A Framework for Robust Measurement-Based Admission Control

Matthias Grossglauser *

INRIA
BP 93
06902 Sophia Antipolis Cedex
France
Matthias.Grossglauser@inria.fr

David Tse†

Dept. of Electrical Engineering and Computer Sciences
University of California
Berkeley, CA 94720
USA
dtse@eecs.berkeley.edu

Abstract

Measurement-based Admission Control (MBAC) is an attractive mechanism to concurrently offer Quality of Service (QoS) to users, without requiring a-priori traffic specification and on-line policing. However, several aspects of such a system need to be clearly understood in order to devise robust MBAC schemes. Through a sequence of increasingly sophisticated stochastic models, we study the impact of parameter estimation errors, of flow arrival and departure dynamics, and of estimation memory on the performance of an MBAC system.

We show that a *certainty equivalence* assumption, i.e., assuming that the measured parameters are the real ones, can grossly compromise the target performance of the system. We quantify the improvement in performance as a function of the memory size of the estimator and a more conservative choice of the certainty-equivalent parameters. Our results yield valuable new insight into the performance of MBAC schemes, and represent quantitative guidelines for the design of robust schemes.

1 Introduction

Integrated-services networks are expected to carry a class of traffic that requires Quality of Service (QoS) guarantees. One of the main challenges consists in providing QoS to users while efficiently sharing network resources through statistical multiplexing. The role of Admission Control (AC) is to limit the number of flows admitted into the network such that each individual flow obtains the desired QoS.

Traditional approaches to admission control require an *a priori* traffic specification in terms of the parameters of a deterministic or stochastic model. The admission decision is then based on the specifications of the existing and the new flow. This approach suffers from several drawbacks. First, it is usually difficult for the user to tightly characterize his traffic in advance [11]. This is true even for stored media such as video-on-demand, as the user is expected to be able to exercise interactive control (such as pause, fast-forward etc.) As a result, traffic specifications can be expected to be quite loose. Second, there exists a modeling tradeoff between the ability to police and the statistical multiplexing gain. Deterministic models such as leaky buckets are easy to police, as they specify the *worst-case* behavior of traffic on a single time-scale, but they fail to provide a sufficient characterization to extract a large fraction of the potential statistical multiplexing gain. While a sequence of leaky buckets can approach such a multiple time-scale characterization, the number of model parameters grows accordingly [10].

Stochastic models such as those based on effective bandwidth [8] are better suited to achieve good statistical multiplexing gain. However, they suffer from two problems. First, it is difficult for the user to come up with the model parameters *a priori*. If he overestimates his requirements, then resources will be wasted in the network. This reduces the network utilization. If he underestimates his requirements, then insufficient resources will be allocated to his flow. The user has to abort the flow or try to adapt to this situation, for example by increasing the degree of compression of a video flow, thereby lowering its perceived quality. Second, it is hard to police traffic according to a statistical model [8]. It is not clear how to ensure that a traffic flow correspond to the specified parameters, without which admission control can easily be "fooled".

Measurement-based Admission Control (MBAC) avoids this problem by shifting the task of traffic specification from the user to the network. Instead of the user explicitly specifying his traffic, the network attempts to "learn" the statistics of existing flows by making on-line measurements. This approach has several important advantages. First, the user-specified traffic descriptor can be trivially simple (e.g. peak rate). Second, an overly conservative specification does not result in an overallocation of resources for the entire duration of the session. Third, when traffic from different flows are multiplexed, the QoS experienced depends often on their *aggregate* behavior, the statistics of which are easier to estimate than those of an individual flow. This is a consequence of the law of the large numbers. It is thus easier to predict aggregate behavior rather than the behavior of an individual flow.

Relying on measured quantities for admission control raises a number of issues that have to be understood in order to develop robust schemes.

- **Estimation error**. There is the possibility of making errors associated with any estimation procedure. In the context of MBAC, the estimation errors can translate into erroneous flow admission decisions. The effect of these decision errors has to be carefully studied, because they add another level of uncertainty to the system, the first level being the stochastic nature of the traffic itself. Assuming *certainty equivalence* upfront, i.e. assuming that the estimated parameters are the real parameters, is dangerous, as we simply ignore its impact on the quality of service. Actually, as we

*This author has been supported in part by a grant from France Telecom/CNET.

†This author has been supported by grant F49620-96-1-0199 from AFOSR, a grant from Pacific Bell and a MICRO grant from the government of California.

will show, estimation error tends to compromise the QoS delivered to users, as there exists a fundamental asymmetry associated with the uncertainty of the flow parameters: the negative effect on QoS of an underestimation of flow parameters - and therefore of an overestimation of the number of permissible flows - far exceeds the positive effect on QoS of an overestimation of flow parameters. Thus, on average, measurement uncertainty works against us, and we shall quantify how they can be compensated for in the MBAC.

- **Dynamics and separation of time-scale**. A MBAC is a dynamical system, with flow arrivals and departures, and parameter estimates that vary with time. It is tempting to make an assumption of burst and flow time-scale separation to conceptually decouple estimation, which measures the in-flow burst statistics, and flow dynamics, which occur on the flow time-scale. However, as we will show, this time-scale separation not only depends on the absolute burst and flow time-scale, but also on the size of the system as well as the flow arrival rate. Thus, the question to the impact of flow arrivals and departures on QoS arises. Intuitively, each flow arrival carries the potential of making a wrong decision. We therefore expect a high flow arrival rate to have a negative effect on performance. On the other hand, the impact of a wrong flow admission decision on performance also depends on how long it takes until this error can be corrected - that is, on flow departure dynamics.

- **Memory**. The quality of the estimators can be improved by using more past information about the flows present in the system. However, memory in the estimation process adds another component to the dynamics of a MBAC. For example, it introduces more correlation between successive flow admission decisions. Moreover, using too much memory will reduce the adaptability of MBAC to non-stationarities in the statistics. A key issue is therefore to determine an appropriate amount of memory to use. For this, a clear understanding of the impact of memory on both estimation errors and flow dynamics is necessary.

The goal of this work is to study the above issues - the impact of estimation error, of flow arrival and departure dynamics, and of measurement memory - in a unified framework. We wish to gain an understanding about how these aspects of a MBAC system interact. To do so, we consider a sequence of increasingly sophisticated models, adding one of the above issues at a time. This sequence culminates in the *continuous load model*, which allows us to derive analytical approximations, as well as an intuitive understanding, about how the above issues fit together. The ultimate goal is to shed insights on the design of robust MBAC schemes which can provide the appropriate QoS to the user even in the presence of the additional uncertainty due to measurements.

It should be stressed that the goal of this paper is not to propose a complete admission control scheme. Rather, we focus only on the above issues, by keeping the models as simple as possible. This has led us to make several assumptions that are clearly not realistic, but not of relevance to the issues we are interested in. These assumptions include stochastic homogeneity across flows, stationarity, and the absence of long-range dependence (LRD) in the flows. Relaxing these assumptions opens up promising directions for future research.

The rest of the paper is organized as follows. In Section 2, we describe the models that will be studied. The analysis of these models is explained in Section 3 and 4. In Sections 5 and 6, we summarize the insights obtained, report some initial simulation results, and discuss how our results relate to previous work in measurement-based admission control. We conclude the paper in Section 7.

2 Models

We begin by briefly describing the basic model. The network resource considered is a bufferless single link with capacity c. Flows arrive over time and, if admitted, stay for a random time. The bandwidth requirements of a flow fluctuate over time while in the system. We assume that the statistics of the bandwidth fluctuations of each flow are identical, stationary and independent of each other, with a mean bandwidth requirement of μ and variance σ^2. An important system parameter is the normalized capacity $n := \frac{c}{\mu}$, which measures the system size in terms of the mean bandwidth of the flows. Resource overload occurs when the instantaneous aggregate bandwidth demand exceeds the link capacity, and the quality of service is measured by the steady-state overflow probability p_f.

To study the various issues outlined in the introduction, we will analyze three variations of this basic model of increasing complexity. In the first variation, an infinite burst of flows arrives at time 0 and admission control decisions are made then, based on the initial bandwidths of the flows. After time 0, no more flows will be accepted and moreover the flows already admitted will stay in the system forever. We call this the *impulsive load* model with *infinite flow holding time*. This model permits us to study the impact of the measurement errors on the number of admitted flows and on the overflow probability, without the need to worry about flow dynamics.

In the second variation, we consider a similar model with flows admitted only at time 0, but now the admitted flows have holding times exponentially distributed with mean T_h. Thus, they will gradually depart from the system. We call this the impulsive load model with *finite* flow holding time. This model allows us to study the impact of flow departures on the overflow probability.

The last variation is the *continuous load* model, where the full flow arrival and departure dynamics are considered. In this model, flows arrive continuously over time with effectively *infinite* arrival rate, i.e. there are always flows waiting to be admitted into the network. Once they are admitted, they stay for an exponentially distributed holding time with mean T_h. The motivation for this model is that a well-designed robust MBAC should work well even for very high flow arrival rates, to cater for times when there is a surge in user demand of the service. Thus, the continuous-load model provides the most stringent test for MBACs.

Several comments about the model are in order. First, we observe that the traffic model is a stationary one. In practice, one of the main reasons for using a measurement-based scheme is to adapt to non-stationarities in the statistics of the traffic, either due to the change in the nature of the flows or change in the statistics within a flow itself. The approach taken in this paper is to use a stationary model to evaluate the performance of schemes with *limited memory*. Thus, the results are valid if the traffic statistics are stationary within the memory time-scale. We view this as a first step towards a full understanding of adaptivity issues.

Second, we consider a resource model without buffers. There are several motivations for this. First, the dynamics leading to the overflow event in a bufferless system is much simpler than that of overflowing in a buffered system, as the event occurs whenever the instantaneous aggregate traffic load exceeds the link capacity. This simplification allows us to focus on the measurement problem that is of central interest in this paper. Second, our recent work on multiple time-scale traffic [6] such as compressed VBR video has indicated that a significant bulk of the statistical multiplexing gain can be obtained by a Renegotiated Constant Bit Rate

(RCBR) service. In this service model, buffering only occurs at the network edge, while sources renegotiate CBR rates from the network over the duration of a flow. Thus, the rates of the users fluctuate over time. Bandwidth renegotiations fail when the current aggregate bandwidth demand exceeds the link capacity, and the renegotiation failure probability is the QoS measure of this service. Thus, our bufferless model is directly applicable to this problem. In any case, the performance of schemes for the bufferless model is a conservative upper bound to the case when there are buffers.

Before we begin the analysis of these models, a few words about the notations in this paper. We use capital letters to denote random variables. The Gaussian distribution will play a central role in our analysis; the probability density function of a zero mean, unit variance Gaussian random variable ($N(0,1)$) is denoted by

$$\phi(x) := \frac{1}{\sqrt{2\pi}} \exp(\frac{-x^2}{2}) \tag{1}$$

and the complementary cumulative distribution function denoted by

$$Q(x) := \int_x^\infty \phi(u)du. \tag{2}$$

3 Impulsive Load Models

3.1 Infinite Flow Holding Time

In this subsection, we study the impulsive load model with infinite flow holding time, when flows are admitted at time 0 and stay in the system forever. The goal here is twofold. First, we wish to illustrate the importance of the additional uncertainty due to *measurement* or *estimation error*, by comparing schemes with perfect knowledge and measurement-based schemes. Second, we wish to lay the groundwork for the more sophisticated models discussed in subsequent sections.

Suppose the stationary bandwidth distribution of each flow i has mean μ and variance σ^2. The number of admissible flows m^* is the largest integer m such that

$$\Pr\left\{\sum_{i=1}^{m} X_i(t) > c\right\} \leq p_q. \tag{3}$$

where $X_i(t)$ is the bandwidth of the ith flow at time t. (Recall that $c := n\mu$ is the total capacity of the link.) For large system size n, the number of admissible calls will be large, and by the Central Limit Theorem,

$$\frac{1}{\sigma\sqrt{m}}\left[\sum_{i=1}^{m} X_i(t) - m\mu\right] \sim N(0,1)$$

Thus, if the parameters μ and σ^2 are known *a priori*, then the number of flows m^* to accept should satisfy:

$$Q\left[\frac{n\mu - m^*\mu}{\sigma\sqrt{m^*}}\right] = p_q. \tag{4}$$

where $Q(\cdot)$ is the ccdf of a $N(0,1)$ Gaussian random variable as defined in eqn. (2). [1] Because the AC has perfect knowledge of the statistics, the actual steady state overflow probability

$$p_f := \Pr\left\{\sum_{i=1}^{m^*} X_i(t) > c\right\}$$

satisfies the QoS requirement. For reasonably large capacities, it follows from solving (4) that m^* is well approximated by:

$$m^* \approx n - \frac{\sigma\alpha_q}{\mu}\sqrt{n} \tag{5}$$

where $\alpha_q := Q^{-1}(p_q)$. Note that n is the number of flows that can be carried on the link if each has constant bandwidth μ. Thus, the term $\frac{\sigma\alpha_q}{\mu}\sqrt{n}$ in the above expression can be interpreted as the safety margin left to cater for the (known) burstiness of the traffic.

Now, consider the situation when a MBAC does not know μ and σ *a priori*, but relies on an estimation of these parameters from the initial bandwidth of the flows and use the estimates in a *certainty equivalent* fashion. More specifically, we assume there are an infinite number of flows waiting for admission at time 0 due to a burst of arrivals. Invoking again the central limit approximation for large systems, the number of flows M_0 the MBAC admits should satisfy:

$$Q\left[\frac{n\mu - M_0\hat{\mu}}{\hat{\sigma}\sqrt{M_0}}\right] = p_q, \tag{6}$$

where

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i(0) \qquad \text{and} \qquad \widehat{\sigma} = \left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i(0) - \widehat{\mu})^2\right]^{\frac{1}{2}}$$

The criterion (6) is the same as (4), but with the true mean μ and standard deviation σ replaced by the *estimated* mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$ respectively.[2] Note that the number of flows M_0 admitted under the MBAC is now random, depending on the random bandwidths of the flows at time 0. This is a consequence of the fact that the admission control decisions are made based on measurements rather than known parameters. Also, the scheme considered here is an example of a *memoryless* MBAC, since the admission control decisions are made based on the current bandwidths only.

We now want to approximate the average overflow probability

$$p_f := \Pr\left\{\sum_{i=1}^{M_0} X_i(t) > c\right\}$$

in steady state (i.e. for t large) and compare it to the target p_q. To do this, we first find an approximation for the distribution of M_0, the number of flows admitted.

For large capacities, by the law of large numbers, the estimated mean $\hat{\mu}$ will be close to the true mean μ, and the estimated variance $\hat{\sigma}^2$ will be close to the true variance σ^2. A more precise approximation of the deviation of these estimated quantities from the true values is given by the Central Limit Theorem:

$$\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} X_i(0) = \mu + \frac{1}{\sqrt{n}}\left\{\frac{1}{\sqrt{n}}\left[\sum_{i=1}^{n} X_i(0) - n\mu\right]\right\} \\ &\approx \mu + \frac{\sigma Y_0}{\sqrt{n}} \end{aligned} \tag{7}$$

for large n. Here, $Y_0 \sim N(0,1)$, and can be interpreted as the scaled aggregate bandwidth fluctuation at time 0 around the mean. Similarly, the estimated standard deviation can be written as:

$$\hat{\sigma} \approx \sigma + \frac{Z_0}{\sqrt{n}} \tag{8}$$

[1] Note that here, as in the sequel, we are ignoring the fact that m^* is an integer and therefore eqn. (4) cannot be satisfied exactly in general. In the regime of large capacities, however, the approximation is good and the discrepancy can be ignored.

[2] Observe here that the estimation is based on n flows. In a more precise model, the estimation should be based on M_0 flows, the number to be admitted. However, in a large system, M_0 will be close to n and the discrepancy in replacing M_0 by n in the estimators are small.

where Z_0 is Gaussian. These two approximations imply that the deviation of the estimates from the respective true quantities is of order $\frac{1}{\sqrt{n}}$. Now, as mentioned earlier, if the estimates were *exactly* equal to their true values, then the number of flows admitted M_0 will be precisely m^*. This suggests that we can approximate the distribution of M_0 by a *linearization* of the relationship (6) around a nominal operating point (m^*, μ, σ) (i.e. the operating point under perfect knowledge):

$$\frac{n\mu - (m^* + \Delta_M)(\mu + \frac{\sigma Y_0}{\sqrt{n}})}{(\sigma + \frac{Z_0}{\sqrt{n}})\sqrt{m^* + \Delta_M}} = \alpha_q$$

Expanding the left hand side, using eqn. (4) and neglecting terms $o(1)$ [3], we get[4]

$$\frac{\Delta_M}{\sqrt{n}} + \frac{\sigma}{\mu} Y_0 \approx 0$$

and hence for large n,

$$M_0 \approx m^* - \frac{\sigma}{\mu} Y_0 \sqrt{n}. \tag{9}$$

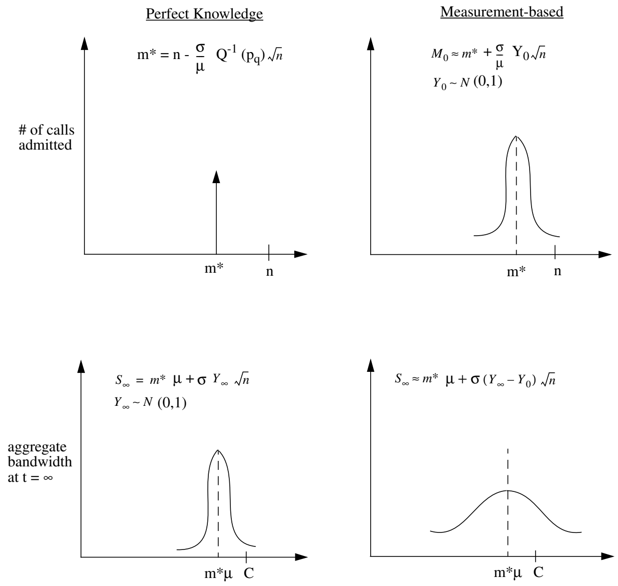

Figure 1: Uncertainty due to fluctuation in the number of flows (top) and in the aggregate bandwidth (bottom), for an admission controller with perfect knowledge (left) and for an MBAC (right).

Thus, we see that the effect of estimation error is an order $\sqrt{n}$ Gaussian fluctuation around m^*, the number of sources admitted under perfect knowledge (cf. top part of Fig 1). Note also that the randomness in the number of flows admitted is due mainly to the error in estimating the mean (Y_0) rather than the error in estimating the standard deviation (Z_0).

Substituting eqn. (5) into (9), we get M_0 in terms of the system size n:

$$M_0 \approx n - \frac{\sigma}{\mu}(Y_0 + \alpha_q)\sqrt{n} \tag{10}$$

[3]The notation $o(1)$ means terms going to zero as n goes to infinity

[4]Here and in the sequel, the notation "$\approx$" refers to an approximation with an error $o(\sqrt{n})$.

Although we have derived the result somewhat heuristically, it can be made precise by the following result, which is proved in the appendix.

Proposition 3.1 *For each system size n, let $M_0^{(n)}$ be the random number of flows admitted under the MBAC when the capacity is $n\mu$. Then the sequence of random variables $\{\frac{M_0^{(n)} - n}{\sqrt{n}}\}$ converges* in distribution *to the random variable $-\frac{\sigma}{\mu}(Y_0 + \alpha_q)$.*

We now proceed with an explicit approximation of the overflow probability. The randomness in the aggregate traffic load at some future time is due both to the randomness in the number of flows admitted as well as the randomness in the bandwidth demands of those flows. This can be approximated with the help of the following lemma, which is an extension of the Central Limit Theorem for a sum of a random number of random variables:

Lemma 3.2 *[1, p. 369, problem 27.14] Let $X_1, X_2, \dots$ be independent, identically distributed random variables with mean μ and variance σ^2, and for each positive n, let V_n be a random variable assuming positive integers as values; it need not be independent of the X_m's. Let $W_n = \sum_{i=1}^{V_n} X_i$. Suppose as $n \to \infty$, $\frac{V_n}{n}$ converges to 1 almost surely. Then as $n \to \infty$,*

$$\frac{W_n - V_n \mu}{\sigma\sqrt{n}}$$

converges in distribution to a $N(0,1)$ random variable.

Applying this lemma, the aggregate load at time t can approximated by:

$$S_t := \sum_{i=1}^{M_0} X_i(t) \approx M_0\mu + \sigma Y_t \sqrt{n} \tag{11}$$

Here $Y_t \sim N(0,1)$ and can be interpreted as an approximation for the scaled aggregate bandwidth fluctuation at time t:

$$\frac{1}{\sigma\sqrt{n}}\left[\sum_{i=1}^{n} X_i(t) - n\mu\right] \approx Y_t \tag{12}$$

Intuitively, eqn. (11) means that the fluctuation of the aggregate load is approximately the linear superposition of two effects: the random number of flows together with the random bandwidth fluctuation around the mean. Substituting eqn. (9), we get

$$S_t \approx n\mu + \sigma(Y_t - Y_0 - \alpha_q)\sqrt{n}$$

Thus, for large n, the overflow probability at time t is:

$$\Pr\{S_t > n\mu\} \approx \Pr\{Y_t - Y_0 > \alpha_q\}$$

This expression gives us an interpretation of how overflow occurs in a MBAC system: it is a combination of an aggregate bandwidth estimation error at admission time (Y_0) and a fluctuation of the aggregate bandwidth (Y_t) at time t after the flows have been accepted. Contrast this with the case with perfect knowledge, where the overflow probability at time t is simply $\Pr\{Y_t > \alpha_q\}$, due to bandwidth fluctuation at time t.

To get the overflow probability in steady state, we set $t = \infty$, in which case Y_∞ is independent of Y_0. Therefore, the difference $Y_\infty - Y_0$ is a Gaussian random variable with mean 0 and variance $2\sigma^2$. The overflow probability is therefore

$$p_f \approx Q\left(\frac{\alpha_q}{\sqrt{2}}\right). \tag{13}$$

We summarize this result more formally in the following proposition:

Proposition 3.3 *Suppose the target overflow probability QoS is p_q. Let $p_f^{(n)}$ be the actual average steady state overflow probability using the certainty equivalent MBAC for capacity $n\mu$. Then as the system size grows:*

$$\lim_{n\to\infty} p_f^{(n)} = Q\left(\frac{Q^{-1}(p_q)}{\sqrt{2}}\right)$$

Note that for the AC with perfect knowledge, the overflow probability is exactly p_q. This is because the aggregate bandwidth fluctuation stems only from the fluctuation of the individual flows' bandwidths (cf. lower left part of Fig. 1). On the other hand, in the measurement-based case, the variance of the aggregate bandwidth is doubled because the number of flows also fluctuates due to measurement error (cf. lower right part of Fig. 1). The $\sqrt{2}$ factor is therefore the effect of measurement error, and has quite a tremendous impact on the overflow probability p_f. For example, if $p_q = 1.0e-5$, then the actual performance in the MBAC system would be $p_f \approx 1.3e-3$, a difference of two orders of magnitude. In other words, if we want to achieve $p_f = p_q$ using a MBAC in this impulsive load model, then we have to adjust the target overflow probability under certainty equivalence.

$$p_{ce} = Q\left(\sqrt{2}\alpha_q\right) \qquad \text{or} \qquad \alpha_{ce} := Q^{-1}(p_{ce}) = \sqrt{2}\alpha_q. \quad (14)$$

Using the approximation $Q(x) \approx \frac{\phi(x)}{x}$ for small $Q(x)$, we see that

$$p_{ce} \approx \frac{\alpha_q}{2\sqrt{\pi}} p_q^2$$

Thus, we see that to achieve a target p_q in this setting, we need to set p_{ce} roughly to be the square of the target probability. This conservatism leads to a loss in system *utilization* compared to the scheme with perfect knowledge of the statistics. The average utilization (in terms of bandwidth) for the certainty equivalent scheme using parameter p_{ce} instead of p_q is given by $E(M_0)\mu$, or $c - \sigma\alpha_{ce}\sqrt{n}$, as implied by eqn. (9). The average utilization for the perfect knowledge scheme, on the other hand, is given by $m^*\mu$, or $c - \sigma\alpha_q\sqrt{n}$, as inferred from (5). Thus, if we pick α_{ce} to be $\sqrt{2}\alpha_q$, this translates to a loss of utilization of $(\sqrt{2}-1)\sigma\alpha_q\sqrt{n}$.

Proposition 3.3 has several surprising aspects. First, it is a *universal* result in the sense that the performance of the certainty equivalent scheme does not depend on the stationary distribution of the flow nor its mean and variance. Second, although the estimators are unbiased, the net impact on the performance of the system is negative. Thus there is an inherent asymmetry between the effects of over-estimation and under-estimation. Third, the impact of the estimation error does not vanish as the system size becomes large, even though the estimates become more and more accurate. Fourth, for a large system, the degradation in performance of the certainty equivalent scheme is due mainly to the estimation error in the *mean* μ of the bandwidth distribution and not to that in the *standard deviation* σ.

To get more insights into the last two phenomena, let us perform the following deterministic sensitivity analysis. Define the following function:

$$p_f(\mu, \sigma, m) := Q\left[\frac{c - m\mu}{\sigma\sqrt{m}}\right]$$

which is the overflow probability when there are m flows in the system each with mean rate μ and variance σ^2. Suppose first that we measure only μ, but that σ is known exactly. The number of flows admitted $m(\widehat{\mu})$ depends on the measured value $\widehat{\mu}$ and is given by the certainty-equivalent admission criterion (compare with (6)):

$$p_f(\widehat{\mu}, \sigma, m(\widehat{\mu})) = p_q. \quad (15)$$

Note that the *actual* overflow probability p_f for a given $m(\widehat{\mu})$ is $p_f(\mu, \sigma, m(\widehat{\mu}))$. The *sensitivity* of the overflow probability with respect to the measured $\widehat{\mu}$ is the deviation of p_f from its target value p_q if $\widehat{\mu}$ deviates slightly from its target value μ. For small deviations, we can simply use the derivative of p_f with respect to $\widehat{\mu}$.

$$s_\mu := \frac{\partial}{\partial\widehat{\mu}} p_f(\mu, \sigma, m(\widehat{\mu}))\bigg|_{\widehat{\mu}=\mu}.$$

Using (15), this derivative can be computed as:

$$s_\mu = -\frac{\phi(\alpha_q)\mu}{\sigma}\sqrt{m^*}.$$

Similarly, the sensitivity with respect to measured $\widehat{\sigma}$, assuming μ known, is given by:

$$s_\sigma = -\frac{\alpha_q\phi(\alpha_q)}{\sigma}$$

Now observe that the sensitivity of the system performance on the knowledge of the standard deviation, s_σ, does not depend on the system size. Therefore, increasing the system size, and therefore improving the quality of the estimator $\widehat{\sigma}$, results in a *diminishing* net impact on the overflow probability. On the other hand, the sensitivity s_μ *increases* with the system size, approximately as $\sqrt{n}$, while the variance of the estimator $\widehat{\mu}$ decreases approximately as $1/\sqrt{n}$. This suggests that the net impact of the uncertainty in the mean bandwidth estimate does not diminish as the system size grows, and also explains why the deviation from p_f from the target overflow probability p_q is asymptotically independent of n: both effects, less estimation error but increased sensitivity to estimation error, cancel out. The increased sensitivity to the mean estimate arises because when there are more flows in the system, and therefore more statistical regularity in the aggregate bandwidth, the system is driven closer to full utilization, which makes it more susceptible to admission mistakes.

The approximations used here are based in the *heavy traffic regime*, where the system size is large and when scaling up the size of the system, we exploit the additional statistical regularity by increasing the system utilization, while keeping the QoS constant. This is in contrast to the *large deviations regime*, where the system utilization is asymptotically constant, but where the QoS-requirement is scaled with the system size. The heavy traffic approximations allow us to linearize the dynamics of the system and to use Gaussian statistics. This will prove even more valuable as we analyze more complex models in the next sections. A large deviations analysis of a related measurement-based admission control problem can be found in [14].

3.2 Finite Holding Time

Now that we have convinced ourselves that estimation error can have an impact that should not be neglected, we want to refine the previous model. More specifically, we now assume that the time-scale separation is finite. There still is a burst of flows arriving at time 0 and demanding admission into the system. However, these flows are now assumed to have *finite duration*. In fact, we assume that the length of a flow (i.e., the time between the flow's admission and the time when it departs from the system) is an exponential random variable with mean T_h, and the lengths of different flows are assumed independent. We let p_t denote the probability that a flow has not departed from the system at time t. It is given by

$$p_t = \exp\left(-\frac{t}{T_h}\right). \quad (16)$$

Furthermore, we let $\rho(t)$ denote a flow's autocorrelation function.

If N_t is the number of flows left in the system at time t, and M_0 is the initial number of flows admitted into the system, then expected number of flows $E[N_t]$ at time t is $p_t E[M_0]$. Using eqn. (9), this implies that

$$E[N_t] \approx p_t n - \frac{p_t \sigma \alpha_q}{\mu}\sqrt{n}$$

We observe that the system size is n, and so approximately a fraction p_t of the total capacity is used at time t. The law of large number suggests that as n becomes large and everything else fixed, the overflow probability at time t actually goes to zero!

Intuitively, this can be explained as follows. When performing certainty-equivalent admission control, we set aside some bandwidth in order to accommodate fluctuations of the aggregate bandwidth. This spare bandwidth is on the order of $\sqrt{n}$ (cf. (5)). On the other hand, the flow departure rate is *proportional* to the number of flows in the system, approximately proportional to n/T_h. Now suppose that at some time instant, the system is close to overloading. How much time is necessary to restore the "safety margin" of $\sqrt{n}$ by letting flows depart? This restore time is on the order of $\sqrt{n}/(n/T_h) = T_h/\sqrt{n}$. Thus, the larger the system, the faster can the safety margin be restored. This means that to cause an overload, the aggregate bandwidth must fluctuate fast enough so that this fluctuation cannot be compensated for by just letting flows depart. However, as the time-scale gets shorter, the aggregate bandwidth tends to be more correlated, thus making such a quick change more and more unlikely.

While the above suggests that for large enough n, the overflow probability gets close to zero, it is clear that the longer the duration T_h of the flows, the larger the system size has to be for this effect to kick in. The above asymptotic analysis is crude in the sense that the flow duration, which may be quite long, does not enter the picture, since all other parameters are kept fixed while n grows large. On the other hand, it can be seen from the above discussion that the restore time $T_h/\sqrt{n}$ is the natural time-scale to analyze the dynamics due to flow departure. To make such analysis more convenient, let us rescale the flow holding time:

$$T_h = \widetilde{T_h}\sqrt{n}$$

where we view $\widetilde{T_h}$ fixed as n grows large. The advantage of this scaling is that it allows us to make approximations for large n but at the same time taking into consideration the actual duration of the flows. More specifically, it can be shown, under this scaling, the flow departure rate can be thought of as constant equal to $\sqrt{n}/\widetilde{T_h}$. Letting $D[0,t]$ be the number of flows departing in $[0,t]$, we have the approximation:

$$D[0,t] \approx \frac{t}{\widetilde{T_h}}\sqrt{n} \tag{17}$$

Using eqn. (9), the number of flows left in the system at time t can therefore be approximated as

$$N_t = M_0 - D[0,t] \approx n - \left[\frac{\sigma}{\mu}(Y_0 + \alpha_q) + \frac{t}{\widetilde{T_h}}\right]\sqrt{n} \tag{18}$$

Using Lemma 3.2, the aggregate load at time t can be approximated as:

$$\begin{aligned} S_t &= \sum_{i=1}^{N_t} X_i(t) \approx N_t\mu + \sigma Y_t\sqrt{n} \\ &\approx n\mu + \sigma\left(Y_t - Y_0 - \frac{\mu t}{\sigma\widetilde{T_h}} - \alpha_q\right)\sqrt{n} \end{aligned} \tag{19}$$

where Y_t is an approximation of the scaled fluctuation of the aggregate bandwidth

$$\frac{1}{\sigma\sqrt{n}}\left[\sum_{i=1}^{n} X_i(t) - n\mu\right].$$

By the Central Limit Theorem applied to pairs of random variables [1], Y_0 and Y_t are jointly Gaussian random variables with zero means, unit variances and covariance $\rho(t)$ (i.e. same as an individual flow). Thus, $Y_t - Y_0 \sim N[0, 2(1-\rho(t))]$.

The overflow probability $p_f(t)$ at time t is given by

$$\begin{aligned} p_f(t) &\approx \Pr\left\{Y_t - Y_0 > \frac{\mu t}{\sigma\widetilde{T_h}} + \alpha_q\right\} \\ &= Q\left(\frac{1}{\sqrt{2(1-\rho(t))}}\left[\frac{\mu}{\sigma}\frac{t}{\widetilde{T_h}} + \alpha_q\right]\right) \end{aligned} \tag{20}$$

From (20), we can see clearly the two effects affecting the overflow probability. For small t, the denominator $\sqrt{2(1-\rho(t))}$ is close to zero, making the overflow probability very small. This is because shortly after the admission decision, due to correlation in the bandwidth of the flows, the aggregate bandwidth does not change much. For large t, $t/\widetilde{T_h}$ makes the argument of the Q-function large as well, i.e. the overflow probability small. This is because enough flows have departed to make overflow unlikely. Intuitively, $\widetilde{T_h}$ defines the *critical time-scale* for this system: it is unlikely that an overflow event occurs at times significantly after $\widetilde{T_h}$. Thus, in the study of this system, we can concentrate on what happens between times of the order of $\widetilde{T_h}$. It is interesting that since $\widetilde{T_h} = T_h/\sqrt{n}$, this critical time-scale depends not only on the average holding time but also the size of the system.

4 The Continuous Load Model

We shall now consider a full-blown dynamical model, where flows arrive *continuously* over time. We assume a worst-case scenario, where the effective arrival rate is infinite, i.e. there are always flows waiting to be admitted into the network. Thus, admission control decisions are made continuously at all times. Clearly, the performance of any admission control algorithm under finite arrival rate will be no worse than its performance in this model. Another advantage of this model is that we need not worry about the specific flow arrival process which may be hard to model in practice. As before, when flows are admitted, they stay for a duration exponentially distributed with mean T_h. In this section, we will look at both memoryless MBAC schemes and schemes with memory and compare their performance.

4.1 Memoryless MBAC

We first look at the scheme that was considered in the impulsive load model, where admission control decisions are made based on estimates of the mean and variance using the current bandwidths of the flows. Assume that the system is in steady-state. Our goal is to find the overflow probability at an arbitrary time t. We do this by first analyzing the dynamics of the number of flows in the system.

Let M_t be the number of flows that the MBAC determines should be in the network at time t; as in (21), M_t is given by:

$$Q\left[\frac{n\mu - M_t\hat{\mu}(t)}{\hat{\sigma}(t)\sqrt{M_t}}\right] = p_q, \tag{21}$$

where

$$\widehat{\mu}(t) = \frac{1}{n}\sum_{i=1}^{n} X_i(t) \qquad \text{and} \qquad \widehat{\sigma}(t) = \left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i(t)-\widehat{\mu}(t))^2\right]^{\frac{1}{2}} \tag{22}$$

Observe that M_t is random and depends only on the current bandwidths $X_i(t)$'s of the flows. Call M_t the *estimated admissible* number of flows at time t. The *actual* number of flows N_t in the system at time t is no less than M_t since there are always flows waiting to be admitted and thus the system is always filled to the limit as currently determined by the MBAC. On the other hand, N_t can be strictly greater than M_t as flows that were admitted earlier stay for a certain duration and thus N_t cannot perfectly track the fluctuations of M_t (see Fig. (2)). To compute N_t, first observe that if s^* is the last time at or before time t that flows were admitted, then the number of flows in the system at time s^* is precisely the same as number of flows admissible at time s^*, i.e. $N_{s^*} = M_{s^*}$. In between time s^* and time t, no new flows were admitted. Hence, if we let $D[s,t]$ be the number of flows departed in time interval $[s,t]$, then

$$N_t = N_{s^*} - D[s^*,t] = M_{s^*} - D[s^*,t] \tag{23}$$

On the other hand, for *any* $s \leq t$,

$$N_t = N_s + A[s,t] - D[s,t] \geq N_s - D[s,t] \geq M_s - D[s,t] \tag{24}$$

where $A[s,t]$ is the number of flows *admitted* during $[s,t]$. Thus we conclude from (23) and (24) that

$$N_t = \max_{s\leq t}\{M_s - D[s,t]\} \tag{25}$$

Eqn. (18) in the previous section tells us that $M_s - D[s,t]$ is the number of flows in the system at time t if there were only a single impulse of flow arrivals at time s. Thus, the effect under a continuous load can be thought of as the worst-case over all impulsive arrival times.

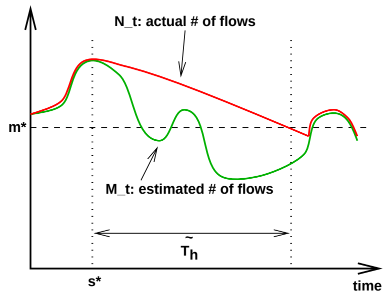

Figure 2: The relationship between the current estimate of admissible number of flows M_t and the actual number of flows N_t. The time-scale $\widetilde{T_h}$ is the typical time for the system to recover from admission errors.

Using formula (25), we can approximate N_t using our approximations for M_s and $D[s,t]$ as discussed in the previous section. Eqn. (9) gives an approximation for M_s:

$$M_s \approx n - \frac{\sigma}{\mu}(Y_s + \alpha_q)\sqrt{n} \tag{26}$$

where $\{Y_t\}$ is a stationary zero-mean Gaussian process with unit variance and auto-correlation function $\rho(t)$ (that of an individual flow), and can be interpreted as the scaled aggregate bandwidth fluctuation of the flows around the mean. Eqn. (17) gives an approximation for $D[s,t]$:

$$D[s,t] \approx \frac{(t-s)}{\widetilde{T_h}}\sqrt{n}$$

An approximation for the number of flows in the system is then given by:

$$\begin{aligned} N_t &\approx \max_{s\leq t}\left\{n - \left[\frac{\sigma}{\mu}(Y_s+\alpha_q) + \frac{(t-s)}{\widetilde{T_h}}\right]\sqrt{n}\right\} \\ &= n + \frac{\sigma\sqrt{n}}{\mu}\max_{s\leq t}\left\{-Y_s - \frac{\mu(t-s)}{\sigma\widetilde{T_h}} - \alpha_q\right\} \end{aligned} \tag{27}$$

In terms of N_t, the aggregate load at time t can be approximated as in (19):

$$\begin{aligned} S_t &\approx N_t\mu + \sigma Y_t\sqrt{n} \\ &\approx n\mu + \sigma\max_{s\leq t}\{Y_t - Y_s - \beta(t-s) - \alpha_q\}\sqrt{n} \end{aligned}$$

where we define for brevity

$$\beta := \frac{\mu}{\sigma\widetilde{T_h}}. \tag{28}$$

The steady-state overflow probability is therefore:

$$p_f = \Pr\{S_t > n\mu\} \approx \Pr\left\{\max_{s\leq t}\{Y_t - Y_s - \beta(t-s)\} > \alpha_q\right\} \tag{29}$$

Interestingly, one can interpret this probability as that of the length of a certain *queue* exceeding α_q. The queue is one which has a constant service rate of β, with the amount of work arriving in time interval $[s,t]$ given by $Y_t - Y_s$.

4.2 Analysis of Overflow Probability

Our next step is to analyze the approximation to the overflow probability given by eqn. (29). Since the process $\{Y_t\}$ is stationary and symmetrically distributed around 0, we can rewrite that as

$$p_f \approx \Pr\left\{\max_{t\geq 0}\{Y_{-t} - Y_0 - \beta t\} > \alpha_q\right\}.$$

This can be interpreted as the *hitting* probability of a Gaussian process $\{Y_{-\tau} - Y_0\}$ on a moving boundary $y = \beta\tau + \alpha_q$. While there is no known closed-form solution to this problem, an approximation can be derived by extending some of the results by [4] on hitting probabilities of *stationary* Gaussian processes to non-stationary ones. Define

$$\sigma^2(t) := E[(Y_{-t} - Y_0)^2] = 2[1-\rho(t)]$$

to be the variance of $Y_{-t} - Y_0$. (Recall that Y_t has zero mean and unit variance.) Assume the single-sided derivatives of $\rho(t)$ at $t=0$ exist and are finite, let $v^+(0)$ be the right derivative of the function $\sigma^2(t)$ at $t=0$. [5] Then an approximation to the hitting probability is given by:

$$\begin{aligned} &\Pr\left\{\max_{t\geq 0}\{Y_{-t} - Y_0 - \beta t\} > \alpha_q\right\} \approx \\ &\approx \frac{1}{2}\int_0^\infty v^+(0)\frac{\alpha_q+\beta t}{\sigma^3(t)}\phi\left(\frac{\alpha_q+\beta t}{\sigma(t)}\right)dt \end{aligned} \tag{30}$$

where $\phi(x)$ is the $N(0,1)$ probability density function. The integrand above can be viewed as an approximation to the first hitting time density at time t; integrating over all t yields the probability that hitting occurs at all. This is an approximation in the sense that as $\alpha_q \to \infty$, the ratio of the

[5] i.e. $v^+(0) := \lim_{t\to 0^+}\frac{\sigma^2(t)-\sigma^2(0)}{t}$.

left-hand and the right-hand sides approaches 1. Hence this approximation is good when p_q is small.

While this yields an approximation that can be computed numerically for general auto-correlation functions, we would like to get more analytical insights. To that end, consider the specific auto-correlation function:

$$\rho(t) = \exp(-\frac{|t|}{T_c}). \tag{31}$$

With this choice of the auto-correlation function, $\{Y_t\}$ is the well-known Ornstein-Uhlenbeck process. The parameter T_c governs the exponential drop-off rate of the correlation function, and is a natural *correlation time-scale* for the burst dynamics of the traffic. Substituting this into the approximation (30) and rescaling the time variable, we get:

$$p_f \approx \gamma \int_0^\infty \frac{(\alpha_q + t)}{[2(1-\exp(-\gamma t))]^{\frac{3}{2}}} \phi\left(\frac{\alpha_q + t}{\sqrt{2(1-\exp(-\gamma t))}}\right) dt \tag{32}$$

where

$$\gamma := \frac{1}{\beta T_c} = \frac{\widetilde{T_h}}{T_c} \cdot \frac{\sigma}{\mu}.$$

One can think of γ as the separation between the flow and burst scales, although note that $\widetilde{T_h}$ is the scaled holding time. If we make a time-scale separation assumption, i.e. $\gamma \gg 1$, then

$$p_f \approx \gamma \int_0^\infty \frac{(\alpha_q + t)}{2^{\frac{3}{2}}} \phi\left(\frac{\alpha_q + t}{\sqrt{2}}\right) dt = \frac{\gamma}{2\sqrt{\pi}} \exp\left(-\frac{1}{4}\alpha_q^2\right) \tag{33}$$

Note that the first approximation is via $\exp(-\gamma t) \approx 0$ for $\gamma \gg 1$.

It is interesting to compare this overflow probability for the continuous-load model with the corresponding result for the impulsive load model under long flow durations, given in Proposition (3.3). To do this, we first use the approximation $\frac{\phi(x)}{x} \approx Q(x)$ and rewrite (33) in terms of the flow parameters as

$$p_f \approx \frac{\widetilde{T_h}}{2T_c} \cdot \frac{\sigma \alpha_q}{\mu} Q\left(\frac{\alpha_q}{\sqrt{2}}\right) \tag{34}$$

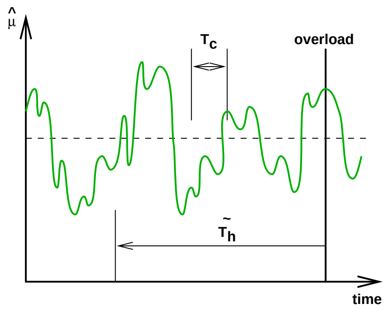

Figure 3: The ratio of correlation time-scale T_c and of the critical time-scale $\widetilde{T_h}$ determines the overflow probability.

For the impulsive load model, the overflow probability is approximately $Q(\frac{\alpha_q}{\sqrt{2}})$. Eqn. (34) tells us that in the regime of separation of time-scales, the corresponding overflow probability can be much worse in the continuous-load model. This is because while estimation errors can occur only at a *single* point of time in the impulsive load model (time 0), in the continuous-load model estimation errors occurring at *any* time in an interval of size roughly $\widetilde{T_h}$ before time t will have a significant impact on the number of flows at time t. The shorter the traffic correlation time-scale T_c, the faster the memoryless mean bandwidth estimates fluctuates, and the larger the probability of having an under-estimation at some time in the interval. Hence, the overflow probability in the continuous-load model increases with the separation of time-scale $\frac{\widetilde{T_h}}{T_c}$. For example, note the multiple peaks (underestimations of μ) within the interval of length $\widetilde{T_h}$ in Fig. 3: each of these peaks could potentially cause overload within the critical time-scale $\widetilde{T_h}$. The lesson is that it's not only important to consider the estimation error at a single time-instant, but also the chance of making error any time in the interval defined by the effective flow holding time-scale $\widetilde{T_h}$. Note also since $\widetilde{T_h}$ decreases as $\frac{T_h}{\sqrt{n}}$, where T_h is the actual mean holding time, the overflow probability decreases roughly as $\frac{1}{\sqrt{n}}$.

We can also write the above approximation as (using again $\frac{\phi(x)}{x} \approx Q(x)$),

$$p_f \approx \frac{\widetilde{T_h}}{\sqrt{2}T_c} \frac{\sigma}{\sqrt{2\pi}\mu} \left(\sqrt{2\pi}\alpha_q p_q\right)^{\frac{1}{2}} \tag{35}$$

4.3 MBAC with Memory

We see that the memoryless scheme suffers from two problems . First, the estimation error at a specific admission time instant is large, and in fact has impact which is of the same order of magnitude as that due to the statistical fluctuations of the bandwidths when the correct number of flows are admitted. Second, the correlation time-scale of the estimation errors is the same as that of the traffic itself; thus, in the regime when the flow holding time is much larger than the traffic correlation time-scale ($\widetilde{T_h} \gg T_c$), the probability of having a large under-estimation of mean bandwidth at *some time* during the time-scale $\widetilde{T_h}$ is high. A strategy which, as we will see, counters both these difficulties is to use more memory in the mean and variance estimators.

To be more concrete, let us consider using the first-order auto-regressive filter with impulse response:

$$h(t) := \frac{1}{T_m} \exp\left(-\frac{t}{T_m}\right) u(t)$$

to estimate both the mean and the variances. (Here, $u(t)$ is the unit step function.) Thus, in place of the memoryless estimators in eqn. (22), the MBAC would use:

$$\begin{aligned} \widehat{\mu}_m(t) &= \int_0^\infty \left[\frac{1}{n}\sum_{i=1}^n X_i(t-\tau)\right] h(\tau) d\tau \\ \widehat{\sigma}_m^2(t) &= \int_0^\infty \left[\frac{1}{n-1}\sum_{i=1}^n \left(X_i(t-\tau) - \widehat{\mu}_m(t)\right)^2\right] h(\tau) d\tau \end{aligned}$$

Note that the estimates are obtained by an exponential weighting of the past bandwidths of the flows. The parameter T_m governs how the past bandwidths are weighted; it can thought of as a measure of the *memory size* of the estimators. The relationship between $\widehat{\mu}_m(t)$ and the memoryless estimator $\widehat{\mu}(t)$ is simply $\widehat{\mu}_m = \widehat{\mu} * h$, where $*$ is the convolution operation.

Corresponding to eqn. (27) in the memoryless case, we can show that the number of flows N_t in the system at time t under the MBAC with memory is approximately

$$N_t \approx n + \frac{\sigma\sqrt{n}}{\mu} \max_{s \le t} \left\{-Z_s - \frac{\mu(t-s)}{\sigma \widetilde{T_h}} - \alpha_q\right\} \tag{36}$$

where $Z_t = (h * Y)_t$, and $\{Y_t\}$ is the scaled aggregate bandwidth fluctuation around the mean. One can interpret Z_t as the error in the *filtered* estimate of the mean bandwidth of a flow at time t. The overflow probability under the MBAC with memory can be approximated by:

$$p_f \approx \Pr\left\{\max_{t\geq 0}(Z_{-t} - Y_0 - \beta t) > \alpha_q\right\}$$

This is again a hitting probability of a Gaussian process ($\{Z_{-t} - Y_0\}$) on a moving boundary, and an approximation of such a probability is given by:

$$p_f \approx \frac{\gamma T_c}{T_c + T_m}\int_0^\infty \frac{(\alpha_q + t)}{[\sigma_m(t)]^3}\phi\left(\frac{\alpha_q + t}{\sigma_m(t)}\right)dt + Q\left(\alpha_q\sqrt{1+\frac{T_c}{T_m}}\right) \tag{37}$$

where

$$\sigma_m^2(t) := E[(Z_{-t} - Y_0)^2] = \frac{2T_c + T_m}{T_c + T_m} - \frac{2T_c}{T_c + T_m}\exp(-\gamma t)$$

Now, under separation of time-scales, $\gamma \gg 1$, we have the approximation that

$$\sigma_m^2(t) \approx \frac{2T_c + T_m}{T_c + T_m}$$

in which case the above integral can be explicitly computed as:

$$\begin{aligned} p_f \quad &\approx \quad \frac{\gamma T_c}{\sqrt{(T_c+T_m)(2T_c+T_m)}} \cdot \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{T_c+T_m}{2(2T_c+T_m)}\alpha_q^2\right) \\ &\quad + Q\left(\alpha_q\sqrt{1+\frac{T_c}{T_m}}\right) \end{aligned} \tag{38}$$

To compare this result to the memoryless case, let us first use the approximation $Q(x) \approx \frac{\phi(x)}{x}$ to rewrite (38) in terms of p_q and also the flow parameters:

$$\begin{aligned} p_f \quad &\approx \quad \frac{\widetilde{T_h}}{\sqrt{(T_c+T_m)(2T_c+T_m)}} \cdot \frac{\sigma}{\sqrt{2\pi}\mu}\left(\sqrt{2\pi}\alpha_q p_q\right)^{\frac{T_c+T_m}{2T_c+T_m}} \\ &\quad + Q\left(\alpha_q\sqrt{1+\frac{T_c}{T_m}}\right) \end{aligned} \tag{39}$$

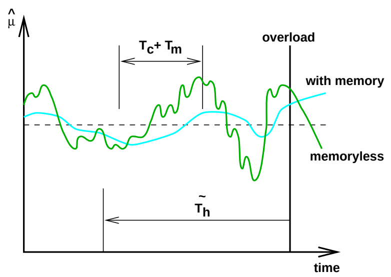

Figure 4: Estimation memory reduces the variance of the bandwidth estimator, and also smoothes its fluctuation.

Comparing eqn. (38) to eqn. (33), we can see explicitly the effect of memory. Let us look at the first term in (38), which corresponds to (35). The exponent is $\frac{T_c+T_m}{(2T_c+T_m)}$ which is $\frac{1}{2}$ when there is no memory (as we had in the memoryless scheme), monotonically increases with T_m, and reaching a value of 1 for infinite memory. This effect can be explained by the fact that the variance of the mean bandwidth estimate, $E[Z_t^2]$, is $\frac{T_c}{T_c+T_m}$ and decreases monotonically to zero with more memory. Thus the inaccuracy in the estimates and hence the inaccuracy in the number of flows accepted decreases (cf. Fig. 4). Furthermore, increasing the amount of memory has an additional effect of *smoothing* the mean bandwidth estimates; thus, not only are the *individual* bandwidth estimates more accurate, they also fluctuate less so that the probability of having an under-estimation at *some time* over an interval of length $\widetilde{T_h}$ is reduced. This is reflected in the smaller pre-factor $\frac{\widetilde{T_h}}{\sqrt{(T_c+T_m)(2T_c+T_m)}}$ in the first term of (39) replacing the factor $\frac{\widetilde{T_h}}{\sqrt{2}T_c}$ in the memoryless case. This can be interpreted as increasing the correlation time-scale by T_m, the estimator memory size.

In the limit for large T_m, we always have exactly the right number of flows in the system and the overflow occurs due only to the fluctuation of bandwidth requirements of flows in the system, and not to the fluctuation of the number of flows in the system. This is now given by the second term in (39).

Although formula (39) gives the overflow probability in terms of the memory size T_m of the estimator, it also depends on the traffic correlation time-scale T_c, which may be hard to estimate in practice as realistic auto-correlation functions are more complex than a pure exponential. Thus, it may not be easy to directly use formula (39) to determine what the appropriate amount of memory to use in the estimator. However, in the case that T_m is chosen large compared to T_c, formula (39) becomes

$$p_f \approx \left(\frac{\widetilde{T_h}}{T_m}\cdot\frac{\sigma\alpha_q}{\mu} + 1\right)p_q \tag{40}$$

which does not depend on T_c. In this regime, the effect of the estimator memory effectively masks the original correlation structure of the traffic. Although this result is derived using the simple exponential auto-correlation function (31), it can be expected that the detailed correlation structure is not relevant and a similar approximation holds for other auto-correlation functions.

Formula (40) can be used to choose the memory size and to adjust the certainty equivalent parameter p_{ce} in the MBAC such that the overflow probability meets the QoS requirement, i,e, choose T_m and p_{ce} such that:

$$\left(\frac{\widetilde{T_h}}{T_m}\cdot\frac{\sigma}{\mu}Q^{-1}(p_{ce}) + 1\right)p_{ce} = p_q$$

The shorter T_m, the more conservative the choice of p_{ce} has to be, resulting in a loss of utilization. This loss of utilization can be quantified. The average utilization (in terms of bandwidth) of the system is given by $\mu E[N_t]$, where N_t is the (stationary) number of flows in the system at time t. Eqn. (36) allows us to approximate this when p_{ce} is used as the certainty-equivalent parameter:

$$\mu E[N_t] \approx n\mu + \sigma\sqrt{n}E\left[\max_{s\leq t}\left\{-Z_s - \frac{\mu(t-s)}{\sigma\widetilde{T_h}}\right\}\right] - \sigma Q^{-1}(p_{ce})\sqrt{n}$$

Since the other terms do not depend on p_{ce}, we see that the difference in utilization in using p_{ce} and p'_{ce} is simply

$$\sigma\sqrt{n}\left[Q^{-1}(p_{ce}) - Q^{-1}(p'_{ce})\right] \tag{41}$$

This allows us to quantify the impact on the utilization on using a more conservative certainty-equivalent parameter.

5 Discussions and Simulations

Our framework yields several interesting qualitative insights about the measurement-based admission control issues we discussed in the introduction:

- Memoryless certainty-equivalent admission control can have very poor performance due to estimation error. The target QoS overflow probability can be missed by several orders of magnitude. The impact of the estimation errors does not diminish as the system gets larger.

- Estimation errors of different statistical parameters can have very different impact on the performance of an MBAC scheme. In the heavy traffic regime, the effect of error in estimating the mean is much more significant than the error in estimating the standard deviation.

- Flow departure dynamics have a significant impact on the performance of an MBAC scheme. The parameter $\widetilde{T_h} = T_h/\sqrt{n}$, where T_h is the average flow holding time and n the system size, defines a *critical time-scale* for which the effect of an admission error persists. This critical time-scale decreases with a shorter holding time or a bigger system because flows can leave the system more rapidly to repair a wrong decision.

- A high flow arrival rate has a detrimental effect on the performance of an MBAC scheme. A robust MBAC not only has to make sure that the estimation error for *each* decision is small, but also that the *worst* estimation error over the critical time-scale is small. Thus, a memoryless scheme which makes decisions based only on estimating *current* bandwidths is not robust; if the traffic correlation time-scale is short compared to the critical time scale $\widetilde{T_h}$, then the bandwidth estimates fluctuate too wildly.

- Increasing the amount of memory in the estimator helps in two ways. First, the individual bandwidth estimates are more accurate because of averaging over a larger number of samples. Second, it smoothes the bandwidth estimates so that they fluctuate less over time. This provides more control to the worst estimation error over the critical time-scale.

These insights are obtained from our analysis, which culminated in *explicit formulas* for evaluating the performance of MBAC schemes in terms of key parameters such as estimator memory size, traffic correlation time scale and average flow duration. Specifically, the main results are the general formula (37) for the overflow probability, the formula (39) specialized to the regime of separation of flow and burst time-scales, and the formula (40) with the further assumption that the memory size is much longer than the traffic correlation time-scale. Moreover, formula (41) yields the impact of a more conservative MBAC scheme on the utilization of the system, and, together with the previous formulas on overflow probability, quantifies the tradeoff between estimator memory size and the conservativeness of the MBAC for a given target QoS.

We now describe some simulations we have performed to validate and make concrete the above insights. We use RCBR (Renegotiated Constant Bit Rate [6]) traffic sources, i.e., the traffic rate produced by a source is constant over time intervals. Rate changes (renegotiations) are source-initiated and occur only on interval boundaries. We use independent homogeneous sources where the marginal rate distribution is Gaussian with $\sigma/\mu = 0.3$. The interval lengths are i.i.d. following an exponential law with mean T_c, which implies that the autocorrelation function of the traffic rate process is precisely as in (31).

We simulate the admission controller under infinite load and we measure the resulting overflow probability p_f. We terminate simulations when (a) the 95% confidence interval is less than +/- 20% of the estimated mean, or (b) the estimated mean plus the confidence interval is at least an order of magnitude below the target overflow probability ($1.0e-3$ in all our simulations). The latter criterion is to terminate simulations within a reasonable time for very small p_f.

We sample p_f at regular intervals of length $2\max(\widetilde{T_h}, T_m, T_c)$. This sample period is long enough to give approximately independent samples of the system, as the "memory" due to flow dynamics, estimation memory, and traffic correlation is taken into account. We also let the system initially warm up to steady state without collecting samples.

The first experiment (cf. Fig. 5) shows p_f for a certainty-equivalent MBAC as a function of memory size T_m, for different system sizes (both simulation results and the approximations via eqn. (38)). It is clear that using little or no memory, the performance of the system can be extremely bad. For example, for $n = 100$ and memoryless estimation, the target overflow probability is exceeded by more than two orders of magnitude! Using more estimation memory clearly improves the performance, but as the $n = 100$ case shows, is not necessarily sufficient for achieving the target QoS: even for $T_m = 100$, p_f is still larger than $1.0e-3$. Fig. 5 also shows that (38) is a slightly conservative approximation of the simulated p_f. Qualitatively, the correspondence is good; in particular, the "knee" in the curve is well matched.

The second experiment (cf. Fig. 6) shows the correction to be applied to p_{ce} in order to reach a robust QoS target of $p_q = 1.0e-3$, by inverting (38). It is another manifestation of the importance of the issues we discuss to achieve robust admission control: note that for small memory sizes, the *corrected* p_{ce} can be as low as $1.0e-10$ for a target of $1.0e-3$!

The third experiment (cf. Fig. 7) finally shows the simulated performance of the robust MBAC where the correction computed in the second experiment is applied. We see that the QoS target of $p_f < 1.0e-3$ is consistently met over the whole parameter range. The slight conservativeness of the approximation (38) carries through: p_f is everywhere between about 0.5 and 1.5 orders of magnitude below the target.

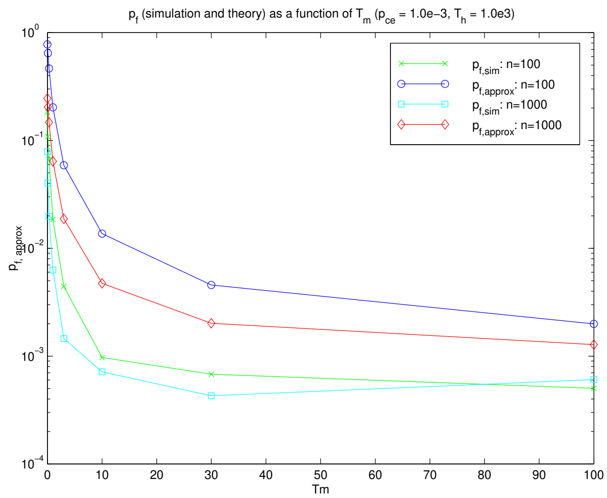

Figure 5: Non-robust MBAC: The simulated and theoretical overflow probability p_f for $p_{ce} = 1.0e-3$.

An important open question is the appropriate choice of the memory window size in practice. While our results provide quantitative insights into the role of estimator memory,

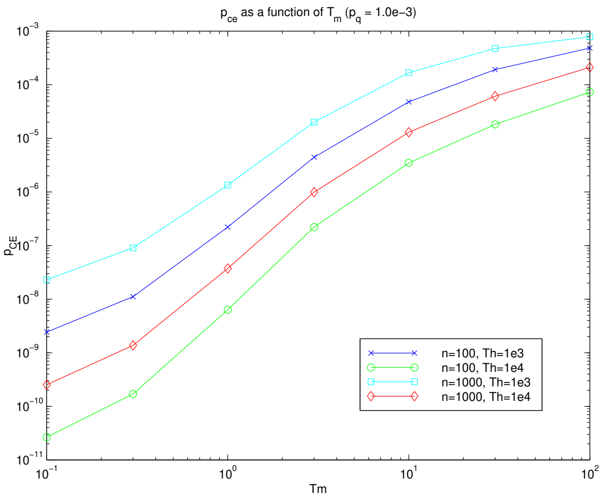

Figure 6: Robustness correction: The p_{ce} as a function of memory T_m to achieve a robust QoS target of $p_q = 1.0e - 3$.

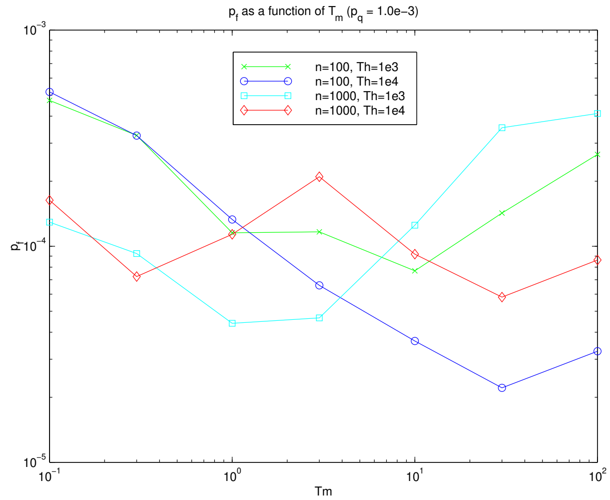

Figure 7: Robust MBAC: The simulated p_f *with* robustness correction, for QoS target $p_q = 1.0e - 3$.

they alone do not answer this question. This is because, as remarked earlier, real traffic is non-stationary but our model is a stationary one. A more comprehensive understanding of this problem requires a non-stationary model to study adaptivity issues and how the non-stationarity time-scale interacts with the various time-scales we studied in this paper. We plan to address this issue in future work.

6 Related Work

Past work on measurement-based admission control [3], [12], [7] have either ignored measurement errors or assumed a static situation where calls do not arrive or depart the system and there is arbitrarily long time to make accurate measurements. Here we discuss two more recent papers which are closer in spirit to our work.

Jamin *et al.*, in [9], presented a specific algorithm for measurement-based admission control of predictive traffic, and evaluated its performance through simulation. The algorithm relies on measurements of the maximum delay and maximum bandwidth over a measurement interval. There are several parameters in the algorithm (sampling window size S, measurement window size T, utilization target, back-off factor λ) that are found to have a significant impact on performance. However, clear guidelines on how to set these parameters are lacking. We believe that our work offers some insight into the impact of these system parameters. In particular, the measurement window size T is very similar to our measurement time-scale T_m. Also, λ is a parameter that controls an *overestimation* of the actual measured delay - in other words, it controls conservativeness, which in our work is represented through the parameter p_{ce}. Therefore, while the details of the models and metrics are not exactly identical, we think that our work helps understand the issues that govern the tuning of the above parameters. Our work has the further advantage that we use a much simpler service model so that we can focus on the issues associated with the measurement process.

Gibbens *et al.* [5] studied *memoryless* measurement-based admission control in a decision-theoretic framework. Their work takes into account the impact of measurement errors on performance and also considers the call dynamics. However, there are some significant differences between theirs and our work. First, a perfect time-scale separation is explicitly built into their model by assuming that the network states seen by successive call arrivals are independent. This makes it difficult to evaluate the performance of MBAC schemes with memory and also the effect of traffic correlation on a system with very high call arrival rates. Indeed they only focused on *memoryless* schemes. Moreover, our results show that the condition for time-scale separation is rather subtle, as it depends, among others parameters, on the system size. Second, while they also observed that a memoryless certainty equivalent scheme can perform poorly, their remedy is quite different. They relied on essentially two mechanisms: the use of a Bayesian prior on the call statistics and network state-independent call rejection. The first mechanism serves to smooth out the fluctuation in successive memoryless estimates, as the observations are weighted by a fixed prior. The second mechanism counters very high call arrival rates, by not accepting calls until one has left the system. In contrast, we propose the use of an appropriate amount of memory in the estimator , which as we have seen deals with both these problems. Our framework, without *a priori* assuming time-scale separation, allows us to evaluate the performance as a function of the amount of memory used. We believe the appropriate use of memory is a natural and effective strategy, particularly when no reliable prior exists.

7 Conclusions

Measurement-based Admission Control simplifies the contract between the user and the network, at the expense of having to deal with additional uncertainty in the system. The benefit of relieving the user of the burden of a-priori traffic specification, and of relieving the network of the burden of policing, far outweighs the costs of this uncertainty, if it can be prevented from compromising the quality of service experienced by the user. This problem has motivated the present work.

In this paper, we have presented a framework for studying the performance of admission control schemes under measurement uncertainty and flow dynamics. Using heavy-traffic approximations, the analysis of the resulting dynamical systems is simplified via linearization around a nominal operating point and by Gaussian approximations of the statistics via central limit theorems. We believe that the insight derived from our models, and the engineering guidelines on the choice of memory and certainty-equivalent target overflow probability, should be directly applicable in the design of robust MBAC schemes. In addition to the problem of memory window size choice we discussed earlier, there are also some additional issues that merit attention. For lack of space, we have to limit ourselves to a brief discussion.

First, there is increasing interest in *adaptive* applications, i.e., applications that are capable of functioning properly even if the QoS falls below the desired level [2]. This interest stems from the inability of the current Internet to guarantee any level of QoS. The QoS metric used here, i.e., the probability that a flow cannot get at least its target bandwidth at time t, is extreme in the sense that it does not account for the fact that getting part of that target bandwidth is still useful to an adaptive application. We are therefore working on a generalization of the QoS metric based on utility functions, inspired by Shenker's work [13]. The goal is to assess the impact of application adaptivity on the admission problem.

Second, we have assumed that *individual flows* are available for measurement. This might actually not be desirable or feasible in practice. Aggregate measurements can be expected to be easier to implement, because no per-flow information has to be maintained. While using only aggregate measurement does not affect the mean estimator, the accuracy of the variance estimator is hampered without per-flow information. We plan to study the effect on QoS of having only aggregate estimates available.

Third, we have assumed that the statistics of the flows are homogeneous. This is essentially a *worst-case* assumption, as it can be shown that if the flows were heterogeneous, the accuracy of the mean estimator remains the same but the variance estimate is only an upper bound to the true variance. Hence, the schemes presented here may be overly-conservative for heterogeneous traffic, and it is interesting to see how they can be improved for that case.

References

[1] P. Billingsley. *Probability and Measure (3rd Ed.)*. Wiley, 1995.

[2] D. Clark, S. Shenker, and L. Zhang. Supporting real-time applications in an integrated services packet network: Architecture and mechanism. In *Proc. ACM SIGCOMM '92*, pages 14–26, 1992.

[3] Costas Courcoubetis et al. Admission Control and Routing in ATM Networks using Inferences from Measured Buffer Occupancy. In *ORSA/TIMS special interest meeting*, Monterey, CA, January 1991.

[4] J. Cuzick. Boundary Crossing Probabilities for Stationary Gaussian Processes and Brownian Motion. *Transactions of the American Mathematical Society*, pages 469–492, February 1981.

[5] 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 of Communications*, pages 1101–1114, August 1995.

[6] M. Grossglauser, S. Keshav, and D. Tse. RCBR: A Simple and Efficient Service for Multiple Time-Scale Traffic. In *Proc. ACM SIGCOMM '95*, pages 219–230, Boston, Mass., August 1995.

[7] I. Hsu and J. Walrand. Dynamic Bandwidth Allocation for ATM Switches. *Journal of Applied Probability*, September 1996.

[8] J.Y. Hui. Resource allocation for broadband networks. *IEEE Journal on Selected Areas of Communications*, December 1988.

[9] S. Jamin, P. B. Danzig, S. Shenker, and L. Zhang. A Measurement-Based Admission Control Algorithm for Integrated Services Packet Networks. In *Proc. ACM SIGCOMM '95*, 1995.

[10] E. Knightly and H. Zhang. Traffic Characterization and Switch Utilization using a Deterministic Bounding Interval Dependent Traffic Model. In *Proc. IEEE INFOCOM '95*, Boston, Mass., April 1995.

[11] E. P. Rathgeb. Policing of Realistic VBR Video Traffic in an ATM Network. *International Journal of Digital and Analog Communications Systems*, 6:213–226, 1993.

[12] H. Saito and K. Shiomoto. Dynamic Call Admission Control in ATM Networks. *IEEE Journal on Selected Areas of Communications*, 9:982–989, 1991.

[13] S. Shenker. Fundamental Design Issues for the Future Internet. *IEEE Journal on Selected Areas of Communications*, 13(7), 1995.

[14] D. Tse and M. Grossglauser. Measurement-Based Call Admission Control: Analysis and Simulation. In *Proc. IEEE INFOCOM '97*, Kobe, Japan, April 1997.

8 Appendix

We use the notations $\xrightarrow{\mathcal{D}}$ and $\xrightarrow{a.s.}$ to denote convergence in distribution and almost sure convergence respectively. The following theorems are standard results in the theory of convergence in distribution.

Theorem 8.1 *(Continuous-Mapping Theorem) Let $\{\vec{Y}^{(n)}\}$ be a sequence of random vectors on $\Re^k$. If $h : \Re^k \to \Re$ is continuous and $\vec{Y}^{(n)} \xrightarrow{\mathcal{D}} \vec{Y}$, then $h(\vec{Y}^{(n)}) \xrightarrow{\mathcal{D}} h(\vec{Y})$.*

Theorem 8.2 *Let $\vec{Y}^{(n)}$'s and $\vec{Z}^{(n)}$'s be random vectors defined on the same probability space. If $\vec{Y}^{(n)} \xrightarrow{\mathcal{D}} \vec{Y}$ and $\vec{Z}^{(n)} \xrightarrow{a.s.} \vec{a}$ where $\vec{a}$ is a constant vector, then $(\vec{Y}^{(n)}, \vec{Z}^{(n)}) \xrightarrow{\mathcal{D}} (\vec{Y}, \vec{a})$.*

Proof of Proposition 3.1:

For each system size n, let $\hat{\mu}_n$ and $\hat{\sigma}_n$ be the estimates of the mean and standard deviation of the bandwidth distribution of the flow, respectively. By definition of the MBAC,

$$M_0^{(n)} = \frac{1}{4\hat{\mu}_n^2}\left(\sqrt{\hat{\sigma}_n^2\alpha_q^2 + 4n\mu\hat{\mu}_n} - \hat{\sigma}_n\alpha_q\right)^2$$

which is obtained by solving eqn. (6) for each n. Thus,

$$\frac{M_0^{(n)} - n}{\sqrt{n}} = \frac{\sqrt{n}(\mu - \hat{\mu}_n)}{\hat{\mu}_n} + \frac{\hat{\sigma}_n^2\alpha_q^2}{2\hat{\mu}_n^2\sqrt{n}} - \frac{\hat{\sigma}_n\alpha_q}{2\hat{\mu}_n^2}\sqrt{\frac{\hat{\sigma}_n^2\alpha_q^2}{n} + 4\mu\hat{\mu}_n}$$

By the strong law of large numbers $\hat{\mu}_n \xrightarrow{a.s.} \mu$ and $\hat{\sigma}_n \xrightarrow{a.s.} \sigma$. For the first term above, $\sqrt{n}(\mu - \hat{\mu}_n) \xrightarrow{\mathcal{D}} -\sigma Y_0$ by the Central Limit Theorem, where $Y_0 \sim N(0,1)$. Also, $\hat{\mu}_n \xrightarrow{a.s.} \mu$ and hence by theorems (8.2) and (8.1) above, the first term converges to $-\frac{\sigma}{\mu}Y_0$ in distribution. The second term converges almost surely to 0, while the third term converges almost surely to $-\frac{\sigma\alpha_q}{\mu}$. Applying the above theorems we now get the desired result:

$$\frac{M_0^{(n)} - n}{\sqrt{n}} \xrightarrow{\mathcal{D}} -\frac{\sigma}{\mu}(Y_0 + \alpha_q)$$

□