor$, it is easy to see that $\pi(j) \leq (\rho/m^-)^{2N\epsilon+1} \pi(\lfloor Nm^- \rfloor)$. Summing this inequality for all such j, we have,

$$\sum_{j=\lfloor Nm^- \rfloor}^{\lfloor Nm^+ \rfloor} \pi(j) \;\leq\; \pi(\lfloor Nm^- \rfloor) \left((2N\epsilon + 1) \left(\frac{\rho}{m^-} \right)^{2N\epsilon+1} \right) \tag{43}$$

The sum $\sum_{j=\lfloor Nm^+ \rfloor}^{\infty} \pi(j)$ is negligible since we have already established Proposition A.6 (iii).

Adding the inequalities (42), (43) we obtain, for large enough N,

$$\begin{aligned} 1 \;\;&=\;\; \textstyle\sum_{j=0}^{\infty} \pi(j) \\ &\leq\;\; \pi(\lfloor Nm^- \rfloor) \Big[\lfloor Nm^- \rfloor \mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{m^-}\}^{-\lfloor Nm^- \rfloor} + 2(N\epsilon + 1) \left(\tfrac{\rho}{m^-} \right)^{2N\epsilon+1} \Big] \end{aligned} \tag{44}$$

From Theorem 3.2, we have $\lim_{N\to\infty} \log \mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{m^-}\} = 0$. Combining this with (44) gives,

$$\liminf_{N\to\infty} N^{-1} \log(\pi(\lfloor Nm^- \rfloor)) \;\geq\; -2\epsilon \log \left(\frac{\rho}{m^*} \right)$$

Now using Lemma A.6 (i) along with the above inequality, we have

$$\liminf_{N\to\infty} N^{-1}\log(\pi(\lfloor Nm^*\rfloor)) \geq -(L(m^*)-L(m^-)) - 2\epsilon\log\left(\frac{\rho}{m^*}\right)$$

The function L is left-continuous; thus $L(m^-)\to L(m^*)$ when $\epsilon\downarrow 0$. We thus conclude,

$$\liminf_{N\to\infty} N^{-1}\log(\pi(\lfloor Nm^*\rfloor)) \geq 0,$$

which establishes (ii) in the special case $m^\bullet = m^* \leq \rho$.

Consider now the case when $m^* > \rho$, i.e., $m^\bullet = \rho$. The proof for this case is similar to that for $m^* \leq \rho$: We bound the sum $\sum_{j=0}^{\infty}\pi(j)$ by separating it into three smaller sums, between the limits $\{0,\lfloor N\rho\rfloor\},\{\lfloor N\rho\rfloor,\lfloor Nm^+\rfloor\}$ and $\{\lfloor Nm^+\rfloor,\infty\}$ (where $m^+ := m^*+\epsilon$ as before). The third sum is negligible, again by (iii).

As in (41), we have here, for $0\leq j\leq\lfloor N\rho\rfloor$,

$$\left(\frac{\pi(\lfloor N\rho\rfloor)}{\pi(j)}\right) \geq \prod_{i=j}^{\lfloor N\rho\rfloor-1} \mathsf{P}\{\widehat{\xi}(t)\in\Delta^*_{i/N}\} \geq \mathsf{P}\{\widehat{\xi}(t)\in\Delta^*_{\lfloor N\rho\rfloor/N}\}^{\lfloor N\rho\rfloor-j}$$

Summing the above inequality gives a bound on the first of the three sums:

$$\sum_{j=0}^{\lfloor N\rho\rfloor}\pi(j) \;\leq\; \pi(\lfloor N\rho\rfloor)\left(\lfloor N\rho\rfloor\mathsf{P}\{\widehat{\xi}(t)\in\Delta^*_{\lfloor N\rho\rfloor/N}\}^{-\lfloor N\rho\rfloor}\right) \tag{45}$$

In order to bound the second term we use the fact that $\pi(j)\leq\pi(\lfloor N\rho\rfloor)$ for $j>\lfloor N\rho\rfloor$ since $N\lambda/j\mu\leq 1$. Thus

$$\sum_{j=\lfloor N\rho\rfloor+1}^{\lfloor Nm^+\rfloor}\pi(j) \leq \pi(\lfloor N\rho\rfloor)(N(m^*+\epsilon-\rho)+1) \tag{46}$$

Adding the inequalities (45) and (46) we obtain for large n,

$$1=\sum_{j=1}^{\infty}\pi(j) \leq \pi(\lfloor N\rho\rfloor)\left(\lfloor N\rho\rfloor\mathsf{P}\{\widehat{\xi}(t)\in\Delta^*_{\lfloor N\rho\rfloor/N}\}^{-\lfloor N\rho\rfloor} + N(m^*+\epsilon-\rho)+2\right)$$

Now recall that in this case $\rho<m^*$, so according to Theorem 3.2, $\mathsf{P}\{\widehat{\xi}(t)\in\Delta^*_{\lfloor N\rho\rfloor/N}\}\to 1$ as $N\to\infty$. Thus we have,

$$\liminf_{N\to\infty} N^{-1}\log(\pi(\lfloor N\rho\rfloor)) \;\geq\; 0,$$

which completes the proof of part (ii). □

Proof of Proposition 3.3: First we prove part (ii). For $m\leq m^*$ the result follows on combining Proposition A.6 (i) and (ii),

$$\begin{aligned}\lim_{N\to\infty}\log(\pi(\lfloor Nm\rfloor)) &= \lim_{N\to\infty}\log\left(\frac{\pi(\lfloor Nm\rfloor)}{\pi(\lfloor Nm^\bullet\rfloor)}\pi(\lfloor Nm^\bullet\rfloor)\right)\\ &= -\left(L(m)-L(m^\bullet)\right)-L(m^\bullet) = -L(m)\end{aligned}$$

For $m > m^*$ the result follows from Proposition A.6 (iii).

To prove part (i) we consider separately the cases $m^* \leq \rho$ and $m^* > \rho$. For $m^* \leq \rho$, we define for fixed $\epsilon > 0$,

$$m^- := m^* - \epsilon, \quad m_- := m^* - \epsilon/2.$$

From Lemma A.5 we have, for $0 \leq j \leq \lfloor Nm^- \rfloor$,

$$\pi(j) \leq \pi(\lfloor Nm_- \rfloor) \left(\frac{\rho}{m_-}\right)^{j-\lfloor Nm_- \rfloor} \mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{\lfloor Nm_- \rfloor/N}\}^{-\lfloor Nm_- \rfloor},$$

giving

$$\sum_{j=0}^{\lfloor Nm^- \rfloor} \pi(j) \leq \pi(\lfloor Nm_- \rfloor) \left(\frac{\rho}{m_-}\right)^{-N\epsilon/2+1} (1 - m_-/\rho)^{-1} \mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{\lfloor Nm_- \rfloor/N}\}^{-\lfloor Nm_- \rfloor}$$

Now from Theorem 3.2, $\mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{\lfloor Nm_- \rfloor/N}\} \to 1$. We also have $\rho/m_- > 1$. Therefore, for large enough N we have

$$\sum_{j=0}^{\lfloor Nm^- \rfloor} \pi(j) \leq \pi(\lfloor Nm_- \rfloor)$$

This combined with Proposition A.6 shows that $\pi(\lfloor Nm^- \rfloor, \lfloor Nm^+ \rfloor) \to 1$ when $m^* \leq \rho$.

For $m^* > \rho$, we apply similar arguments. We define $\rho^- := \rho - \epsilon$, $\rho_- := \rho - \epsilon/2$, $\rho^+ := \rho + \epsilon$ and $\rho_+ := \rho + \epsilon/2$. We then have, for $j \leq \lfloor N\rho^- \rfloor$,

$$\pi(j) \leq \left(\frac{\rho}{\rho_-}\right)^{j-\lfloor N\rho_- \rfloor} \pi(\lfloor N\rho_- \rfloor) \mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{\lfloor N\rho_- \rfloor/N}\}^{-\lfloor N\rho_- \rfloor}.$$

And as in the case $m^* \leq \rho$,

$$\lim_{N\to\infty} \sum_{j=0}^{\lfloor N\rho^- \rfloor} \pi(j) = 0. \tag{47}$$

Now for $j \geq \lfloor N\rho^+ \rfloor$, we have $\pi(j) \leq \left(\frac{\rho}{\rho_+}\right)^{j-\lfloor N\rho_+ \rfloor} \pi(\lfloor N\rho_+ \rfloor)$, giving the bound

$$\sum_{j=\lfloor N\rho^+ \rfloor}^{\infty} \pi(j) \leq \pi(\lfloor N\rho_+ \rfloor) \left(\frac{\rho}{\rho_+}\right)^{N\epsilon/2-1} (1 - \rho/\rho_+)^{-1}.$$

Since $\rho/\rho_+ < 1$, we conclude that

$$\lim_{N\to\infty} \sum_{j=\lfloor N\rho^+ \rfloor}^{\infty} \pi(j) = 0. \tag{48}$$

The inequalities (47) and (48) imply that $\pi(\lfloor N\rho^- \rfloor, \lfloor N\rho^+ \rfloor) \to 1$ when $m^* > \rho$. □

Proof of Theorem 3.4: The proof of Theorem 3.4 is divided into two parts: we first establish the asymptotics of the steady-state overflow probability $\eta^*(N)$ under $\mathcal{A}^*$, and then show that the infimum in (24) is achieved at m^* when $m^* \leq \rho$.

Although the statement of the theorem is an affirmation of the usual large deviations principle that the term with the smallest exponent dominates the rate of decay of a sum, the proof is more complicated since the infimization (over m) of the exponents is not a finite minimization.

Since $\eta^* \geq \pi_{Nm} q_{Nm}$ for every m, we have using Proposition 3.3,

$$\liminf_{N\to\infty} \frac{1}{N} \log \eta^* = \liminf_{N\to\infty} \frac{1}{N} \log \left(\sum_{j=0}^{\infty} \pi(j) q_j \right) \geq - \inf_{m>0} (L(m) + m \bar{I}_{\Gamma^\circ}(C/m))$$

Thus we only need to establish the reverse inequality,

$$\limsup_{N\to\infty} \frac{1}{N} \log \left(\sum_{j=0}^{\infty} \pi(j) q_j \right) \leq - \inf_{m>0} (L(m) + m \bar{I}_{\Gamma^\circ}(C/m)) \tag{49}$$

We divide the proof into two cases, $m^* \leq \rho$ and $m^* > \rho$. We begin with the simpler case $m^\bullet = m^* \leq \rho$.

First we show that $\inf_{m>0}\{L(m)+m\bar{I}_{\Gamma^\circ}(C/m)\}$ is achieved at m^*, as required by the theorem Since $L(m) = \infty$ for $m > m^*$, it is sufficient to show that $L(m) + m\bar{I}_{\Gamma^\circ}(C/m)$ is a decreasing function on $(0, m^*]$. In fact, the function L is decreasing on $[0, m^*]$ from the definition (21), and $m\bar{I}_{\Gamma^\circ}(C/m)$ is also decreasing in m by Lemma A.1.

To obtain (49), first note that the right hand side of (49) is $-(L(m^*) + m^*\bar{I}_{\Gamma^\circ}(C/m^*)) = -m^*\bar{I}_{\Gamma^\circ}(C/m^*)$. From the definition (23) we obtain the bound, for any $\epsilon > 0$,

$$\sum_{j=0}^{\lfloor Nm^+ \rfloor} \pi(j) q_j \leq q_{\lfloor Nm^+ \rfloor}$$

where $m^+ = m^* + \epsilon$. As in the proof of Proposition 3.3 we can apply Proposition A.6 to argue that the sum from $\lfloor Nm^+ \rfloor$ to infinity is negligible, giving

$$\limsup_{N\to\infty} \frac{1}{N} \log \left(\sum_{j=0}^{\infty} \pi(j) q_j \right) \leq -m^+ \bar{I}_{\Gamma^\circ}(C/m^+))$$

On letting $\epsilon \downarrow 0$ this bound implies (49) since the rate function $\bar{I}_{\Gamma^\circ}$ is convex, which implies that $m\bar{I}_{\Gamma^\circ}(C/m))$ is continuous in a neighborhood of m^*.

The case $m^* > \rho$ is treated similarly. Define a partition of the interval $[0, m^*]$, with $m_0 = 0$, $m_p = m^*$ and $m_l = \rho$ for some l, such that $|L(m_{k+1}) - L(m_k)| \leq \epsilon, k = 0, \ldots, p-1$. This is possible because the restrictions of L to $[0, \rho]$ and $[\rho, m^*]$ are continuous and monotone functions. Then for $k \leq l$ and $j \in \{\lfloor Nm_k \rfloor, \ldots, \lfloor Nm_{k+1} \rfloor\}$, we have $\pi(j) \leq \pi(\lfloor Nm_{k+1} \rfloor) \mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{\lfloor Nm_{k+1} \rfloor / N}\}^{-\lfloor Nm_{k+1} \rfloor}$. We thus have, for $k \leq l$,

$$\sum_{j=\lfloor Nm_k \rfloor}^{\lfloor Nm_{k+1} \rfloor} \pi_j q_j \leq (N(m_{k+1} - m_k) + 1) \pi_{\lfloor Nm_{k+1} \rfloor} q_{\lfloor Nm_{k+1} \rfloor} \mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{\lfloor Nm_{k+1} \rfloor / N}\}^{-\lfloor Nm_{k+1} \rfloor}.$$

Similarly for $k > l$ and $j \in [\lfloor Nm_k \rfloor, \lfloor Nm_{k+1} \rfloor]$, $\pi_j \leq \pi_{\lfloor Nm_k \rfloor} \mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{\lfloor Nm_k \rfloor/N}\}^{\lfloor Nm_k \rfloor}$ and $q_j \leq q_{\lfloor Nm_{k+1} \rfloor}$. Thus for $k > l$,

$$\sum_{j=\lfloor Nm_k \rfloor}^{\lfloor Nm_{k+1} \rfloor} \pi_j q_j \leq (N(m_{k+1} - m_k) + 1)\pi_{\lfloor Nm_k \rfloor} q_{\lfloor Nm_{k+1} \rfloor} \mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{\lfloor Nm_k \rfloor/N}\}^{\lfloor Nm_k \rfloor}.$$

We thus obtain the bound,

$$\begin{aligned} \sum_{j=0}^{\lfloor Nm^* \rfloor} \pi(j) q_j \quad &\leq \quad \sum_{k \leq l} (N(m_{k+1} - m_k) + 1)\pi_{\lfloor Nm_{k+1} \rfloor} q_{\lfloor Nm_{k+1} \rfloor} \mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{\lfloor Nm_{k+1} \rfloor/N}\}^{\lfloor Nm_{k+1} \rfloor} \\ &+ \quad \sum_{k > l} (N(m_{k+1} - m_k) + 1)\pi_{\lfloor Nm_k \rfloor} q_{\lfloor Nm_{k+1} \rfloor} \mathsf{P}\{\widehat{\xi}(t) \in \Delta^*_{\lfloor Nm_k \rfloor/N}\}^{\lfloor Nm_k \rfloor} \qquad (50) \end{aligned}$$

Using Theorem 3.2, Proposition 3.3 and the fact that $|L(m_{k+1}) - L(m_k)| \leq \epsilon$, we have,

$$\begin{aligned} \limsup_{N \to \infty} \frac{1}{N} \log \left(\sum_{j=0}^{\lfloor Nm^* \rfloor} \pi(j) q_j \right) &\leq - \min_k \{L(m_k) + m_k \bar{I}_{\Gamma^\circ}(C/m_k)\} + \epsilon \\ &\leq - \inf_{m > 0} \{L(m) + m \bar{I}_{\Gamma^\circ}(C/m)\} + \epsilon \end{aligned}$$

and from Proposition A.6 (iii) we can again conclude that

$$\limsup_{N \to \infty} \frac{1}{N} \log \left(\sum_{\lfloor Nm^* \rfloor}^{\infty} \pi(j) q_j \right) = 0.$$

which establishes the asymptotics of $\eta^*(N)$ in the case $m^* > \rho$. □

The proofs for Proposition 4.1, Proposition 4.2 and Proposition 4.3 are identical to those for Proposition 3.1, Theorem 3.2 and Proposition 3.3 respectively. The proof of Theorem 4.4 is also similar to that of Theorem 3.4, with some minor differences that are described below.

Proof of Theorem 4.4. Although the asymptotic expression for $\eta^*(N)$ in Theorem 4.4 is different from the corresponding one in Theorem 3.4, the proof of Theorem 3.4 carries over almost exactly to Theorem 4.4. This is because the proof of the asymptotics of η_M depends on the following: (i) The structural properties of the function L, (ii) The fact that q_j is monotone increasing in j, and (iii) The fact that $m\bar{I}_{\Gamma^\circ}(C/m)$ is lower semi-continuous for $m > m^*$. The function L also appears in Theorem 4.4, so that the stuctural properties of L can be used here as well. Also the monotonicity of q_j is easy to see in the buffered case as well.

In the buffered model, the function $m\bar{I}_{\Gamma^\circ}(C/m)$ is replaced by $\inf_{t \geq 0} m\bar{I}_{\Gamma,t}((Ct + B)/m)$. It remains to prove that this last function is lower semi-continuous for $m > m^*$. Firstly, from Assumption A3, this can be written as $\inf_{0 \leq t \leq T_{\max}} m\bar{I}_{\Gamma,t}((Ct+B)/m)$. Now from Assumption A2 we can conclude that $(C+B)/m^* < R$, i.e., $m^* > R^{-1}(C+B)$. Consequently, $m^* > (Rt)^{-1}(Ct+B)$ for any $t \geq 1$. Therefore we must have $m\bar{I}_{\Gamma,t}((Ct+B)/m) = mI_{\Gamma,t}((Ct+B)/m)$ for all $t \geq 1$ and for $m > m^*$. We therefore have, for $m > m^*$,

$$\inf_{t \geq 0} m\bar{I}_{\Gamma,t}((Ct + B)/m) = \inf_{0 \leq t \leq T_{\max}} mI_{\Gamma,t}((Ct + B)/m)$$

Finally, since $I_{\Gamma,t}$ is a lower semi-continuous function of m for any t, and the minimum of a finite number of lower semi-continuous functions is also lower semi-continuous, we conclude that $\inf_{t\geq 0} m\bar{I}_{\Gamma,t}((Ct+B)/m)$ is lower semi-continuous on $[m^*,\infty)$.

In the proof of Theorem 3.4, the fact that the infimum is achieved at m^* when $m^* \leq \rho$ relies on the monotone decreasing nature of $m\bar{I}_{\Gamma^\circ}(C/m)$. Here, since $m\bar{I}_{\Gamma,t}((Ct+B)/m)$ is monotone decreasing in m for each t, $\inf_t m\bar{I}_{\Gamma,t}((Ct+B)/m)$ is also monotone decreasing in m. Thus the proof of Theorem 3.4 carries over to this case as well. □

References

[1] G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57:33–45, 1962.

[2] D. Bertsimas and J. Sethuraman. Moment problems and semidefinite optimization. In *Handbook of semidefinite programming*, volume 27 of *Internat. Ser. Oper. Res. Management Sci.*, pages 469–509. Kluwer Acad. Publ., Boston, MA, 2000.

[3] L. Vandenberghe, S. Boyd, and K. Comanor. Generalized Chebyshev bounds via semidefinite programming. Submitted to SIAM Review, Problems and Techniques Section, January 2004.

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

[5] F. Brichet and A. Simonian. Conservative Gaussian models applied to measurement-based admission control. In *Proceedings of IWQoS*, Napa, CA, May 1998.

[6] M. Chen, I.-K. Cho, and S.P. Meyn. Reliability by design in a distributed power transmission network. To appear, *Automatica*, 2005. (invited), 2004.

[7] H. Chen and D.D. Yao. *Fundamentals of Queueing Networks: Performance, Asymptotics and Optimization*. Springer-Verlag, New York, 2001.

[8] C. Courcoubetis, G. Kesidis, A. Ridder, J. Walrand, and R. R. Weber. Admission control and routing in ATM networks using inferences from measured buffer occupancy. *IEEE Transactions on Communications*, 43:1778–1784, April 1995.

[9] C. Courcoubetis and R. Weber. Buffer overflow asymptotics for a buffer handling many traffic sources. *J. Appl. Prob.*, 33:886–903, 1996.

[10] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer, New York, 2nd edition, 2000.

[11] N. G. Duffield. A large deviation analysis of errors in measurement based admission control to buffered and bufferless resources. *Queueing Syst. Theory Appl.*, 34(1-4):131–168, 2000.

[12] P. W. Glynn and D. Ormoneit. Hoeffding's inequality for uniformly ergodic Markov chains. *Statistics and Probability Letters*, 56:143–146, 2002.

[13] I. Kontoyiannis, L. A. Lastras-Montaño, and S.P. Meyn. Relative entropy and exponential deviation bounds for general Markov chains. In *Proceedings of the International Symposium on Information Theory (ISIT), 2005*, June 2005. Submitted for publication.

[14] D.Y. Eun and N.B. Shroff. A measurement-analytic approach for QoS estimation in a network based on the dominant time scale. *IEEE/ACM Transactions on Networking*, 11(2):222–235, April 2003.

[15] S. Floyd. Comments on measurement-based admission control for controlled-load services. Technical report, Lawrence Berkeley National Laboratory, July 1996.

[16] R. Gibbens and F. Kelly. Measurement-based connection admission control. 15th International Teletraffic Congress, 1997.

[17] R. Gibbens, F. Kelly, and P. Key. A decision-theoretic approach to call admission control in ATM networks. *IEEE JSAC*, 13(6):1101–1114, August 1995.

[18] P.W. Glynn and D. Ormoneit. Hoeffding's inequality for uniformly ergodic markov chains. *Statist. Probab. Lett.*, 56(2):143–146, 2002.

[19] J. W. Roberts. A survey on statistical bandwidth sharing. *Comput. Networks*, 45(3):319–332, 2004.

[20] M. Grossglauser and D. N. C. Tse. A time-scale decomposition approach to measurement-based admission control. *IEEE/ACM Trans. Netw.*, 11(4):550–563, 2003.

[21] M. Grossglauser, S. Keshav, and D.N.C. Tse. RCBR: a simple and efficient service for multiple time-scale traffic. *IEEE/ACM Transactions on Networking*, 5(6):741–755, 1997.

[22] M. Grossglauser and D. Tse. A framework for robust measurement-based admission control. *IEEE/ACM Transactions on Networking*, 7(3):293–309, June 1999.

[23] F. Guillemin and R. Mazumdar. Extremal traffic and bounds on the loss probability in buffers fed with regulated traffic. In *Proceedings of the Allerton Conference*, October 2001.

[24] S.G. Henderson. *Variance Reduction Via an Approximating Markov Process*. PhD thesis, Stanford University, Stanford, California, USA, 1997.

[25] S.G. Henderson, S. P. Meyn, and V. Tadic. Performance evaluation and policy selection in multiclass networks. *DEDS*, 13:149–189, 2003. Special issue on learning and optimization methods (invited).

[26] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[27] J. Huang, C. Pandit, S. Meyn, M. Medard, and V. Veeravalli. Entropy, inference, and channel coding. In Prathima Agrawal, Matthew Andrews, Philip J. Fleming, George Yin, and Lisa Zhang, editors, *Proceedings of the Summer Workshop on Wireless Networks (To appear.)*, IMA volumes in Mathematics and its Applications, New York, 2005. Springer-Verlag.

[28] J. Huang and S. P. Meyn. Characterization and computation of optimal distribution for channel coding. *IEEE Trans. Inform. Theory*, 51(7):1–16, 2005.

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

[30] F. P. Kelly and R. J. Williams. Fluid model for a network operating under a fair bandwidth-sharing policy. *Ann. Appl. Probab.*, 14(3):1055–1083, 2004.

[31] G. Kesidis and T. Konstantopoulos. Extremal shape-controlled traffic patterns in high-speed networks. *IEEE Transactions on Communications*, 48(5):813–819, 2000.

[32] M. G. Kreĭn. The ideas of P. L. Čebyšev and A. A. Markov in the theory of limiting values of integrals and their future developments. *Translations of the American Mathematical Society*, 12:1–121, 1959.

[33] C. Pandit. *Robust Statistical Modeling Based On Moment Classes With Applications to Admission Control, Large Deviations and Hypothesis Testing.* PhD thesis, University of Illinois at Urbana Champaign, University of Illinois, Urbana, IL, USA, 2004.

[34] C. Pandit and S. P. Meyn. Worst-case large-deviations with application to queueing and information theory. To appear, *Stoch. Proc. Applns.*, 2005.

[35] J. Qiu and E.W. Knightly. Measurement-based admission control with aggregate traffic envelopes. *IEEE/ACM Transactions on Networking*, 9(2):199–210, 2001.

[36] M. Reisslein. Measurement-Based Admission Control for Bufferless Multiplexers. *Int. Journal of Communication Systems*, 14(8): 735-761, 2001.

[37] J. Roberts and L. Massoulié. Bandwidth sharing and admission control for elastic traffic. *ITC Specialists Seminar*, 1998.

[38] H. Saito and K. Shiomoto. Dynamic call admission control in ATM networks. *IEEE JSAC*, 9(7):982–989, September 1991.

[39] S. Shakkottai, R. Srikant, N. Brownlee, A. Broido and K. C. Claffy. The RTT distribution of TCP flows on the Internet and its impact on TCP based flow control, *CAIDA Tech Report number tr-2004-02*, January 2004