A FRAMEWORK FOR ROBUST MEASUREMENT-BASED ADMISSION CONTROL

by

Matthias Grossglauser and David Tse

Memorandum No. UCB/ERL M98/17

14 April 1998

A FRAMEWORK FOR ROBUST MEASUREMENT-BASED ADMISSION CONTROL

by

Matthias Grossglauser and David Tse

Memorandum No. UCB/ERL M98/17

14 April 1998

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

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 new insights 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

*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.

of the existing and the new flow. However, it is usually difficult for the user to tightly characterize his traffic in advance [17]. For many types of traffic, such as variable-bit-rate (VBR) compressed video, it is very difficult to define adequate traffic descriptors that take into account the widely varying source characteristics occurring at a slow time-scale (such as due to scene changes). 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.) Traffic descriptors such as the leaky bucket that has been proposed in standard bodies like the ATM Forum is only adequate for describing the short time-scale burstiness of the variable-rate traffic. As a result, traffic specifications can be expected to be quite loose, resulting in conservative use of resources.

Stochastic models such as those based on effective bandwidth [14] 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 [14]. 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 these problems 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* up-front, i.e. assuming that the estimated parameters are the real parameters, is dangerous, as we simply ignore its impact on the quality of service.
- **Dynamics and separation of time-scale**. A MBAC is a dynamical system, with flow arrivals and departures, and parameter estimates that vary with time. Since the estimation process measures

the in-flow burst statistics, while the admission decisions are made for each arriving flow, MBAC inherently links the flow and burst time-scale dynamics. Thus, the question of 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 large a memory window will reduce the adaptability of MBAC to non-stationarities in the statistics. A key issue is therefore to determine an appropriate memory window size 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.

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 Section 5, we summarize the insights obtained and use them to study the problem of choosing the appropriate memory window size. We also report some simulation results on real and synthetic traffic. In 6, we 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 [10] 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.

Third, the flows are assumed to have homogeneous statistics. How this assumption can be relaxed will

be discussed in Section 5.4.

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

[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.

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:

$$m^* = n - \frac{\sigma\alpha_q}{\mu}\sqrt{n} + o(\sqrt{n}) \tag{5}$$

where $\alpha_q := Q^{-1}(p_q)$ and $o(\sqrt{n})$ denotes a term which grows slower than $\sqrt{n}$. 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

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i(0) \qquad \text{and} \qquad \hat{\sigma} = \left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i(0) - \hat{\mu})^2\right]^{\frac{1}{2}} \tag{7}$$

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.

[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.

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:

$$\begin{aligned}\hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} X_i(0) = \mu + \frac{1}{\sqrt{n}}\left\{\frac{1}{\sqrt{n}}\left[\sum_{i=1}^{n} X_i(0) - n\mu\right]\right\} \\ &= \mu + \frac{\sigma Y_0}{\sqrt{n}} + o(\frac{1}{\sqrt{n}}) \end{aligned} \tag{8}$$

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:

$$\hat{\sigma} = \sigma + \frac{Z_0}{\sqrt{n}} + o(\frac{1}{\sqrt{n}}) \tag{9}$$

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), we get

$$\frac{\Delta_M}{\sqrt{n}} + \frac{\sigma}{\mu} Y_0 = o(1)$$

and hence

$$M_0 = m^* - \frac{\sigma}{\mu} Y_0 \sqrt{n} + o(\sqrt{n}). \tag{10}$$

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 (10), we get M_0 in terms of the system size n:

$$M_0 = n - \frac{\sigma}{\mu}(Y_0 + \alpha_q)\sqrt{n} + o(\sqrt{n}) \tag{11}$$

Although we have derived the result somewhat heuristically, it can be made precise by the following result, which is proved in Appendix 7.

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 $-\frac{\sigma}{\mu}(Y_0 + \alpha_q)$.*

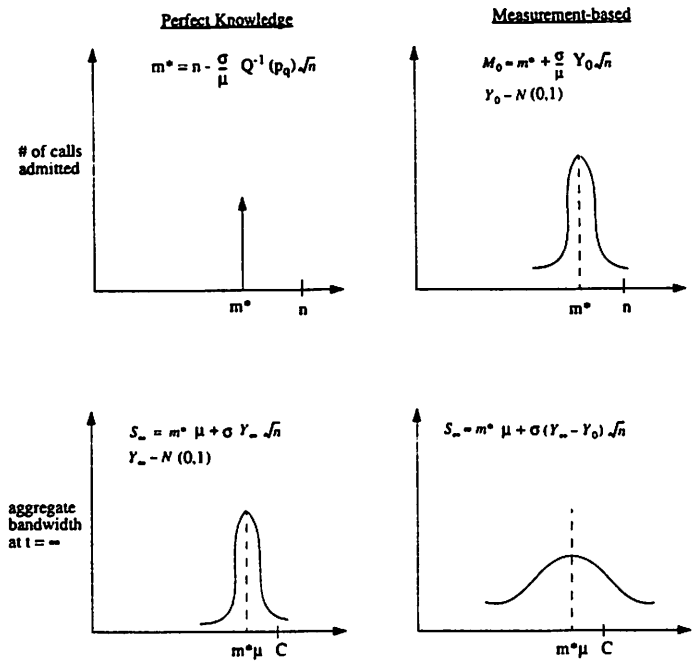

Figure 1: Uncertainty due to fluctuation in the number of flows (top) and in the aggregate bandwidth (bottom), for an admission controller with perfect knowledge (left) and for an MBAC (right).

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 *[3, p. 369, problem 27.14] Let $X_1, X_2, \ldots$ 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:

$$S_t := \sum_{i=1}^{M_0} X_i(t) = M_0\mu + \sigma Y_t\sqrt{n} + o(\sqrt{n}) \tag{12}$$

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

$$\frac{1}{\sigma\sqrt{n}}\left[\sum_{i=1}^{n} X_i(t) - n\mu\right] \tag{13}$$

Intuitively, eqn. (12) 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. (10), we get

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

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

$$p_f \approx Q\left(\frac{\alpha_q}{\sqrt{2}}\right). \tag{14}$$

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.

$$p_{ce} = Q\left(\sqrt{2}\alpha_q\right) \qquad \text{or} \qquad \alpha_{ce} := Q^{-1}(p_{ce}) = \sqrt{2}\alpha_q. \tag{15}$$

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. (10). 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)):

$$p_f(\widehat{\mu}, \sigma, m(\widehat{\mu})) = p_q. \tag{16}$$

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 $\widehat{\mu}$.

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

Using (16), 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 [21].

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 holding time 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 holding times 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

$$p_t = \exp\left(-\frac{t}{T_h}\right). \tag{17}$$

Furthermore, we let $\rho(t)$ denote a flow's auto-correlation function, where

$$\rho(t) := \frac{E[(X_i(0) - \mu)(X_i(t) - \mu)]}{\sigma^2}.$$

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. (10), this implies that

$$E[N_t] = p_t n - \frac{p_t \sigma \alpha_q}{\mu}\sqrt{n} + o(\sqrt{n})$$

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:

$$D[0,t] = \frac{t}{\widetilde{T_h}}\sqrt{n} + o(\sqrt{n}) \tag{18}$$

Using eqn. (10), the number of flows left in the system at time t can therefore be given by:

$$N_t = M_0 - D[0,t] = n - \left[\frac{\sigma}{\mu}(Y_0 + \alpha_q) + \frac{t}{\widetilde{T_h}}\right]\sqrt{n} + o(\sqrt{n}) \tag{19}$$

Using Lemma 3.2, the aggregate load at time t is:

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

$$\approx \quad n\mu + \sigma\left(Y_t - Y_0 - \frac{\mu t}{\sigma\widetilde{T_h}} - \alpha_q\right)\sqrt{n} \tag{20}$$

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 [3], 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

$$\begin{aligned} p_f(t) &\approx \Pr\left\{Y_t - Y_0 > \frac{\mu t}{\sigma\widetilde{T_h}} + \alpha_q\right\} \\ &= Q\left(\frac{1}{\sqrt{2(1-\rho(t))}}\left[\frac{\mu}{\sigma}\frac{t}{\widetilde{T_h}} + \alpha_q\right]\right) \end{aligned} \tag{21}$$

From (21), 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 starts at time 0. Our goal is to find the overflow probability at an arbitrary time t, particularly at $t = \infty$ which yields the steady-state overflow probability. 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 (22), M_t is given by:

$$Q\left[\frac{n\mu - M_t\hat{\mu}(t)}{\hat{\sigma}(t)\sqrt{M_t}}\right] = p_q, \tag{22}$$

where

$$\widehat{\mu}(t) = \frac{1}{n}\sum_{i=1}^{n} X_i(t) \qquad \text{and} \qquad \widehat{\sigma}(t) = \left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i(t) - \widehat{\mu}(t))^2\right]^{\frac{1}{2}} \tag{23}$$

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

$$N_t = N_{s^*} - D[s^*,t] = M_{s^*} - D[s^*,t] \tag{24}$$

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

$$N_t = N_s + A[s,t] - D[s,t] \geq N_s - D[s,t] \geq M_s - D[s,t] \tag{25}$$

where $A[s,t]$ is the number of flows *admitted* during $[s,t]$. Thus we conclude from (24) and (25) that

$$N_t = \sup_{0\leq s\leq t}\{M_s - D[s,t]\} \tag{26}$$

Eqn. (19) 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 superposition of the effects from all impulsive arrival times, starting at time 0.

Using formula (26), we can approximate N_t using our approximations for M_s and $D[s,t]$ as discussed in the previous section. Eqn. (10) gives an approximation for M_s:

$$M_s = n - \frac{\sigma}{\mu}(Y_s + \alpha_q)\sqrt{n} + o(\sqrt{n}) \tag{27}$$

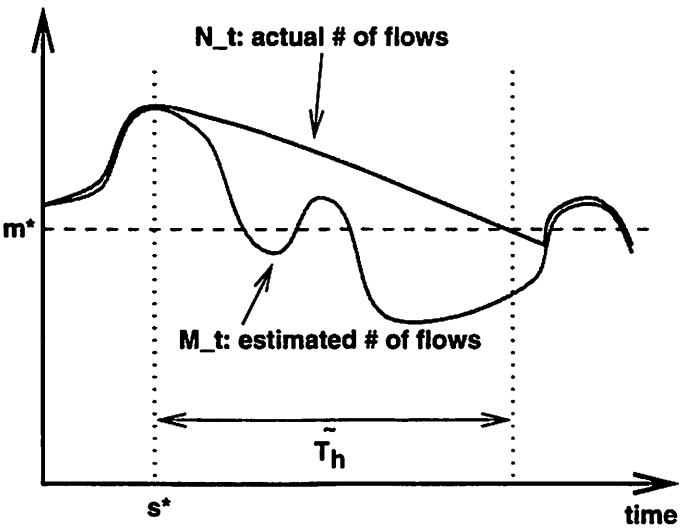

Figure 2: The relationship between the current estimate of admissible number of flows M_t and the actual number of flows N_t. The time-scale $\widetilde{T_h}$ is the typical time for the system to recover from admission errors.

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. (18) suggests an approximation for $D[s,t]$:

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

The following limit theorem makes the approximation of N_t precise.

Theorem 4.1 *For the system of size n, let $\{N_t^{(n)}\}$ be the process describing the evolution of the number of flows in the system. Assume condition B.6 is satisfied. If we scale the flow holding time as $T_h^{(n)} = \widetilde{T_h}\sqrt{n}$, where $\widetilde{T_h}$ is a fixed constant, then as $n \to \infty$, for each t, $\frac{N_t^{(n)}-n}{\sqrt{n}}$ converges in distribution to*

$$\frac{\sigma}{\mu} \sup_{0 \le s \le t} \left\{ -Y_s - \frac{\mu(t-s)}{\sigma \widetilde{T_h}} - \alpha_q \right\}$$

where $\{Y_t\}$ is defined as above.

Since we are now dealing with random *processes* rather that random variables as in the last two sections, the proof of this theorem is more technical and involves the notion of *weak convergence*. It is given in Appendix B. Condition B.6 contains mild technical assumptions on the individual flow processes; these are also stated in the appendix. These assumptions hold for a very broad class of models. For example, they hold for if each individual flow is a Markov modulated fluid

Once we obtained an approximation for N_t, we can immediately deduce an approximation for the aggregate load S_t at time t and hence the steady-state overflow probability p_f, using the same argument as for the impulsive load model.

Proposition 4.2 *For the system of size n, let $S_t^{(n)}$ be the aggregate load t time t and $p_f^{(n)}(t)$ be the overflow probability at time t. Then $\frac{S_t^{(n)}-n\mu}{\sigma\sqrt{n}}$ converges in distribution as $n \to \infty$ to*

$$\sup_{0\le s\le t}\left\{Y_t - Y_s - \frac{\mu}{\sigma\widetilde{T}_h}(t-s) - \alpha_q\right\}$$

and the overflow probability $p_f^{(n)}(t)$ converges to:

$$Pr\left\{\sup_{0\le s\le t}\left\{Y_t - Y_s - \frac{\mu}{\sigma\widetilde{T}_h}(t-s)\right\} > \alpha_q\right\}.$$

For brevity, we will define the important parameter:

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

The steady-state overflow probability can then be approximated by taking $t = \infty$ in Prop. 4.2 and using stationarity of $\{Y_t\}$ to get:

$$p_f \approx \Pr\left\{\sup_{s\le 0}\{Y_0 - Y_s + \beta s\}\right\} \tag{29}$$

Interestingly, one can interpret the limiting overflow probability at time t as that of the length of a certain *queue* at time t 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

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

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 obtained by applying results due to Braker [11, 12] on hitting probabilities of locally stationary Gaussian processes, extending the results by [7] for stationary processes. 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 $t = 0$. [3] Then an approximation to the hitting probability is given by:

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

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 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

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

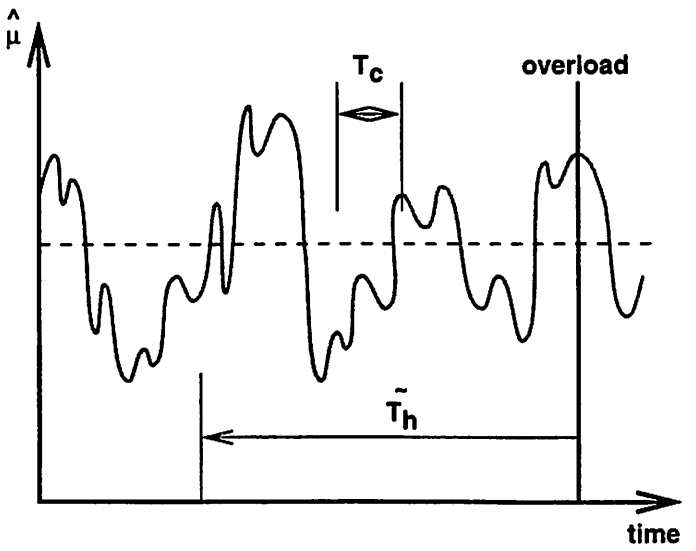

Figure 3: The ratio of correlation time-scale T_c and of the critical time-scale $\widetilde{T_h}$ determines the overflow probability.

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 $\frac{\phi(x)}{x} \approx Q(x)$),

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

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

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. (23), the MBAC would use:

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

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 Theorem 4.1 and Prop. 4.2 in the memoryless case, we can show:

Theorem 4.3 *For the system of size n, let $\{N_t^{(n)}\}$ be the process describing the evolution of the number of flows in the system. Assume condition B.6 is satisfied. If we scale the flow holding time as $T_h^{(n)} = \widetilde{T_h}\sqrt{n}$, where $\widetilde{T_h}$ is a fixed constant, then as $n \to \infty$, for each t, $\frac{N_t^{(n)}-n}{\sqrt{n}}$ converges in distribution to*

$$\frac{\sigma}{\mu} \sup_{0\le s\le t}\left\{-Z_s - \frac{\mu(t-s)}{\sigma\widetilde{T_h}} - \alpha_q\right\} \tag{36}$$

where $Z_t = (h*Y)_t$ and $\{Y_t\}$ is a zero-mean, unit-variance stationary Gaussian process with autocorrelation function same as that of an individual flow. The overflow probability $p_f^{(n)}(t)$ at time t converges to:

$$Pr\left\{\sup_{0\le s\le t}\left\{Y_t - Z_s - \frac{\mu}{\sigma\widetilde{T_h}}(t-s)\right\} > \alpha_q\right\}.$$

One can interpret Z_t as the error in the *filtered* estimate of the mean bandwidth of a flow at time t. The steady-state overflow probability under the MBAC with memory can therefore be approximated by:

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

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 [11, 12]:

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

where

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

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

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

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

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

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:

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

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

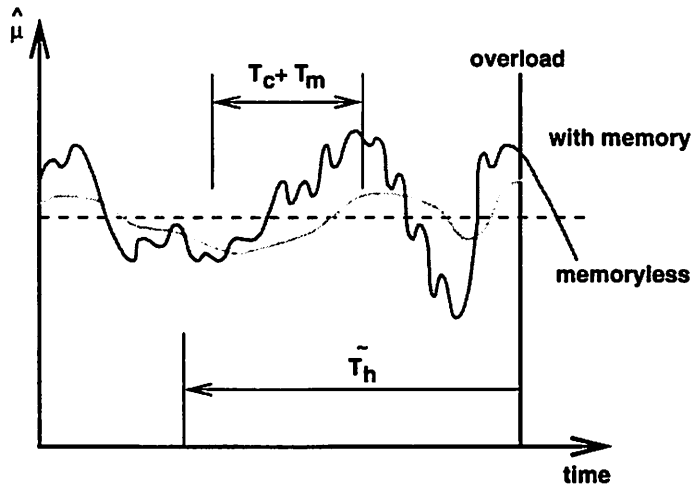

Figure 4: Estimation memory reduces the variance of the bandwidth estimator, and also smoothes its fluctuation.

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 $\frac{\widetilde{T_h}}{\sqrt{2T_c}}$ 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).

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:

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

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

$$\sigma\sqrt{n}\left[Q^{-1}(p_{ce}) - Q^{-1}(p'_{ce})\right] \tag{40}$$

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

5 Discussions and Simulations

We begin this section by summarizing the results obtained from our model. We then report some simulation results. We then illustrate how robust measurement-based admission control can be performed using our results. Finally, we investigate the sensibility of our approach to several of the assumptions in our model.

5.1 Summary of Results

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. This is in spite of the fact that the estimation errors are unbiased. 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.
- 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 reduces the overflow probability 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, and the formula (39) specialized to the regime of separation of flow and burst time-scales Moreover, formula (40) 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.

5.2 Simulation Results

We use RCBR (Renegotiated Constant Bit Rate [10]) 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 two orders of magnitude below the target overflow probability. The latter criterion is to terminate simulations within a reasonable time for very small p_f. In that case, we report an estimated p_f obtained by keeping track of the empirical mean $\widehat{\mu}$ and variance $\widehat{\sigma^2}$ of the aggregate bandwidth at the sample points and computing p_f as $Q(\frac{c-\widehat{\mu}}{\widehat{\sigma}})$.

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.

We now describe some simulations we have performed to validate the above insights. In particular, we wish to verify that our formulas can be used to perform robust measurement-based admission control. We proceed in two steps. First, we compare the overflow probability p_f obtained through simulation to the value predicted by theory. Second, we invert (38) to obtain an adjusted target overflow probability p_{ce} such that $p_f(T_m', T_c', p_{ce}) = p_q$. We then simulate the system with this adjusted target overflow probability in order to check if the overflow probability p_f really is close to the target overflow probability p_q regardless of the other parameters.

Figure 5 shows the overflow probability p_f as a function of the memory window size T_m. We observe that the overflow probability predicted by theory is conservative with respect to the simulated value. We attribute this offset to the assumptions in our model, such as ignoring the discreteness of the number of flows. However, the shape of the graphs correspond very well; in particular, the knee, corresponding to the value of T_m beyond which using a longer memory window size has little additional benefit, is well matched. Figure 6 and 7 demonstrate that our formulas can be used to perform robust measurement-based admission control. We see that with an adjusted overflow probability target, the actual overflow probability is slightly smaller than p_q over the whole range of parameters (cf. Fig. 7). Note that for small T_m, the adjusted

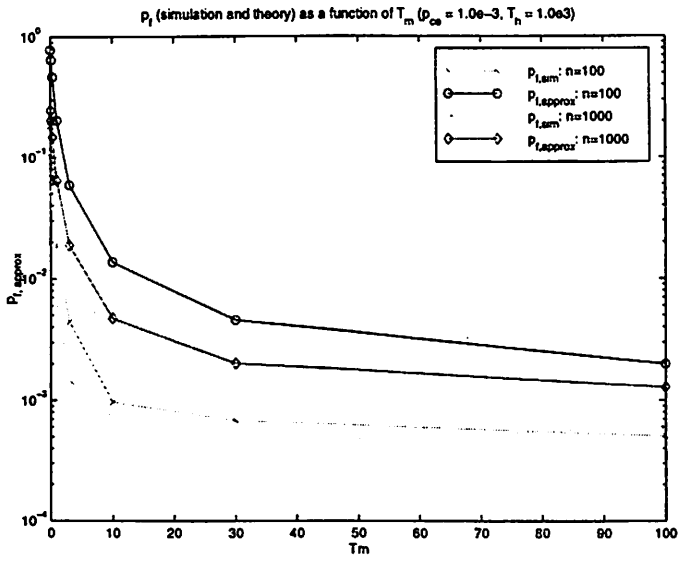

Figure 5: The overflow probability p_f as predicated by theory (eqn. (38)) and obtained by simulation ($T_h = 1000$, $T_c = 1.0$, $p_{ce} = 1.0e-3$).

target overflow probability p_{ce} can be very small ($< 1e-10$) with respect to the target overflow probability p_q of $1e-3$.

5.3 Memory Window Size and Robust MBAC

So far, we have assumed that the correlation time-scale parameter T_c and the flow holding time T_h are known. In practice, it is usually not very difficult to obtain a good estimate of the average holding time T_h of flows. On the other hand, the correlation time-scale T_c and more generally the correlation structure of the traffic is hard to estimate in practice, as realistic auto-correlation functions are more complex than a pure exponential. Therefore, we would like to design the MBAC such that its performance is good over a wide range of values for T_c. We claim that this can be accomplished by choosing the memory window length T_m on the order of the critical time-scale $\widetilde{T_h}$. For concreteness, let us pick the window size T_m to be $\widetilde{T_h}$ and examine the performance of the system for a range of T_c.

First, assume T_c is small with respect to $\widetilde{T_h}$. This is the separation of time-scale regime and formula (39) applies and holds for all T_m. Using the fact that $T_m = \widetilde{T_h} \gg T_c$, we get the further approximation:

$$p_f \approx \left(\frac{\sigma\alpha_q}{\mu} + 1\right) p_q \tag{41}$$

which is of the order of p_q. In this regime, the effect of the estimator memory effectively smoothes the fluctuations of the traffic and obtain a reliable estimate of the mean traffic rate. Although this result is derived using the simple exponential auto-correlation function (31), it can be easily shown that in this

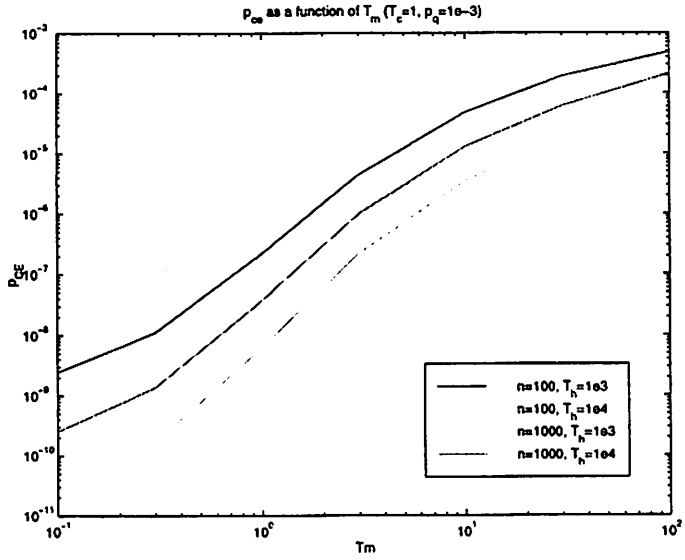

Figure 6: The adjusted target overflow probability by inversion of (38), for $n = \{100, 1000\}$, $T_h = \{1e3, 1e4\}$, and $p_q = 1.0e - 3$.

regime, the detailed correlation structure is not relevant and a similar approximation holds for other auto-correlation functions. We call this the *masking regime* (cf. Figure 8) because the memory window size masks the impact of the parameter T_c on the overflow probability p_f; the fluctuation time-scale of the mean estimator is determined by T_m alone.

Next, let us consider the other extreme, when T_c is much longer than $\widetilde{T_h}$. In this case, $\gamma = \frac{\widetilde{T_h}}{T_c}\frac{\sigma}{\mu} \ll 1$, and we have the approximation:

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

Substituting this into the general formula (37) and evaluating the integral, we get:

$$p_f \approx \frac{1}{\sqrt{2\pi}} \frac{T_c}{\widetilde{T_h}} \frac{\sigma}{\mu} \exp\left[-\left(\frac{T_c}{\widetilde{T_h}}\right)^2 \alpha_q^2 \right]$$

which definitely meets the target QoS since $T_c \gg \widetilde{T_h}$ in this regime. In contrast to the masking regime, the time-scale of the estimator fluctuation is dominated by T_c. The memory window is effectively useless in this regime, as it does not reduce estimation errors. However, the fluctuation of the estimators around their mean is at a time-scale longer than the critical time-scale. This is precisely the regime where the repair effect makes overflow unlikely. Therefore, we call this the *repair regime* (cf. Fig. 8).

For T_c in between the two extremes, there is no closed-form expression for the overflow probability, and we resort to a numerical integration of the formula (37) to study the performance of the MBAC. This is shown in Fig. 9, where we plot the overflow probability as a function of $T_m/\widetilde{T_h}$ and T_c. We see that while for small $T_m/\widetilde{T_h}$ the performance is not robust, the QoS is satisfied over a wide range of T_c once the

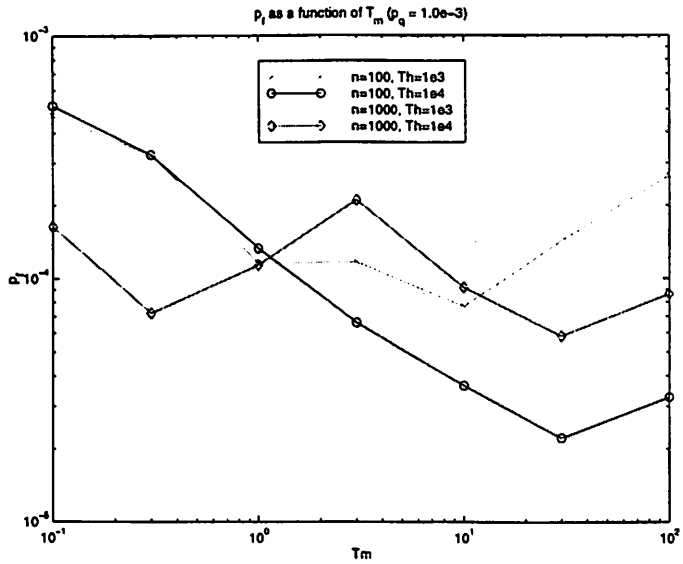

Figure 7: The simulated overflow probability p_f using the adjusted target overflow probability shown in the previous graph.

memory window size is chosen to be a significant fraction of $\widetilde{T_h}$. This is further corroborated by simulation results using RCBR shown in Fig. 10.

Another way of understanding these two regimes is through a slightly different viewpoint of the MBAC. So far, we have assumed that the goal of the measurement process is to estimate stationary traffic parameters (μ and σ^2) in order to keep the number of admitted flows as close as possible to the number of flows m^* we would admit if μ and σ^2 were known exactly. The goal of this process is to maintain enough spare bandwidth such that the target overflow probability is not exceeded. However, we have seen that flow departures reduce the time-scale over which admission errors persist to the critical time-scale $\widetilde{T_h}$. The system has a relaxation period over which it "forgets" past admission mistakes. What we effectively do by setting the memory window size to $\widetilde{T_h}$ is to let the estimators *track* the traffic fluctuation over this relaxation time-scale. This is appropriate as we only need to accurately *predict* the parameters over a time-scale of $\widetilde{T_h}$. Longer-term fluctuations are implicitly absorbed by the system.

The above analysis and simulations are based on a traffic model with correlation at a single time-scale. In practice, traffic fluctuations may occur at multiple time-scales. In particular, several studies of various types of network traffic have found phenomenon of long-range dependence (LRD) [16, 8, 1, 6]. However, based on the intuition gained from the single time-scale model , we expect that a memory window size on the order of $\widetilde{T_h}$ is again appropriate here. As before, flow departures dictate a critical time-scale $\widetilde{T}_h$ over which the statistics of the future behavior of the traffic has to be predicted. A memory window of $\widetilde{T}_h$ allows the simultaneous *smoothing* of the fluctuations faster than $\widetilde{T}_h$ for reliable estimation and the *tracking* of fluctuations at a time-scale larger than $\widetilde{T}_h$. The statistics of the long-term fluctuations of long-range

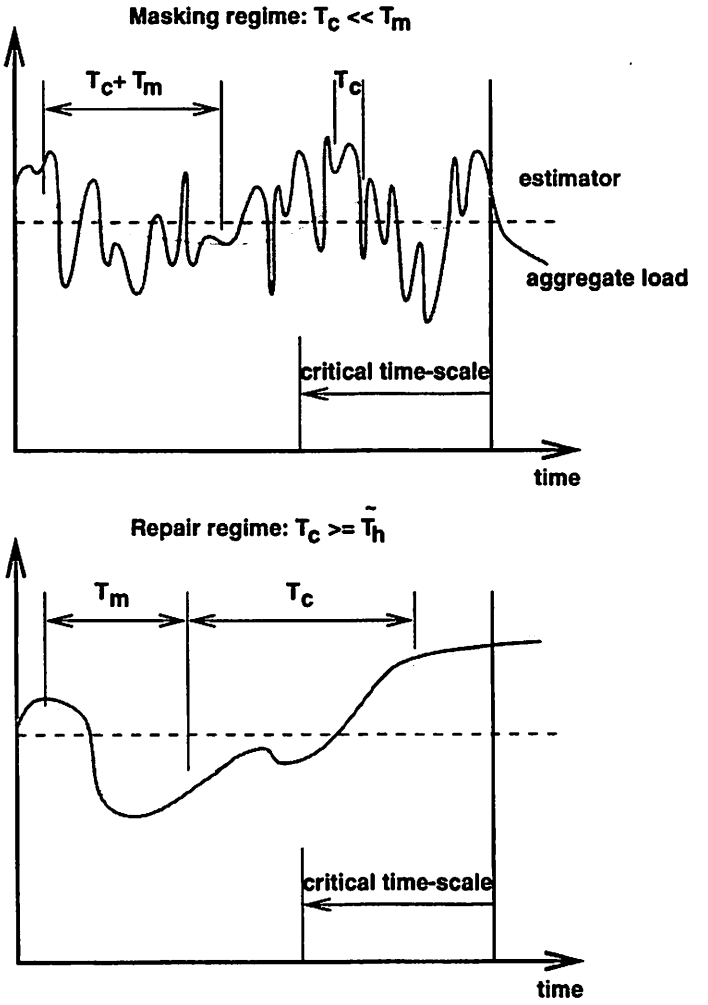

Figure 8: An illustration of the masking regime and of the repair regime.

dependence is therefore irrelevant.

To provide some support for this hypothesis, we present simulation results on an actual traffic trace. Figure 11 and 12 show the overflow probability when the flow is a piecewise CBR version of the MPEG-1 encoded Starwars movie [10]. This particular trace has been shown to exhibit long-range dependence [8]. We vary the average flow holding time and plot the overflow probability as a function of $1/\widetilde{T_h}$. As with the synthetic traffic above, we see that the performance is not robust under memoryless estimation. When $\widetilde{T_h}$ is large (corresponding to small T_c in Fig. 9), the performance misses the target by 1 or 2 orders of magnitude. On the other hand, we note that with the choice of memory window size $T_m = \widetilde{T_h}$, the MBAC is robust (cf. Fig. 12). Apparently, the strong long-term fluctuations of this traffic do not degrade the performance of the MBAC.

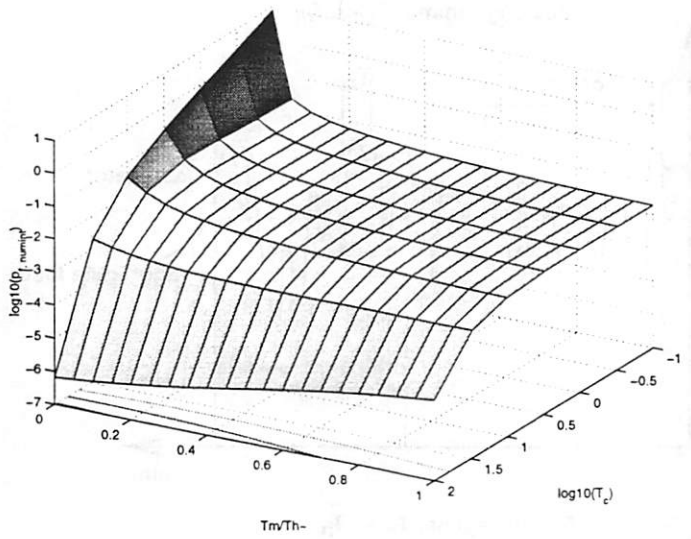

Figure 9: The overflow probability p_f obtained by numerical integration of (37), as a function of the normalized memory window size $T_m/\widetilde{T_h}$ and of the correlation time-scale T_c.

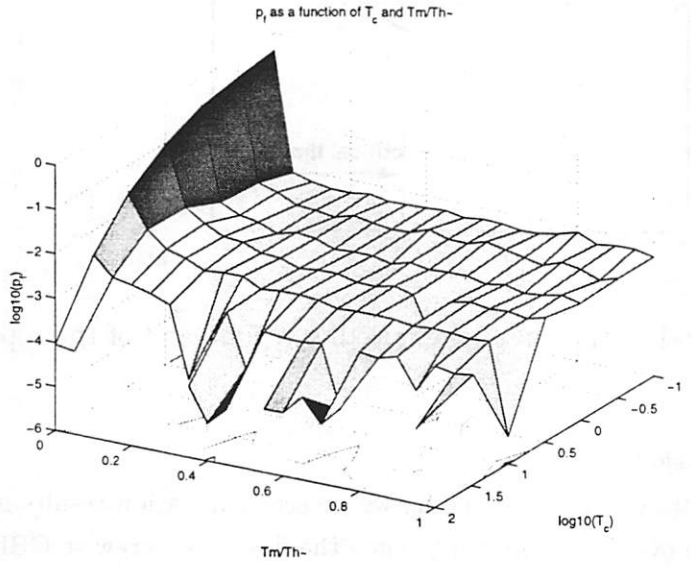

Figure 10: The simulated p_f over the same parameter range as in Fig. 9.

5.4 Heterogeneous Flows

The approach taken in this paper is to use a Gaussian approximation (justified by a heavy-traffic limit) for the *aggregate* flow and focus on estimating the mean and variance of the aggregate fluctuation. We did however make the assumption of *homogeneous* individual flows, each with same mean μ, variance σ^2 and

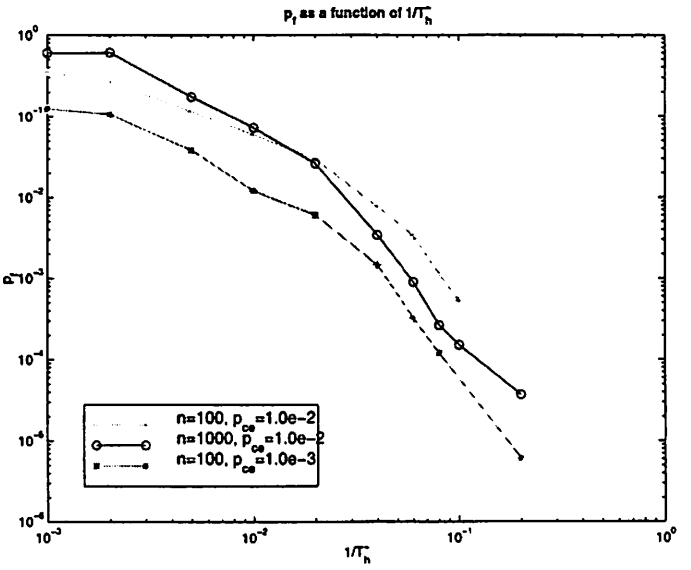

Figure 11: The overflow probability for Starwars sources with memoryless estimation ($T_m = 0$).

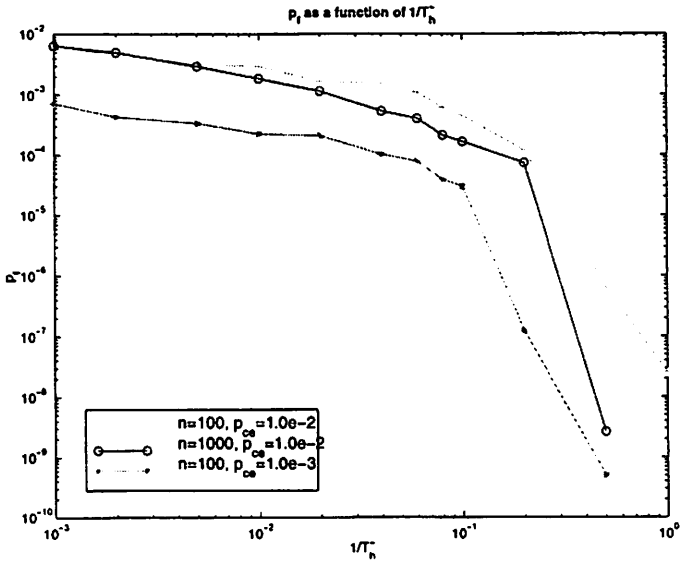

Figure 12: The overflow probability for Starwars sources with $T_m = \widetilde{T_h}$.

mean holding time T_h. A straightforward extension of our analysis shows that the results carry through for the situation when the average holding times of the flows are different, with the critical time-scale $\widetilde{T_h}$ given by the reciprocal of the departure rate averaged over all flows in the system.

Heterogeneous flows with different mean bandwidth requirements has also no effect on the mean esti-

mator used in the paper, provided that a Gaussian approximation of the aggregate is still valid. This is because the mean estimator only makes use of the aggregate and not the individual rates anyway. However, this is not true for the *variance* estimator, as it makes use of the individual rates in the estimate. In fact, it is not difficult to show that the variance estimator (7), which treats each flow as though they have the same mean, is always biased when the flows have different mean and over-estimates the variance. Thus, in a heterogeneous environment, the scheme proposed here will be conservative, leading to some loss in utilization. However, it is robust with respect to the homogeneity assumption. Of course, if classification of the flows is available to the MBAC, one can modify the variance estimator, using a different mean estimate for each class.

6 Related Work

Past work on measurement-based admission control [5], [18], [13] 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 [15], 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.* [9] 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. However, there are also some additional issues that merit attention.

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 [4]. 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 [19]. 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.

References

[1] J. Beran, R. Sherman, and W. Willinger. Long Range Dependence in Variable Bit Rate Video Traffic. *IEEE Trans. on Communications*, 43(3):1566–1579, February 1995.

[2] P. Billingsley. *Convergence of Probability Measures.* Wiley,New York, 1968.

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

[4] 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.

[5] 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.

[6] M. Crovella and A. Bestavros. Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes. In *Proc. ACM Sigmetrics '96*, pages 160–169, Philadelphia, PA, May 1996.

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

[8] M. W. Garrett and Walter Willinger. Analysis, Modeling and Generation of Self-Similar VBR Video Traffic. In *Proc. ACM SIGCOMM '94*, pages 269–280, London, UK, August 1994.

[9] R.J. Gibbens, F.P. Kelly, and P.B. Key. A Decision-theoretic Approach to Call Admission Control in ATM Networks. *IEEE JSAC, Special issue on Advances in the Fundamentals of Networking*, August 1995.

[10] M. Grossglauser, S. Keshav, and D. Tse. RCBR: A Simple and Efficient Service for Multiple Time-Scale Traffic. *IEEE/ACM Transactions on Networking*, December 1997.

[11] H. U. Bräker. High boundary excursions of locally stationary Gaussian processes. In *Proc. of the Conference on Extreme Value Theory and Applications*, Gaithersburg, Maryland, USA, May 1993.

[12] H. U. Bräker. *High boundary excursions of locally stationary Gaussian processes.* PhD thesis, Universität Bern, Switzerland, 1993.

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

[14] J.Y. Hui. Resource Allocation for Broadband Networks. *IEEE Journal on Selected Areas in Communications*, 6(9), December 1988.

[15] 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.

[16] Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson. On the Self-Similar Nature of Ethernet Traffic (Extended Version). *IEEE/ACM Trans. on Networking*, 2(1):1–15, February 1994.

[17] 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.

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

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

[20] D.W. Stroock and S.R.S. Varadhan. *Multidimensional Diffusion Proceeses.* Springer Verlag, 1979.

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

A Proof of Proposition 3.1:

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 A.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 A.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,

$$M_0^{(n)} = \frac{1}{4(\hat{\mu}^{(n)})^2}\left(\sqrt{(\hat{\sigma}^{(n)})^2\alpha_q^2 + 4n\mu\hat{\mu}^{(n)}} - \hat{\sigma}^{(n)}\alpha_q\right)^2 \tag{42}$$

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

$$\frac{M_0^{(n)} - n}{\sqrt{n}} = \frac{\sqrt{n}(\mu - \hat{\mu}^{(n)})}{\hat{\mu}^{(n)}} + \frac{(\hat{\sigma}^{(n)})^2\alpha_q^2}{2(\hat{\mu}^{(n)})^2\sqrt{n}} - \frac{\hat{\sigma}^{(n)}\alpha_q}{2(\hat{\mu}^{(n)})^2}\sqrt{\frac{(\hat{\sigma}^{(n)})^2\alpha_q^2}{n} + 4\mu\hat{\mu}^{(n)}} \tag{43}$$

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 (A.2) and (A.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:

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

□

B Weak Convergence Results for Heavy-Traffic Approximation

In this section, we will prove Theorem 4.1, giving a formal justification of the heavy traffic approximations we used. To begin, we will specify the space in which the sample paths of the processes live, and define the notion of weak convergence.

Definition B.1 *The space $\mathcal{D}[0,\infty]$ is the space of all real-valued functions on $\mathcal{D}[0,\infty]$ that are continuous from the right and have limits from the left. There is a metric (Skorohod metric) defined on $F[0,\infty]$ such that it is complete and separable.*

Definition B.2 *Let $\{Z_t^{(n)}\}$ be a sequence of processes whose sample paths are in $\mathcal{D}[0,\infty]$. $\{Z_t^{(n)}\}$ is said to* converges weakly *to $\{Z_t\}$ if for every continuous function $f : \mathcal{D}[0,\infty] \to \Re$, $E[f(\{Z_t^{(n)}\})] \to E[f(\{Z_t\})]$.*

With a slight abuse of notation, we will use $\xrightarrow{\mathcal{D}}$ to denote weak convergence of processes as well as convergence in distribution for random variables. We shall use the following theorem to verify weak convergence.

Theorem B.3 *A sequence of processes $\{Z_t^{(n)}\}$ converges weakly to $\{Z_t\}$ if all finite-dimensional distributions converge and $\{Z_t^{(n)}\}$ is* tight, *i.e.*

1) For every $\eta > 0$, there exists an $a > 0$ such that

$$Pr\left\{|Z_0^{(n)}| > a\right\} \leq \eta \qquad \forall n.$$

2) For every $T > 0$, $\epsilon, \eta > 0$, there exists a $\delta \in (0,1)$ and an integer n_0 such that

$$Pr\left\{\sup_{|t_1-t_2|\leq\delta, 0\leq t_1,t_2\leq T} |Z_{t_1}^{(n)} - Z_{t_2}^{(n)}| > \epsilon\right\} \leq \eta \qquad \forall n \geq n_0.$$

We will use the following theorems [2], which can be viewed as process-level analogs to A.1 and A.2.

Theorem B.4 *(Continuous-Mapping Theorem for Processes) Let $\{Z_t^{(n)}\}$ be a sequence of processes whose sample paths are in $\mathcal{D}[0,\infty]$. If $h : \mathcal{D}[0,\infty] \to \mathcal{D}[0,\infty]$ is continuous and $\{Z_t^{(n)}\} \xrightarrow{\mathcal{D}} \{Z_t\}$, then $g(\{Z_t^{(n)}\}) \xrightarrow{\mathcal{D}} g(\{Z_t\})$.*

Theorem B.5 *Let $\{W_t^{(n)}\}$ and $\{Z_t^{(n)}\}$'s be processes defined on the same probability space, and $g : D[0.\infty] \times D[0,\infty] \to D[0,\infty]$ is continuous. If $\{W_t^{(n)}\} \xrightarrow{\mathcal{D}} \{W_t\}$ and $\{Z_t^{(n)}\}$ converges weakly to a deterministic process $\{Z_t\}$, then $g(\{W_t^{(n)}\}, \{Z_t^{(n)}\}) \xrightarrow{\mathcal{D}} g(\{W_t\}, \{Z_t\})$.*

We need the following technical conditions on the flow processes.

Assumptions B.6 *1) The sample paths of the individual flow processes $\{X_i(t)\}$ are in $\mathcal{D}[0,\infty]$.*

2) The mean bandwidth estimates $\{\hat{\mu}_s^{(n)}\}$ converges weakly to the constant process μ.

3) The standard deviation estimates $\{\hat{\sigma}_s^{(n)}\}$ converges weakly to the constant process σ.

4) If we define

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

to be the scaled and centered sum of the individual flows, then as $n \to \infty$, $\{Y_t^{(n)}\}$ converges weakly to $\{Y_t\}$, which is a stationary zero-mean Gaussian process with unit variance and auto-correlation function $\rho(t)$ (that of an individual flow).

The second and third conditions say that the estimates as processes are consistent. The fourth condition says that the aggregation of the individual flows satisfies a functional central limit theorem. It holds for a very broad class of models for the individual sources. For example, it can be shown [20] that the condition holds if $\{X_i(t)\}$ is a K-state continuous-time Markov fluid, in which case the limiting process $\{Y_t\}$ is a linear functional of a $K-1$-dimensional diffusion process.

To prove the main theorem, we need the following lemma, which can be viewed as a functional law of large number for the process describing the evolution of the number of flows in the system.

Lemma B.7 *The process $\{\frac{N_t^{(n)}}{n}\}$ converges weakly to the deterministic process taking on a constant value of 1 for all t.*

Proof. Using eqn. (42), assumptions (2) and (3) in B.6, together with Theorem B.5, we can see that the process $\{\frac{M_s^{(n)}}{n}\}$ converges to the process taking on a constant value. Now, for all $t \geq 0$,

$$\frac{M_t^{(n)}}{n} \leq \frac{N_t^{(n)}}{n} \leq \sup_{0 \leq s \leq t} \frac{M_s^{(n)}}{n}$$

Since $\{\frac{M_s^{(n)}}{n}\}$ converges weakly to the constant process 1, so does the process $\{\sup_{0\leq s\leq t} \frac{M_s^{(n)}}{n}\}$, by the continuous mapping theorem. Hence $\{\frac{N_t^{(n)}}{n}\}$ must converge weakly to the constant process 1. □

Proof of Theorem 4.1

From eqn. (43), we get for each s,,

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

By assumption B.6, we know that $\{\sqrt{n}(\mu - \hat{\mu}_s^{(n)})\} \xrightarrow{\mathcal{D}} -\sigma Y_s$, where $\{Y_s\}$ is a zero mean Gaussian process with auto-correlation function ρ. Also, $\{\hat{\mu}_s^{(n)}\}$ converges weakly to the constant process μ and $\{\hat{\sigma}_s^{(n)}\}$ converges weakly to the constant process σ. By Theorem B.5,

$$\{\frac{M_s^{(n)} - n}{\sqrt{n}}\} \xrightarrow{\mathcal{D}} \{-\frac{\sigma}{\mu}(Y_s + \alpha_q)\} \tag{44}$$

Next, we will show that for fixed $t > 0$, $\{D[s,t]\}$ as a process in s converges weakly to the deterministic process $\{\frac{s-t}{\hat{T}_h}\}$ on $[0,t]$. First, let us fix an $s < t$. Define now two random variables $D^u[s,t]$ and $D^l[s,t]$. $D^l[s,t]$ is the number of flows departing from the system when there are $N(s)$ flows in the system at time s and no more flows enter the system in $[s,t]$; $D^u[s,t]$ is the number of flows departing from the system when there are $W := \sup_{\tau\in[s,t]} N(\tau)$ flows at time s and no more flows enter the system in $[s,t]$. It can be seen that for every x,

$$\Pr\left\{D^l[s,t] \geq x\right\} \leq \Pr\left\{D[s,t] \geq x\right\} \leq \Pr\left\{D^u[s,t] \geq x\right\} \tag{45}$$

Using Chebyshev's bound, we have for every $\epsilon > 0$,

$$\Pr\left\{\left|\frac{D^u[s,t]}{\sqrt{n}} - \frac{s-t}{\hat{T}_h}\right| > \epsilon\right\} \leq \frac{\mathrm{E}\left[\left(\frac{D^u[s,t]}{\sqrt{n}} - \frac{s-t}{\hat{T}_h}\right)^2\right]}{\epsilon^2}$$

The expectation can be computed using the fact that the flows have exponential holding time and departs from the system independently:

$$\begin{aligned} & \mathrm{E}\left[\left(\frac{D^u[s,t]}{\sqrt{n}} - \frac{s-t}{\hat{T}_h}\right)^2\right] \\ = \; & \frac{\mathrm{E}[W]}{n}q + \frac{\mathrm{E}[W^2] - \mathrm{E}[W]}{n}q^2 - \frac{(s-t)^2}{\hat{T}_h^2} \end{aligned} \tag{46}$$

where q is the probability that a given flow leaves the system some time in $[s,t]$, and is given by

$$q := 1 - \exp(\frac{-t}{\hat{T}_h\sqrt{n}}) \tag{47}$$

By Lemma B.7 and the continuous mapping theorem, as $n \to \infty$,

$$\mathrm{E}\left[\frac{W}{n}\right] \to 1 \qquad \mathrm{E}\left[\frac{W^2}{n^2}\right] \to 1$$

Substituting this into (46) shows that

$$\lim_{n\to\infty} \mathrm{E}\left[\left(\frac{D^u[s,t]}{\sqrt{n}} - \frac{s-t}{\hat{T}_h}\right)^2\right] = 0$$

Hence $\frac{D^u[s,t]}{\sqrt{n}}$ converges in probability, and hence in distribution, to $\frac{s-t}{\hat{T}_h}$. Using a similar argument, one can show the same thing for $D^l[s,t]$. By (45), this imply that for fixed s and t, $\frac{D[s,t]}{\sqrt{n}} \xrightarrow{\mathcal{D}} \frac{s-t}{\hat{T}_h}$. Using Theorem (A.2), this implies that for all k and $s_1,\ldots,s_k \in [0,t]$, $(\frac{D[s_1,t]}{\sqrt{n}},\ldots\frac{D[s_k,t]}{\sqrt{n}}) \xrightarrow{\mathcal{D}} (\frac{s_1-t}{\hat{T}_h},\ldots,\frac{s_k-t}{\hat{T}_h})$, i.e. finite-dimensional distributions converge. To show weak convergence as a process, we need to verify

tightness, according to Theorem B.3. The first condition is trivially satisfied. For the second condition,

$$\begin{aligned}
& \Pr\left\{\sup_{|s_1-s_2|\le\delta, 0\le s_1,s_2\le t} \frac{D[s_1,s_2]}{\sqrt{n}} > \epsilon\right\} \qquad (48)\\
\le\ & \Pr\left\{\sup_{0\le k\le \frac{t}{\delta}} \frac{D[k\delta,(k+1)\delta]}{\sqrt{n}} > \epsilon\right\}\\
\le\ & (\frac{t}{\delta}+1)\sup_k \Pr\left\{\frac{D[k\delta,(k+1)\delta]}{\sqrt{n}} > \epsilon\right\}\\
\le\ & (\frac{t}{\delta}+1)\frac{1}{\epsilon^2}\sup_k \mathrm{E}\left[\frac{1}{n}(D[k\delta,(k+1)\delta])^2\right]\\
\le\ & (\frac{t}{\delta}+1)\frac{1}{\epsilon^2}\left(\frac{\mathrm{E}[U]}{n}p + \frac{\mathrm{E}[U^2]-\mathrm{E}[U]}{n}p^2\right) \qquad (49)
\end{aligned}$$

where $U := \sup_{\tau\in[0,t]} N_\tau^{(n)}$ and

$$p = \Pr\{\text{a flow departs in time } [k\delta,(k+1)\delta]\} = 1-\exp(\frac{\delta}{\hat{T}_h\sqrt{n}}).$$

By direct calculation, (49) is in turn equal to

$$\frac{1}{\epsilon^2}(\frac{t}{\delta}+1)(\frac{\delta^2}{\hat{T}_h^2}+o(1))$$

where the $o(1)$ term goes to zero as $n\to\infty$. Thus, by appropriate choice of n and δ, (48) can be made arbitrarily small. this verifies the tightness of $\{\frac{D[s,t]}{\sqrt{n}}\}$ and hence its weak convergence.

Combining the weak convergence of $\{\frac{D[s,t]}{\sqrt{n}}\}$ and $\{\frac{M_t^{(n)}-n}{\sqrt{n}}\}$, it follows that

$$N_t^{(n)} \xrightarrow{\mathcal{D}} \sup_{0\le s\le t}\{-\frac{\sigma}{\mu}(Y_s - \frac{\mu(t-s)}{\sigma\hat{T}_h} + \alpha_q)\}$$

□