A Time-Scale Decomposition Approach to Measurement-Based Admission Control

Matthias Grossglauser
AT&T Labs - Research

David N. C. Tse
University of California at Berkeley

Abstract— **We propose a time-scale decomposition approach to measurement-based admission control (MBAC). We identify a *critical time-scale* $\widetilde{T_h}$ such that: 1) aggregate traffic fluctuation slower than $\widetilde{T_h}$ can be tracked by the admission controller and compensated for by flow admissions and departures; 2) fluctuations faster than $\widetilde{T_h}$ have to be absorbed by reserving spare bandwidth on the link. The critical time-scale is shown to scale as $T_h/\sqrt{n}$, where T_h is the average flow duration and n is the size of the link in terms of number of flows it can carry. A MBAC design is presented which filters aggregate measurements into low and high frequency components separated at the cutoff frequency $1/\widetilde{T_h}$, using the low frequency component to track slow time-scale traffic fluctuations and the high frequency component to estimate the spare bandwidth needed. Our analysis shows that the scheme achieves high utilization and is robust to traffic heterogeneity, multiple time-scale fluctuations and measurement errors. The scheme uses only measurements of aggregate bandwidth and does not need to keep track of per-flow information.**

I. Introduction

The main drawback of traditional admission control is the inability of the user or application to come up with tight traffic descriptors *before* establishing the flow. *Measurement-based admission control* (MBAC) avoids this problem by shifting the task of traffic specification from the application to the network, so that admission decisions are based on traffic measurements instead of an explicit specification [8], [5], [6] (cf. Fig. 1). This approach has several important advantages. First, the application-specified traffic descriptor can be trivially simple (e.g., a 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.

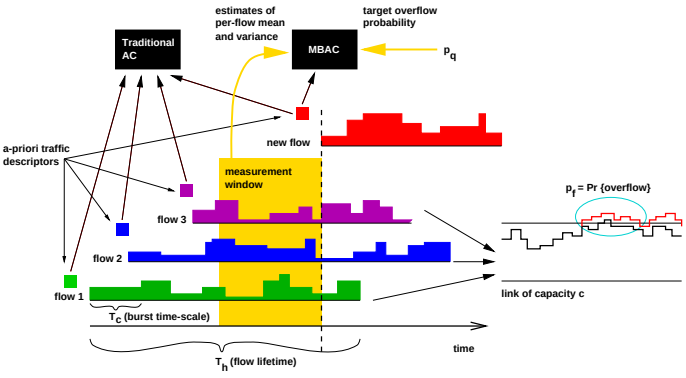

Fig. 1: Traditional admission control makes decisions based on the a-priori traffic descriptors of the existing and the new flow. Measurement-based admission control (MBAC) only uses the new flow's traffic descriptor, but *estimates* the behavior of the existing flows.

In order for an MBAC approach to be successful in practice, it has to fulfill several requirements.

- **Robustness:** An MBAC must be able to ensure a quality of service on behalf of applications in the same way as its a-priori descriptor based counterpart does. This is not trivial, as measurement inevitably has some uncertainty to it, leading to admission errors. The quality of service should also be robust to flow heterogeneity and to fluctuations on many time-scales that are a general property of network traffic [9], [10], [1], [4].
- **Resource utilization:** The quality of service can be improved by being very conservative in admission control, thereby allocating more resources per flow than necessary. Obviously, being too conservative is undesirable, as one also wants to maximize resource utilization, and admit as many flows as possible under the given QoS constraints.
- **Implementation:** The cost of deploying an MBAC system must be smaller than its benefits cited above. For this, the MBAC should be modular, in the sense that adding the measurement machinery to the existing infrastructure should be as nonintrusive as possible. Also, the computational complexity of the algorithm used to make admission decisions needs to be scalable in the flow arrival rate and in the link capacity.

In this paper, we propose an MBAC design that fulfills the above requirements. Our design is robust to fluctuations on multiple time-scales in the traffic and to flow heterogeneity, and achieves high link utilization despite the inherent measurement uncertainty. The scheme is also easy to implement as it only relies on *aggregate* bandwidth information, which implies that the MBAC does not have to maintain per-flow information and to inspect packet headers to identify flows.

Our proposed design is based on a *time-scale decomposition* approach. Flow arrival and departure dynamics are explicitly taken into account. The fact that flows only remain in the system for a finite time gives admission decisions a certain time-horizon, which we call the *critical time-scale*. This critical time-scale determines the fluctuations in the aggregate bandwidth that can be compensated through flow admissions and departures. For example, a slow increase in the aggregate bandwidth may be compensated for simply by rejecting new flows and waiting for some existing flows to depart, thereby avoiding overload. A slow decrease in the aggregate bandwidth may be compensated for by admitting more flows to benefit from the released bandwidth. The MBAC design exploits this by decomposing the aggregate bandwidth fluctuation into a fast time-scale and a slow time-scale component with respect to the critical time-scale. The fast time-scale component is used to estimate the spare bandwidth to be set aside to absorb short-term fluctuations that cannot be "followed" by flow arrivals and departures. The slow time-scale component is used to track fluctuations that do not need spare bandwidth, but are compensated by flow arrivals and departures. This results in higher utilization than a scheme which sets aside spare bandwidth for fluctuations at *all*

time-scales. We will show that an appropriate critical time-scale is $T_h/\sqrt{n}$, where T_h is the average flow duration in the system and n is the size of the system in terms of the number of flows it can carry.

In our earlier work on MBAC [6], the main issue we addressed was robustness with respect to measurement uncertainty. Using a simple, analytical model of an idealized MBAC, we studied the impact of measurement errors on the quality of service. The main insight gained from that model was an understanding of the complicated dynamics that arise as a result of bandwidth fluctuations, measurement uncertainty, flow arrivals and departures, and estimation memory. These insights motivate the MBAC design presented in this paper and serve as a basis for its performance analysis.

In the performance analysis of our proposed MBAC, we relax two assumptions made in our earlier work. First, we assume that the admission controller only has information about the evolution of the *aggregate bandwidth* available to make admission decisions. This is in contrast with our earlier work, where we assumed that the bandwidth of each individual flow is known. Basing admission decisions only on aggregate information is appealing from an implementation viewpoint, as we do not require the MBAC to gather and maintain per-flow information. This improves the MBAC's scalability, as it is not necessary to inspect packet or cell headers in order to determine flow identifiers. This also allows the MBAC to be architecturally more decoupled from the data path. Our goal in this paper is therefore to achieve the same *robustness* for an MBAC relying only on aggregate measurements that we achieve for an MBAC with per-flow information. To achieve this goal, we seek a clear understanding of the impact of errors associated with aggregate measurements.

Second, we consider the situation when flows are *heterogeneous*. As we consider the problem of admission control in the context of an integrated services packet network, flows can represent many different types of media (e.g., audio or video), they can be encoded at very different levels of quality, and they can use different end-to-end control mechanisms. Therefore, we must expect that flows are very heterogeneous in their statistical behavior. On the other hand, an *individual flow* corresponds typically to a single instance of an application (such as a videoconference), of an encoding method, and of a control mechanism. Therefore, we expect an *individual* flow to be well modeled as a stationary and ergodic random process. We will show that the proposed MBAC scheme performs well without *a priori* classfication of flows into different classes and relies only on aggregate measurements of all the flows in the system.

The paper is structured as follows. In Section II, the basic model is introduced. In the next two sections, we focus on two issues that are central to understanding the proposed MBAC design. In Section III, we first study the impact on performance of admission decisions only on aggregate bandwidth information, as opposed to per-flow bandwidth information. In Section IV, we identify the critical time-scale through a study of the dynamics of the system that arise due to fluctuations of the aggregate bandwidth of flows in the system, and due to flow arrivals and departures. Combining the insights obtained in these two sections, we present our MBAC design in Section V. In Section VI, we analyze the performance of the proposed MBAC scheme under both homogeneous and heterogeneous traffic models. Section VII contains our conclusions.

II. Basic Model

We will first outline the basic model which we will use throughout the paper to study various basic measurement-based admission control issues, to motivate our MBAC design and finally to analyze its performance.

The network resource considered is a bufferless single link with capacity c. Flows arrive over time, requesting service. Once admitted, the bandwidth requirement of a flow $\{X_i(\cdot)\}$ fluctuates over time while in the system. We assume that the flow holding time in the system is exponentially distributed with mean T_h; the departures of the flows are independent of each other and independent of the bandwidth processes $\{X_i(\cdot)\}$.

An admission control scheme decides whether to accept or reject a new flow requesting service; a *measurement-based* admission control (MBAC) scheme makes decisions based solely on observation of the past traffic flows.[1] Resource overload occurs when the instantaneous aggregate bandwidth demand S_t exceeds the link capacity, and the QoS is measured by the steady-state overflow probability $p_f := \Pr\{S_t > c\}$. The goal of an admission control scheme is to meet a desired QoS objective p_q (i.e. $p_f \leq p_q$) while maintaining a high average utilization $\mathrm{E}[S_t]$ of the link.

Several processes are of importance in this paper. We denote $\{M_t\}$ as the *estimated* number of flows deemed *admissible* by a MBAC scheme at time t, and $\{N_t\}$ as the *actual* number of flows in the system at time t. The interpretation of M_t is that the MBAC will continue admitting flows until N_t is greater than M_t. Because M_t is determined by past measurements, $\{M_t\}$ is a random process and so is $\{N_t\}$. Furthermore, $\mathcal{F}_t$ denotes the set of flows in the system at time time t. Obviously, $|\mathcal{F}_t| = N_t$.

Our design and analysis is based on the assumption of a large link in which many flows can be accommodated and no single flow dominates. The performance analysis is asymptotic in the link size c.

III. Aggregate versus Individual Flow Measurements

In [6], we have analyzed the impact of measurement errors for MBAC schemes which can measure the individual flow rates $\{X_i(\cdot)\}$. In this paper, we would like to design a scheme which only makes use of the past aggregate flow information, i.e. $\{S_t\}$. This section serves to quantify the performance loss associated with this coarser granularity of information. The insights gained here prepare us for the MBAC design in Section V, and are also interesting on their own right.

In the analysis of this section, we do not deal directly with flow arrivals and departures. We focus on the effect of measurement uncertainty on the number of admissible flows M_t, and then study the resulting impact on the QoS objective if M_t flows were admitted onto the link and remained in the system. A simple MBAC scheme is used as a vehicle for this purpose. Analysis of the complete model with flow dynamics will be done in Section VI after the full MBAC design is proposed in Section V.

Suppose the bandwidth processes of the flows are statistically independent and identical, and the stationary band-

[1] In practice, rough information such as the peak rate of the new flow is used as well. This can be incorporated in an obvious way in our proposed scheme.

width distribution of each flow has mean μ and variance σ^2. The capacity of the link is scaled as $c := n\mu$, where n can be thought of as the system size. When the system size n is large , the number of flows m in the system will be large, and by the Central Limit Theorem,

$$\frac{1}{\sqrt{m}}\left[\sum_{i=1}^{m} X_i(t) - m\mu\right] \sim N(0,\sigma^2)$$

irrespective of the statistics of the individual flows.

Consider then the following hypothetical admission control scheme with perfect knowledge of the parameters μ and σ^2 *a priori*: accept n^* flows with n^* satisfying:

$$Q\left[\frac{n\mu - n^*\mu}{\sigma\sqrt{n^*}}\right] = p_q. \tag{1}$$

where $Q(\cdot)$ is the complementary cdf of a $N(0,1)$ Gaussian random variable and p_q is the QoS objective[2]. For large capacities, it follows from solving (1) that

$$n^* = n - \frac{\sigma\alpha_q}{\mu}\sqrt{n} + o(\sqrt{n}) \tag{2}$$

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, $\frac{\sigma\alpha_q}{\mu}\sqrt{n}$ is the amount of bandwidth margin left to cater for the (known) burstiness.

This motivates the following *certainty-equivalent* MBAC scheme using aggregate flow information. Based on estimates $\hat{\mu}$ and $\hat{\sigma}$ of the statistics, it allows M_0 flows in the system at time 0, with M_0 satisfying:

$$Q\left[\frac{n\mu - M_0\hat{\mu}}{\hat{\sigma}\sqrt{M_0}}\right] = p_q, \tag{3}$$

where the estimates are given by:

$$\hat{\mu} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n}S_{t_k} \qquad \hat{\sigma}^2 = \frac{1}{K-1}\sum_{k=1}^{K}\frac{1}{n}\left(S_{t_k} - n\hat{\mu}\right)^2$$

and

$$S_{t_k} = \sum_{i=1}^{n} X_i(t_k)$$

is the aggregate load of flows in the system at time $t_k < 0$. [3]

Note that M_0 is now a random quantity, a function of K samples of the aggregate load ($K \geq 2$). We are interested in the distribution of M_0. For ease of analysis, let us assume that the sample times $\{t_k\}$ are spaced sufficient far apart such that the loads at distinct times are independent. For large n, by the Central Limit Theorem,

$$S_t = n\mu + Y_t\sqrt{n} + o(\sqrt{n}) \qquad t < 0$$

where $Y_t \sim N(0,\sigma^2)$. Hence, the mean and variance estimators are given by:

$$\hat{\mu} = \mu + \frac{1}{\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^{K} Y_{t_k}\right) + o(\frac{1}{\sqrt{n}}) \tag{4}$$

$$\hat{\sigma}^2 = \hat{\sigma}_K^2 + o(1) \tag{5}$$

where

$$\hat{\sigma}_K^2 = \frac{1}{K-1}\sum_{k=1}^{K}\left(Y_{t_k} - \frac{1}{K}\sum_{l=1}^{K} Y_{t_l}\right)^2.$$

For a fixed K, the variance estimate approaches $\hat{\sigma}_K^2$ for large system size n. Note however that this estimate remains random, unlike the mean estimate which approaches μ, the true mean.

The randomness in the estimators translates into the randomness in the number of flows admitted, via eqn. (3). By performing a linearization around the nominal perfect-knowledge operating point given by (1), it can be shown that (as in the proof of Prop. 3.1 in [6]):

$$M_0 = n - \frac{\sqrt{n}}{\mu}\left(\frac{1}{K}\sum_{k=1}^{K} Y_{t_k} + \alpha_q\hat{\sigma}_K\right) + o(\sqrt{n}). \tag{6}$$

More formally:

Proposition III.1: As $n \to \infty$, $\frac{M_0 - n}{\sqrt{n}}$ converges *in distribution* to the random variable

$$\frac{-1}{\mu}\left(\frac{1}{K}\sum_{k=1}^{K} Y_{t_k} + \alpha_q\hat{\sigma}_K\right) \tag{7}$$

It can be seen that the fluctuation in M_0 is due to both the randomness in the mean and variance estimators, when they are based only on aggregate loads. Contrast this with the case when individual flow measurements are available, when the uncertainty is due only to the measurement error in the mean bandwidth estimator [6]. In that case,

$$M_0 = n - \frac{\sqrt{n}}{\mu}\left(\frac{1}{K}\sum_{k=1}^{K} Y_{t_k} + \alpha_q\sigma\right) + o(\sqrt{n}). \tag{8}$$

Comparing eqn. (8) with (6), we see that the uncertainty in the standard deviation σ disappears with individual flow measurements. This is because individual flow measurements yield n samples per time instance for estimating the variance, while aggregate measurements yield only one. For large n, the effect of error in the variance estimator vanishes in the former case but not the latter.

It is also interesting to observe that M_0 is much more sensitive to errors in the mean estimator than in the variance estimator. This can be seen from eqn. (4), (5) and (7). The error in the mean estimator is magnified by a factor of $\sqrt{n}$, while the randomness in the variance estimator enters directly in (7). This is not very surprising, considering that the mean is a first-order statistic and the variance is second-order. Fortunately, the mean estimator is much more accurate than the variance estimator when only aggregate flow information is available, and this compensates exactly for the difference in order of magnitude of the sensitivities.

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

[3] Observe here that the estimation is based on n flows. In the actual model with flow dynamics, this should be the actual number of flows in the system which fluctuates around n. However, in a large system, this number will be close to n and the discrepancy in replacing it by n in the estimators are of a negligible effect.

We next investigate the effect of this randomness in the number of admitted flows M_0 on the QoS performance of the system. To this end, consider the aggregate load at some future time $t > 0$ after admitting M_0 flows and without future admissions. This is a sum of a random number of random variables, and using a version of Central Limit Theorem ([3, p. 369, problem 27.14], we get the following asymptotic approximation:

$$S_t := \sum_{i=1}^{M_0} X_i(t) = M_0\mu + Y_t\sqrt{n} + o(\sqrt{n}) \tag{9}$$

Here again $Y_t \sim N(0, \sigma^2)$ and can be interpreted as an approximation for the scaled aggregate bandwidth fluctuation at time t:

$$\frac{1}{\sqrt{n}}\left[\sum_{i=1}^{n} X_i(t) - n\mu\right] \tag{10}$$

Substituting eqn. (6), we get

$$S_t = n\mu + \left(Y_t - \frac{1}{K}\sum_{k=1}^{K} Y_{t_k} - \alpha_q\hat{\sigma}_K\right)\sqrt{n} + o(\sqrt{n}) \tag{11}$$

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

$$\Pr\{S_t > n\mu\} \approx \Pr\left\{\frac{1}{\hat{\sigma}_K}\left(Y_t - \frac{1}{K}\sum_{k=1}^{K} Y_{t_k}\right) > \alpha_q\right\} \tag{12}$$

Now since the Y_{t_k}'s are $N(0, \sigma^2)$, the random variables $\frac{1}{K}\sum_{k=1}^{K} Y_{t_k}$ and $\hat{\sigma}_K^2/\sigma^2$ can be interpreted as unbiased estimates of the mean and variance of a $N(0, \sigma^2)$ distribution based on K independent observations. As is well-known (see for example [2]), the two estimates are independent, and

$$\frac{K-1}{\sigma^2}\hat{\sigma}_K^2 \sim \chi_{K-1},$$

a Chi-square distribution with $K-1$ degrees of freedom. If we now make the further assumption that the time t is sufficiently large such that $X_i(t)$ (and therefore Y_t) is independent of $X_i(t_1), \ldots, X_i(t_K)$, then $Y_t - \frac{1}{K}\sum_{k=1}^{K} Y_{t_k}$ is independent of $\hat{\sigma}_K$ and is distributed as $N(0, \frac{K+1}{K}\sigma^2)$ and hence

$$\sqrt{\frac{K}{K+1}}\frac{1}{\hat{\sigma}_K}\left(Y_t - \frac{1}{K}\sum_{k=1}^{K} Y_{t_k}\right) \sim \mathcal{T}_{K-1}$$

where $\mathcal{T}_{K-1}$ is the student-t distribution with $K-1$ degrees of freedom [2].

We summarize this formally in the following.

Proposition III.2: Suppose the target overflow probability QoS is p_q. Then as the system size grows:

$$\lim_{n\to\infty} \Pr\{S_t > n\mu\} = F_{K-1}\left(\sqrt{\frac{K}{K+1}}Q^{-1}(p_q)\right), \tag{13}$$

where F_K is the complementary cdf of the $\mathcal{T}_{K-1}$ distribution.

Note that this limit does not depend on the true mean and variance, but only on the target QoS p_q.

It is interesting to compare with the corresponding result when individual flow measurements are available. A simple generalization of Proposition 3.3 in [6] says that with n independent individual flow measurements at each of K time instants, the asymptotic overflow probability is given by

$$Q\left(\sqrt{\frac{K}{K+1}}Q^{-1}(p_q)\right). \tag{14}$$

To appreciate the difference, it is instructive to examine the density of the $\mathcal{T}_{K-1}$ distribution:

$$f_{K-1}(x) = \frac{\Gamma\left(\frac{K}{2}\right)}{\sqrt{\pi(K-1)}\Gamma\left(\frac{K-1}{2}\right)}\left(1 + \frac{x^2}{K-1}\right)^{-\frac{K}{2}} \tag{15}$$

where $\Gamma(\cdot)$ is the Gamma function. For small K, this distribution has a slow (polynomially) decaying tail as compared to the doubly exponentially decaying tail of the Gaussian distribution. Thus, for small K, the target overflow probability is missed significantly more in the case when only aggregate measurements are available; see Fig. 2. For $K = 5$, the actual overflow probability p_f is very far away from p_q and decreases very slowly with the latter (the upper curve), while p_f is quite close to the target with individual flow measurements. As expected, as $K \to \infty$, the performance approaches p_q under both aggregate and individual flow measurements.

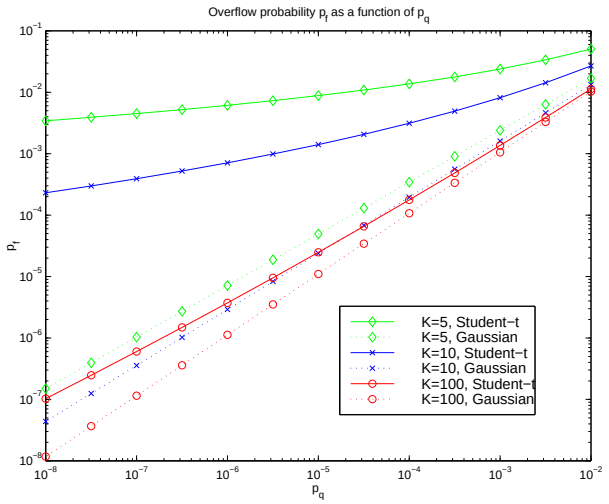

Fig. 2: The overflow probability p_f as a function of the target overflow probability p_q, for various K (Student-t corresponds to aggregate measurements according to (13), Gaussian to per-flow measurements according to (14)).

The significant degradation observed above for small K under aggregate load measurements can be attributed to errors in estimation of the *variance.* With non-negligible probability, the variance can be significantly under-estimated. In that case, the certainty-equivalent admission control scheme will be very aggressive in accepting flows, reserve very little bandwidth margin to cater for the burstiness. This results in high overflow probability when the flows are actually admitted.

IV. The Critical Time-Scale $\widetilde{T_h}$

The goal of this section is to identify the *critical time-scale* $\widetilde{T_h}$ discussed in the introduction. This notion is the

cornerstone of our time-scale decomposition approach: aggregate bandwidth fluctuation slower than $\widetilde{T_h}$ is tracked and compensated for by flow admissions and departure; bandwidth fluctuation faster than $\widetilde{T_h}$ are absorbed by allocating spare bandwidth in the link.

Suppose for the moment that the number of flows in the system is fixed at n, and call the aggregate bandwidth of these n flows S_t^n. Also assume that flows are independent, identically distributed random processes with mean μ and variance σ^2, which share a link of capacity $c = n\mu$. By flow independence, the fluctuation of S_t^n around $n\mu$ is on the order of $\sigma\sqrt{n}$. In the *tracking* regime, we want to *compensate* for this fluctuation by controlling the number of flows N_t over time. In other words, when S_t^n happens to be larger than $n\mu$, i.e., exceeding the link capacity, we want to lower the number of flows $N_t < n$ such that the aggregate bandwidth of N_t flows does not exceed the link capacity. Let S_t denote the aggregate bandwidth of these N_t flows. Then the fluctuation of S_t around its mean has two components, one due to the bandwidth fluctuation, and one due to the fluctuation of the number of flows in the system:

$$S_t = S_t^n - (n - N_t)\mu + o(\sqrt{n}) = n\mu + \sqrt{n}\sigma W_t - (n - N_t)\mu + o(\sqrt{n}) \quad (16)$$

where $\{W_t\}$ is a zero mean, unit variance Gaussian process with autocorrelation function $e^{-t/T_h}\rho(t)$, where $\rho(t)$ is the autocorrelation function of an individual flow. The factor e^{-t/T_h} is because the set of flows $\mathcal{F}_t$ present in the system changes over a time-scale of T_h, and the bandwidth of two different flows is independent.

Because flows cannot be preempted from the system once admitted, the number of flows can only be lowered by letting flows depart from the system. The rate at which flows depart from the system in turn is approximately n/T_h, where T_h is the average flow holding time. This corresponds to a "bandwidth departure rate" of $n\mu/T_h$. An increase in S_t^n can therefore be compensated by flow departures only if the rate of change of S_t^n does not exceed $n\mu/T_h$.

First, assume that the aggregate bandwidth S_t^n fluctuates over a single time-scale T_c[4]. It is therefore unlikely that the rate of change of S_t^n exceed $O(\sigma\sqrt{n}/T_c)$. As a result, the tracking regime is possible whenever $\sigma\sqrt{n}/T_c \ll n\mu/T_h$, or $T_c \gg \frac{\sigma}{\mu}\frac{T_h}{\sqrt{n}}$. We therefore identify

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

as the *critical time-scale* of the system. Thus, in the case when $T_c \gg \widetilde{T_h}$, the aggregate bandwidth fluctuation is completely compensated for by flow admissions and departures, resulting in a near full utilization of the link. See the first column of Fig. 3.

On the other hand, in the *overbooking regime* where $T_c \ll \widetilde{T_h}$, $S_t \approx n^*\mu + \sqrt{n}\sigma W_t^H$. In this case, the amount of spare bandwidth is given by $\mu(n - \mathrm{E}[N_t]) = \mu(n - n^*)$. See the second column of Fig. 3. In this regime, full link utilization cannot be achieved.

More generally, aggregate bandwidth fluctuates over multiple time-scales. The components having time-scale $T_c^L \gg T_h/\sqrt{n}$ can be compensated for through flow admissions and departures, while the components having time-scale $T_c^H \ll T_h/\sqrt{n}$ have to be absorbed through overbooking. See the last column of Fig. 3.

That the critical time-scale $\widetilde{T_h}$ is proportional to the average flow duration T_h is not surprising. What is more subtle is the scaling of $\widetilde{T_h}$ with $1/\sqrt{n}$. The reason for this is that the aggregate flow departure rate grows linearly with n, while the fluctuations grow only like $\sqrt{n}$. As a result, as the system scales, there are more fluctuations can be compensated for by flow departures, manifesting in a short critical time-scale.

Although the discussion here is informal, the main point is to motivate the MBAC design to be presented in the next section. The importance of the critical time-scale will be demonstrated more precisely in the performance analysis of the proposed MBAC (Section VI).

V. The MBAC Design

A. Basic Architecture

Our proposed MBAC design derives directly from the observation in the previous section that fluctuations on a time scale slower than $\widetilde{T_h}$ should be absorbed by tracking, and fluctuations on a time-scale faster than $\widetilde{T_h}$ by overbooking. This suggests decomposing the aggregate bandwidth process S_t into a high-frequency component S_t^H and a low-frequency component S_t^L such that $S_t = S_t^H + S_t^L$, both with a cutoff frequency of $1/\widetilde{T_h}$. We can obtain such a decomposition through a low-pass and a high-pass filter, both with cutoff frequency $1/\widetilde{T_h}$ (cf. Fig. 4).

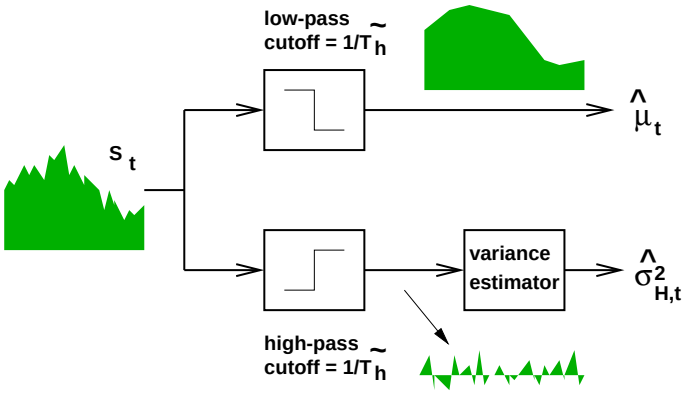

Fig. 4: The decomposition of the measured aggregate bandwidth into a high-frequency component for the variance estimator and a low-frequency component for the mean estimator.

Then the high-frequency process S_t^H is used in order to estimate the amount of spare bandwidth that has to be put aside in order to accommodate fast time-scale fluctuations through overbooking. Hence, we wish to estimate the variance σ_H^2 of S_t^H. The low-frequency process S_t^L is used to estimate the "current mean" $\widehat{\mu}_t$ of the flows. This determines the current number of flows that should be in the system in order to accommodate the slow time-scale fluctuations through tracking.

B. Variance Estimator

How should we estimate the variance σ_H^2 of the high-frequency component of the aggregate traffic? Recall now the main insight we gained from Section III:

- With only aggregate measurements, the performance of a MBAC can be quite poor if there are only a small

[4] Informally, this means that the power of the process $\{W_t\}$ is concentrated around $1/T_c$ in its power spectral density.

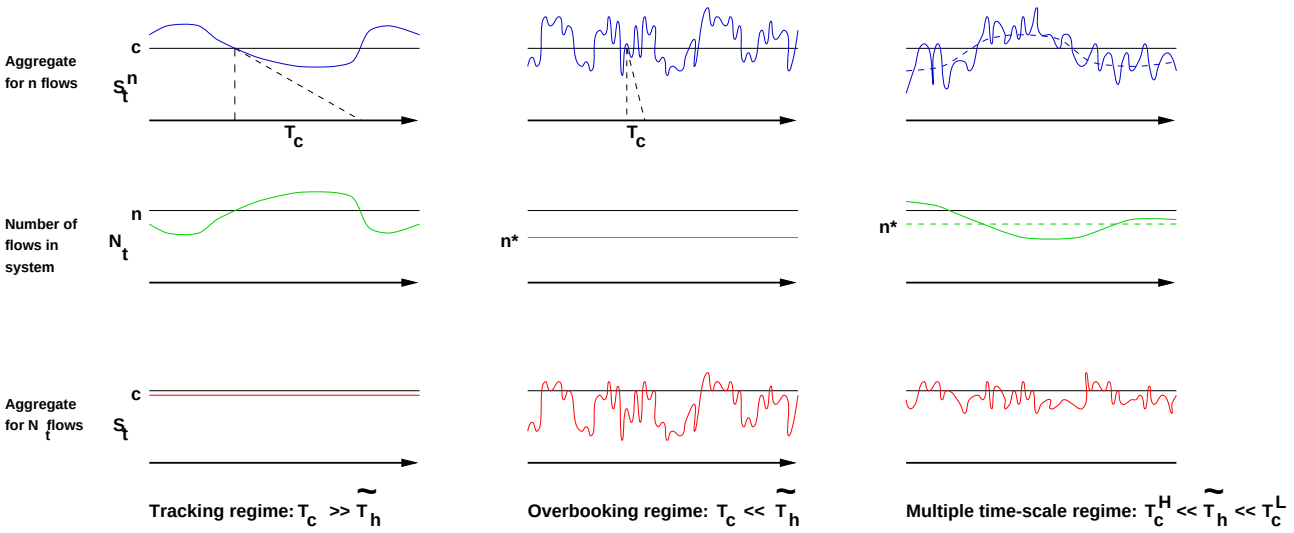

Fig. 3: The tracking and the overbooking regime. In the tracking regime, bandwidth fluctuation is absorbed by a corresponding fluctuation of the number of flows in the system; in the overbooking regime, bandwidth fluctuation is absorbed by overbooking resources, i.e., setting spare bandwidth aside to accommodate the fluctuation of the aggregate load.

number K of independent load measurements. Either the target is missed significantly, or a very conservative admission control scheme is needed to compensate for the measurement errors. This effect is mainly due to estimation error in the *variance*.

This suggests that a long measurement window for estimating the variance σ_H^2 is needed for robust performance and high link utilization. Essentially, we need more measurements *over time* to make up for the lack of measurements *over individual flows*. Since the fast fluctuations by definition occur at time-scale $\widetilde{T_h}$ or shorter, one can expect to get roughly independent measurements of $\widetilde{T_h}$ spaced at $\widetilde{T_h}$ apart. The above observation thus translates into the need of a measurement window with length much larger than $\widetilde{T_h}$. Since $\widetilde{T_h} = \frac{T_h}{\sqrt{n}}$, a natural choice of the measurement window length with the desired scaling property is T_h, the average holding time of a flow.

With this choice of measurement window size, a natural question is the robustness to non-stationarities, especially due to heterogeneity of flows entering and leaving the network. We will address this issue when we analyze the performance of the MBAC design under a heterogeneous traffic model.

C. A Specific MBAC Scheme

We have identified the basic architecture of the MBAC design: a low-pass filter at cutoff $\frac{1}{\widetilde{T_h}}$ to track the slow time-scale fluctuation of the aggregate traffic, and a measurement filter of memory length T_h to estimate the variance of the fast time-scale fluctuations. To analyze performance, we have to fix specific filters. So consider a low-pass filter with impulse response

$$g_t := \frac{1}{\widetilde{T_h}} \exp(-\frac{t}{\widetilde{T_h}}) u_t \tag{17}$$

where u_t is the unit step function. Let

$$h_t := \frac{1}{T_h} \exp(-\frac{t}{T_h})$$

be the filter for estimating the variance. If S_t is the aggregate load at time t, the estimated (slow time-scale) mean is then :

$$\hat{\mu}_t = \int_0^\infty \frac{S_{t-\tau}}{N_{t-\tau}} g_\tau d\tau, \tag{18}$$

where N_t is the number of flows in the system at time t. The high-pass component of the aggregate load is:

$$S_t^H := S_t - \int_0^\infty S_{t-\tau} h_\tau d\tau$$

and the estimate of the high-pass variance is given by:

$$\hat{\sigma}_t^H = \int_0^\infty \frac{1}{N_{t-\tau}} \left[S_{t-\tau}^H - \int_0^\infty S_{t-u}^H h_u du \right]^2 h_\tau d\tau. \tag{19}$$

The number of flows M_t admissible by the MBAC in time t is given by

$$Q\left(\frac{c - M_t \hat{\mu}_t}{\sqrt{M_t \hat{\sigma}_t^H}} \right) = p_q. \tag{20}$$

VI. Performance of MBAC Scheme

We now sketch the performance analysis of the MBAC scheme proposed above in a fully dynamical model with flow arrival and departures. 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. Due to space limitations, the details of the analysis are described in the journal version of this paper [7].

A. Homogeneous Flows

We first consider the homogeneous case when the bandwidth process $\{X_i(\cdot)\}$ of each flow is identically distributed, stationary and ergodic. The mean rate of each flow is μ and the covariance function is $\rho(t) := \mathrm{E}\left[(X_i(0) - \mu)(X_i(t) - \mu)\right]$. The capacity c is scaled as $n\mu$.

Recall that M_t is the number of flows the MBAC determines *should* be admissible in the link at time t. We can analyze the distribution of this process in a similar way as in Section III using Central Limit Theorem, and obtain:

$$M_t = n - \frac{\sqrt{n}}{\mu}\left(Z_t + \alpha_q \hat{\sigma}_t^H\right) + o(\sqrt{n}). \tag{21}$$

where $Z_t = (g * Y)_t$ ($*$ represents the convolution operation), and $\{Y_t\}$ is a zero-mean Gaussian process with covariance function $\rho(t)$, representing the (scaled) fluctuation of the aggregate bandwidth. The process $\{Z_t\}$ is the low-pass version of $\{Y_t\}$. The number of admissible flows at time t is a random quantity with fluctuation of order $\sqrt{n}$ due to the randomness in the statistical estimators $\hat{\mu}_t$ and $\hat{\sigma}_t^H$. The term $-\sqrt{n}Z_t$ represents the tracking of the slow time-scale fluctuations by the MBAC; the term $-\sqrt{n}\alpha_q\hat{\sigma}_t^H$ represents the spare bandwidth catered for the fast time-scale fluctuations.

It can be shown, as in [6], that the number of flows N_t *actually* in the system is given by:

$$N_t = \sup_{s \le t}\{M_s - D[s,t]\} \tag{22}$$

This relationship quantifies precisely how much control the admission scheme has on the number of flows in the system. At time t, the ideal number of flows desired in the system is M_t. But N_t is close to M_t only if the flow departure rate is very high. For finite departure rates, N_t exceeds M_t and to still provide the desirable level of QoS, spare bandwidth has to be allocated in the admission scheme.

We now scale up the system by letting $n \to \infty$ with the critical time-scale $\widetilde{T_h}$ fixed, such that the average flow holding time scales as $T_h = \sqrt{n}\widetilde{T_h}$. Under this scaling, the number of flows departed in $[s,t]$ can be calculated to be:

$$D[s,t] = \frac{t-s}{\widetilde{T_h}}\sqrt{n} + o(\sqrt{n}). \tag{23}$$

Also, the variance estimate $\hat{\sigma}_t^H$ can be shown to converge:

Proposition VI.1: As $n \to \infty$ and with the flow holding time scaling as $T_h = \sqrt{n}\widetilde{T_h}$, $\{\hat{\sigma}_t^H\}$ converges in distribution to a constant σ_H, where

$$\sigma_H^2 = \mathrm{Var}\left[X_i(0) - (g * X_i)(0)\right]$$

is the variance of the high-frequency component of a flow bandwidth process.

The result can be understood intuitively as follows. The high-frequency component has fluctuations at time-scale $\widetilde{T_h}$ or shorter, so roughly samples spaced at $\widetilde{T_h}$ apart are independent. The measurement window for the variance estimator is of time-scale $T_h = \sqrt{n}\widetilde{T_h}$. For large n, the estimate of the power in the high-frequency component will be very accurate. This is analogous to taking a large number K of independent measurements of the aggregate load in the simple model studied in Section III.

Using now equations (21), (22), (23), Proposition VI.1 and similar argument as in Section III, we can obtain the asymptotic distribution of the aggegrate load as $n \to \infty$:

$$S_t = n\mu + \sqrt{n}\sup_{s \le t}\left\{Y_t - Z_s - \frac{\mu}{\widetilde{T_h}}(t-s) - \alpha_q\sigma_H\right\} + o(\sqrt{n}) \tag{24}$$

and the corresponding overflow probability p_f converges to:

$$\Pr\left\{\sup_{s \le 0}\left\{Y_0 - Z_s + \frac{\mu}{\widetilde{T_h}}s\right\} > \alpha_q\sigma_H\right\}. \tag{25}$$

The expression (25) can be interpreted as a *hitting probability* of a Gaussian process $(\{Y_0 - Z_s\})$ on a moving boundary, and an approximation of such a probability can be obtained, given the covariance function $\rho(t)$ [6].

Let us consider two specific examples to obtain a better intuitive understanding of these general results.

1) **Single Time-Scale Traffic:** Suppose now the individual flow has covariance function

$$\rho(t) = \sigma^2 \exp(-\frac{|t|}{T_c})$$

with correlation at a single time-scale T_c. Consider first the regime when $T_c \ll \widetilde{T_h}$; this can be considered as a separation between the burst and flow time-scales. In this case, the variance of the high-pass component σ_H is the same as the variance σ of an individual flow, and the overflow probability is given by:

$$\begin{aligned} & \Pr\left\{\sup_{s \le 0}\left\{Y_0 + \frac{\mu}{\widetilde{T_h}}s\right\} > \alpha_q\sigma_H\right\} \\ = \;& \Pr\left\{Y_0 > \alpha_q\sigma\right\} = p_q. \end{aligned}$$

Thus the target QoS is met using our scheme. In this case, the traffic fluctuations are all of a faster time-scale than $\widetilde{T_h}$ and resources has to be overbooked to absorb them. If we overbook by any amount less than the full variance σ^2 of the fluctuation, the QoS target would not have been met.

For general T_c, we numerically compute the overflow probability using the formula in Section 4.3 of [6]. This is plotted in Fig. 5 for two values of p_q. We see that the actual overflow probability p_f is close to the target p_q across the whole range of T_c. As T_c increases beyond the critical time-scale $\widetilde{T_h}$, the spare bandwidth reserved to absorb the high-frequency burstiness is reduced accordingly, thus maximizing utilization while still meeting the target QoS. Contrast this with the performance of the per-flow scheme considered in [6], which always reserves spare bandwidth equal to σ, the total variance. When T_c is of the order or larger than $\widetilde{T_h}$, this results in over-allocation of resources, as seen in the drop in p_f.

2) **Multiple Time-Scale Traffic:** Let us now consider the situation when an individual flow has correlation at two time-scales T_f and T_s, with covariance function:

$$\rho(t) = \sigma_f^2 \exp\left(-\frac{|t|}{T_f}\right) + \sigma_s^2 \exp\left(-\frac{|t|}{T_s}\right).$$

For any T_f, T_s, the overflow probability can again be computed numerically as before, but it is perhaps more insightful to look at the scenario where $T_f \ll \widetilde{T_h}$ and $T_s \gg \widetilde{T_h}$. Using eqn. (25) and (24), it can be seen that the overflow probability meets the target QoS p_q, while the average utilization $E[S_t] \approx n\mu - \sigma_f\alpha_q\sqrt{n}$. In this scenario, the admission controller tracks the slow time-scale (T_s) traffic fluctuation perfectly, and leaves a spare bandwidth of $\sigma_f\alpha_q\sqrt{n}$ to absorb the fast time-scale fluctuation (T_f). The

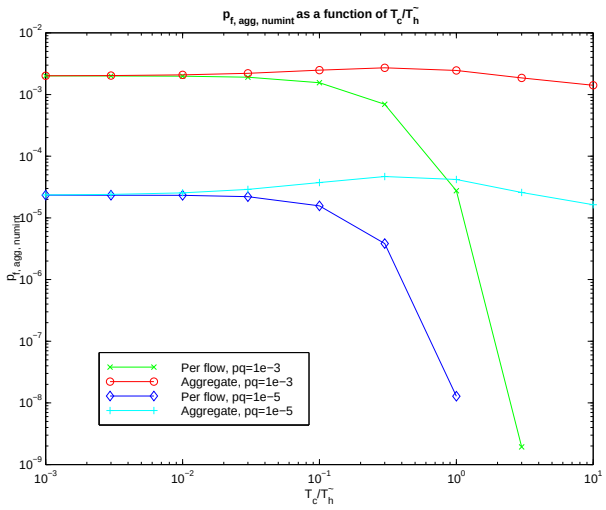

Fig. 5: Overflow probability of the proposed aggregate scheme and a per-flow scheme which always overbooks at σ.

choice of $\widetilde{T_h}$ as the memory time-scale of the low-pass filter is important to keep the utilization high. For if the memory time-scale T_m of the low-pass filter is chosen to be larger than $\widetilde{T_h}$ and close to T_s, some of the slow fluctuations is filtered into the high-frequency component, resulting in a larger than necessary spare bandwidth. In the extreme case when $T_m \gg T_s$, a spare bandwidth of $\sqrt{\sigma_s^2+\sigma_f^2}\alpha_q\sqrt{n}$ is kept, resulting in an over-conservative scheme.

B. Heterogeneous Flows

Consider the following heterogeneous traffic model: the ith flow is given by

$$X_i(t) = \mu_i + \sigma_i U_i(t),$$

where μ_i and σ_i are random variables, identically distributed and independent from flow to flow. The processes $\{U_i(\cdot)\}$ are independent, identically distributed with zero mean and unit variance, and are stationary and ergodic with covariance function $\rho_U(t)$; they are also independent of μ_i's and σ_i's. They represent the in-flow statistical fluctuations. The random variables μ_i and σ_i^2 represent the long-term mean and variance of the flow; they differ from flow to flow but remain fixed once the flow is in progress. The processes $\{U_i(\cdot)\}$ represent the in-flow statistical fluctuations which we model as statistically identical and independent of μ_i's and σ_i's for simplicity. The random variables μ_i and σ_i^2 has the following statistics:

$$\mathrm{E}\left[\mu_i\right] = \mu, \qquad \mathrm{Var}\left[\mu_i\right] = v^2, \qquad \mathrm{E}\left[\sigma_i^2\right] = \sigma^2.$$

One can think of the distribution of (μ_i, σ_i^2) as modeling the typical flow mix. At any time, the composition of flows in the network may deviate from this typical mix.

The aggregate load in the system is given by

$$S_t = N_t\mu + \sum_i(\mu_i - \mu) + \sqrt{n}V_t + o(\sqrt{n}), \qquad (26)$$

where $\{V_t\}$ is zero-mean Gaussian process with covariance function the same as the process $\{\sigma_i U_i(\cdot)\}$. We decompose the load into three terms: 1) $N_t\mu$, which can be thought of as the aggregate load if all flows are transmitting at their average rate μ_i and the flow mix is exactly the same as the typical mix; 2) $\sum_i(\mu_i-\mu)$, where the sum is over the flows currently in the system, is the deviation of the current mix of the flows from the typical mix; 3) $\sqrt{n}V_t$, which is the fluctuation of the flows from their long-term average rates.

Using the Central Limit Theorem for a random number of summands, we can approximate the second term by $\sqrt{n}L_{t/T_h}$, where $\{L_t\}$ is a zero mean Gaussian process with covariance function

$$\rho_L(t) = v^2\exp(-t),$$

as a consequence of the flow departure process. This is the slow time-scale fluctuation in the aggregate load due to the change in flow mix over time. The scaling by T_h emphasizes the fact that this process is evolving at the time-scale of the flow arrivals and departures.

The covariance function of the Gaussian process $\{V_t\}$ is given by:

$$\rho_V(t) = \mathrm{E}\left[\sigma_i^2 U_i(0)U_i(t)\right] = \sigma^2\rho_U(t).$$

We note that the time fluctuation in the *flow variances* due to heterogeneity has disappeared in the approximation (26) of the aggregate load; only the typical variance σ^2 matters. On the other hand, the fluctuation in the mean rates $\{\sqrt{n}L_{t/T_h}\}$ remains. The reason is that the aggregate load is much more sensitive to the mean fluctuation, a first-order effect, than variance fluctuation, a second order effect. We have in fact seen this phenomenon in Section III, where we performed a measurement error analysis.

Continuing on the performance analysis, the low-pass mean estimator is given by (via eqn. (18)):

$$\hat{\mu}_t = \mu + \frac{1}{\sqrt{n}}L_{t/T_h} + \frac{1}{\sqrt{n}}Z_t + o(\frac{1}{\sqrt{n}})$$

where $Z_t = (g * V)_t$. We note that the filter can track the slow time-scale fluctuation $\{L_{t/T_h}\}$ perfectly; this is because the filter has a much shorter time-scale $\widetilde{T_h}$ than $T_h = \sqrt{n}\widetilde{T_h}$.

The number of admissible flows is given by:

$$M_t = n - \frac{\sqrt{n}}{\mu}\left(L_{t/T_h} + Z_t + \alpha_q\hat{\sigma}_t^H\right) + o(\sqrt{n}) \qquad (27)$$

where $\hat{\sigma}_t^H$ is the high-pass variance estimator given by eqn. (19). We have the following convergence result.

Proposition VI.2: As $n\to\infty$ and with the flow holding time scaling as $T_h = \sqrt{n}\widetilde{T_h}$, $\{\hat{\sigma}_t^H\}$ converges in distribution to a constant σ_H, where

$$\sigma_H^2 = \mathrm{Var}\left[V_0 - (g*V)_0\right] = \sigma^2\mathrm{Var}\left[U(0) - (g*U)(0)\right].$$

Although this proposition is identical to the corresponding one (Prop. VI.1) for the homogeneous case, the reason why it is true is more subtle. Recall that the memory time-scale for the high-pass variance estimator is T_h. Hence, the heterogeneous mix of flows actually changes significantly during this time. However, the low sensitivity of the aggregate load to the fluctuation of the variances ensures that the variance estimator remains accurate.

Combining eqns. (26), (27) and Prop. VI.2, the aggregate load and the overflow probability can be computed to

be:

$$S_t = n\mu + \sqrt{n}\sup_{s\le t}\left\{V_t - Z_s - \frac{\mu(t-s)}{\widetilde{T_h}} - \alpha_q\sigma_H\right\} + o(\sqrt{n})$$

and

$$\Pr\{S_0 > n\mu\} \approx \Pr\left\{\sup_{s\le 0}\left\{V_0 - Z_s + \frac{\mu s}{\widetilde{T_h}}\right\} > \alpha_q\sigma_H\right\}$$

Comparing these results with (24), we observe that the (asymptotic) overflow probability and the utilization for the heterogeneous model are the same as those for a *homogeneous* model where each flow has the same mean rate μ and the same variance σ. There are two reasons for this. First, the process $\{L_{t/T_h}\}$ describing the change of the mean rates of the flow mix in the system is completely filtered into the low-frequency component and perfectly compensated for by admission control. Second, the fluctuation due to change in flow variances σ_i has an insignificant impact on the aggregate load and the overflow probability. This ensures that although the memory time-scale for estimating the high-pass variance is of the order of T_h, the estimates will not be significantly corrupted by outdated data.

The above performance analysis of the proposed scheme under a heterogeneous traffic model gives further evidence to the efficiency and robustness of the design, particularly in the choice of $\widetilde{T_h}$ as the filter time-scale for tracking the low-pass mean and T_h as the memory time-scale for estimating the high-pass variance. For example, if the low-pass filter time-scale were chosen to be of the order of T_h and not $\widetilde{T_h}$, then unnecessary spare bandwidth will have to be reserved for the slow time-scale fluctuations due to flow heterogeneity. In the extreme case when the filter time-scale is much larger than T_h, an excess bandwidth proportional to v, the standard deviation of μ in the flow mix, is needed. This corresponds to the case when very conservative admission control is performed, solely based on prior knowledge of flow statistics and without benefiting from the on-line measurements.

VII. Conclusion

Admission control schemes generally make a time-scale separation assumption between the burst time-scale and the flow arrival and departure time-scale. Under this assumption, admission control only relies on burst time-scale statistics, such as the effective bandwidth, to make admission decisions. The flow lifetime does not enter the picture, which considerably simplifies the analysis of the system, as one faces essentially a static situation: based on the burst statistics, the goal is to compute the admissible number of flows such that the QoS target is met.

By taking flow arrival and departure dynamics into account, we have shown the fundamental importance of the critical time-scale. Fluctuations on a shorter and on a longer time-scale with respect to this critical time-scale can be controlled in different ways. Fast time-scale fluctuations have to be absorbed by the overbooking regime, i.e., setting sufficient spare bandwidth aside to accommodate them. Slow time-scale fluctuations can be compensated for by tracking, i.e., adjusting the number of flows in the system through flow arrivals and departures. This decomposition has led to a specific MBAC approach based on a time-scale decomposition of bandwidth fluctuations through low and high-pass filtering.

We have then evaluated the performance of our MBAC approach for both homogeneous and heterogeneous flow models, under the assumption that the window length for the variance estimation is on the order of the average flow holding time. This analysis allows us to account for the measurement uncertainty in the mean estimator, and to compute a corrected QoS target p'_q such that the overflow probability is close to the target.

Our MBAC scheme based on aggregate time-scale decomposition fulfills all of the requirements stated in the introduction. First, it is robust with respect to flow heterogeneity as well as bandwidth fluctuations on multiple time-scales. Also, we have explicitly quantified the correction to be applied to the QoS target p_q in order to compensate for measurement uncertainty, rather than relying on unspecified external parameters that need to be adjusted. Second, it achieves high resource utilization. In fact, exploiting the tracking regime allows us to maximize resource utilization, as we do not set aside spare bandwidth unnecessarily for slow time-scale fluctuations. This has been illustrated in Figure 5, where it can be seen that the overflow probability remains close to the target overflow probability p_q when T_c grows large, which means that it does not unnecessarily sacrifice bandwidth for fluctuations which fall into the tracking regime. Third, our scheme is amenable to an efficient implementation. Aggregate bandwidth information can be obtained in a switch or router simply by counting packets, without inspecting packet headers and without maintaining per-flow information.

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. J. Bickel and K. A. Doksum. *Mathematical Statistics:Basic Ideas and Selected Topics*. Holden-Day, 1977.

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

[4] 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.

[5] R. J. Gibbens, F. P. Kelly, and P. B. Key. A Decision-theoretic Approach to Call Admission Control in ATM Networks. *IEEE Journal on Selected Areas of Communications*, pages 1101–1114, August 1995.

[6] M. Grossglauser and D. Tse. A Framework for Robust Measurement-Based Admission Control. In *Proc. ACM SIGCOMM '97*, Cannes, France, September 1997.

[7] M. Grossglauser and D. N. C. Tse. A Time-Scale Decomposition Approach to Measurement-Based Admission Control. *submitted for publication*, January 1999.

[8] S. Jamin, P. B. Danzig, S. Shenker, and L. Zhang. A Measurement-Based Admission Control Algorithm for Integrated Services Packet Networks. *IEEE/ACM Transactions on Networking*, 5(1), February 1997.

[9] 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.

[10] V. Paxson and S. Floyd. Wide Area Traffic: The Failure of Poisson Modeling. *IEEE/ACM Trans. on Networking*, 3(3):226–244, June 1995.