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

Matthias Grossglauser, *Member, IEEE,* and David N. C. Tse, *Member, IEEE*

***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; and 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. An 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.**

***Index Terms*—Admission control, measurement, resource allocation, time scales.**

I. Introduction

IN ORDER TO make quality-of-service (QoS) guarantees, a network must exercise flow admission control. Admission decisions are based on some traffic characterization, such as effective bandwidths [7], [15] or leaky bucket descriptors [18]. The traditional approach to admission control assumes that a traffic descriptor is provided by the user or application for each flow prior its establishment [19]. However, this approach suffers from several problems. Chief among them is the inability of the user or application to come up with tight traffic descriptors *before* establishing the flow. This is especially so when the bandwidth fluctuates over multiple time scales. Another problem is that this traffic descriptor and the associated QoS guarantee defines a *contract* between the application and the network and, therefore, a need to police this traffic specification. This is difficult for statistical traffic descriptors. Also, the need for a policer makes the network architecturally more complex.

Measurement-Based Admission Control (MBAC) avoids these problems by shifting the task of traffic specification from the application to the network [9], [11], [14]. Instead of the application explicitly specifying the traffic, the network attempts to "learn" the statistics of existing flows by making on-line measurements. 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 large numbers. It is thus easier to predict aggregate behavior rather than the behavior of an individual flow.

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 QoS 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 QoS should also be robust to flow heterogeneity, to the fluctuations on many time scales that are a general property of network traffic [1], [6], [16], [17], as well as to very heavy offered loads, e.g., due to "flash crowds."
- **Resource utilization:** The QoS of admitted flows could be improved by being overly conservative in admission control, thereby allocating more resources per flow than necessary. This is undesirable, because the secondary goal for the MBAC is to maximize link utilization, subject to the QoS constraint for the admitted flows.[1]
- **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.

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

Manuscript received August 28, 2000; revised April 4, 2001 and October 23, 2002; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor G. De Veciana. This work was supported in part by the National Science Foundation under Grant ANI-9814567.

M. Grossglauser was with AT&T Labs—Research, Florham Park, NJ 07932 USA. He is now with the Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland (e-mail: matthias.grossglauser@ica.epfl.ch).

D. N. C. Tse is with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720 USA (e-mail: dtse@eecs.berkeley.edu).

Digital Object Identifier 10.1109/TNET.2003.815289

[1]It is important to note that we define QoS as the performance experienced by admitted flows; we do not view link utilization as a QoS metric *per se*. The goal of the MBAC is to admit as many flows as possible, subject to satisfying the QoS constraints.

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 simply by departing flows to avoid resource 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 [11], the main issue we addressed was 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 the mathematical machinery developed in [11] serves 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. Therefore, we seek a clear understanding of the impact of errors associated with aggregate measurements.

Second, we consider the situation when flows are *heterogeneous*. Flows may represent many different types of media (e.g., audio or video), they may be encoded at different levels of quality, and they may 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 in the presence of heterogeneous flows, even without any *a priori* classfication of flows.

Most MBAC schemes that have been proposed in the literature—including ours—are capable of a broad range of operating points in terms of link utilization and quality of service. In fact, it has recently been noted that most MBACs are essentially equivalent in terms of the set of operating points that they admit, given identical traffic characteristics [5]. However, MBACs differ significantly in their ability to achieve a desired QoS robustly, i.e., with little a-priori knowledge of traffic characteristics and without excessive tuning. A perfectly robust MBAC would have the QoS target itself as the only parameter and would require no tuning at all, because the actual QoS would be equal to this target *regardless of the traffic characteristics*. In practice, it is not possible to completely decouple performance from traffic characteristics, and all MBACs possess additional tuning parameters. Tuning essentially amounts to searching in a possibly multidimensional parameter space. Therefore, the main benefit of an analytic model such as the one discussed in this paper is that it replaces this search with an explicit relationship between traffic characteristics and tuning parameters.

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 based 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, and provide some simulation results. Section VII discusses how the MBAC scheme can be modified for a distributed implementation within the framework of diffserv. Section VIII contains the 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 flow i has been admitted, its bandwidth requirement $\{X_i(\cdot)\}$ 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; an MBAC scheme makes decisions based solely on observation of the past traffic flows.[2] 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 $E[S_t]$ of the link.

Several processes are of importance in this paper. We denote $\{M_t\}$ as the *estimated* number of flows deemed *admissible* by an MBAC scheme at time t, and $\{N_t\}$ as the *actual* number

[2]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.

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 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 [11], we 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 focuses on a simplified model 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.

The analysis in this section does not deal directly with flow arrivals and departures. We consider only the simple case of flows with homogeneous statistics. We focus on the effect of past measurement uncertainty on the number of admissible flows M_0 at time 0, and then study the resulting impact on the QoS objective at a future time t if M_0 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 and heterogeneous flows will be done in Section VI after the full MBAC design is proposed in Section V. This present section can be viewed as a parallel to [10, sec. II].

Suppose the bandwidth processes of the flows are statistically independent and identical, and the stationary bandwidth distribution of each flow has mean μ and variance σ^2. The capacity of the link is c. If we let $n := c/\mu$, then 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 the equation

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

where $Q(\cdot)$ is the complementary cumulative distribution function (CDF) of a $N(0,\, 1)$ Gaussian random variable and p_q is the QoS objective.[3] For large capacities, it follows from solving (1) and substituting $n = c/\mu$ that

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

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, $(\sigma\alpha_q/\mu)\sqrt{n}$ is the (normalized) amount of spare bandwidth left to cater for the (known) burstiness. We also observe that the number of flows admitted is deterministic in this perfect knowledge scenario.

The above scheme motivates the following *certainty-equivalent* MBAC, when the statistics of the flows are not known *a priori* but can only be estimated from aggregate flow information. Based on estimates $\hat{\mu}$ and $\hat{\sigma}^2$ of the mean and variance, the MBAC scheme 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

$$\hat{\mu} := \frac{1}{K}\sum_{k=1}^{K}\frac{S^n_{t_k}}{n}, \quad \hat{\sigma}^2 := \frac{1}{K-1}\sum_{k=1}^{K}\frac{\left(S^n_{t_k} - n\hat{\mu}\right)^2}{n} \tag{4}$$

and

$$S^n_{t_k} := \sum_{i=1}^{n} X_i(t_k)$$

is the aggregate load of flows in the system at time $t_k < 0$.[4]

The estimates $\hat{\mu}$ and $\hat{\sigma}^2$ are obtained by averaging over K samples of the aggregate load ($K \geq 2$). Note that M_0 is now a random quantity, being a function of the estimates $\hat{\mu}$ and $\hat{\sigma}^2$. We are interested in the distribution of M_0 for large n but fixed K. For ease of analysis, let us assume that the sample times $\{t_k\}$ are spaced sufficiently far apart such that the loads at distinct times are independent. For large n, by the Central Limit Theorem

$$S^n_t = n\mu + Y_t\sqrt{n} + o\left(\sqrt{n}\right) \qquad t < 0 \tag{5}$$

where $Y_t \sim N(0,\, \sigma^2)$.[5]

Substituting this into (4) yields the following expressions for the mean and variance estimators:

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

$$\hat{\sigma}^2 = \hat{\sigma}^2_K + o(1) \tag{7}$$

where

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

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

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

[4] 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.

[5] The Central Limit Theorem states that $(S^n_t - n\mu)/\sqrt{n}$ converges in distribution to a $N(0,\, 1)$ Gaussian random variable Y_t. By Skorohod's theorem [4, p. 333, Th. 25.6], one can in fact put the random variables in the same probability space such that $(S^n_t(\omega) - n\mu)/\sqrt{n} \to Y_t(\omega)$ for every sample point ω. Thus, in (5), the $o(\sqrt{n})$ term refers to a sequence of random variables $\{A_n(\omega)\}_n$ such that $A_n(\omega)/\sqrt{n} \to 0$ for all ω. This is consistent with and in fact a generalization of our usage of the $o(\sqrt{n})$ notation in (2).

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

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

This is given more formally in the following proposition.

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

$$\frac{-1}{\mu}\left(\frac{1}{K}\sum_{k=1}^{K} Y_{t_k} + \alpha_q \hat{\sigma}_K\right) \quad (9)$$

where $Y_{t_1}, \ldots, Y_{t_K}$ are independent, identically distributed (i.i.d.) $N(0, \sigma^2)$ random variables.

Proof: The details of the proof are similar to that of [10, Prop. 3.1] for the case of individual flow measurements. ∎

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 [10]. In that case

$$M_0 = n - \frac{\sqrt{n}}{\mu}\left(\frac{1}{K}\sum_{k=1}^{K} Y_{t_k} + \alpha_q \sigma\right) + o\left(\sqrt{n}\right). \quad (10)$$

Comparing (10) with (8), 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. The first term in (6), $1/K\sum_{k=1}^{K} Y_{t_k}$, is due to the estimation error in the mean. From (6)

$$1/K\sum_{k=1}^{K} Y_{t_k} = \sqrt{n}\,(\hat{\mu} - \mu) + o(1)$$

so we see that the effect of the mean estimation error on the variability of M_0 is magnified by a factor of $\sqrt{n}$. On the other hand, the randomness in the variance estimator enters directly in (9). 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 (the former of order $1/\sqrt{n}$ and the latter of order 1), and this compensates exactly for the difference in order of magnitude of the sensitivities. These observations will have implications in Section VI-B.

We next investigate the effect of the variability 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 the Central Limit Theorem [11, Lemma II.2], we get the following asymptotic approximation:[6]

$$S_t := \sum_{i=1}^{M_0} X_i(t) = M_0\mu + Y_t\sqrt{n} + o\left(\sqrt{n}\right). \quad (11)$$

Here again $Y_t \sim N(0, \sigma^2)$. Substituting (8), we get

$$S_t = n\mu + \left(Y_t - \frac{1}{K}\sum_{k=1}^{K} Y_{t_k} - \alpha_q\hat{\sigma}_K\right)\sqrt{n} + o\left(\sqrt{n}\right). \quad (12)$$

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

$$\Pr\{S_t > n\mu\} \approx \Pr\left\{\frac{1}{\hat{\sigma}_K}\left(Y_t - \frac{1}{K}\sum_{k=1}^{K} Y_{t_k}\right) > \alpha_q\right\}. \quad (13)$$

Now, since the Y_{t_k}s are $N(0, \sigma^2)$, the random variables $(1/K)\sum_{k=1}^{K} Y_{t_k}$ 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, [3]), the two estimates are independent, and

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

which is 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 $Y_t - (1/K)\sum_{k=1}^{K} Y_{t_k}$ is independent of $\hat{\sigma}_K$ and is distributed as $N(0, ((K+1)/K)\sigma^2)$ 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 [3].

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

$$\lim_{n\to\infty} \Pr\{S_t > n\mu\} = F_{K-1}\left(\sqrt{\frac{K}{K+1}}\,Q^{-1}(p_q)\right) \quad (14)$$

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 [11, Prop. II.3] says that with n independent individual flow measurements at each of the K time instants, the asymptotic overflow probability is given by

$$Q\left(\sqrt{\frac{K}{K+1}}\,Q^{-1}(p_q)\right). \quad (15)$$

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

$$f_{K-1}(x) = \frac{\Gamma\left(\frac{K}{2}\right)}{\sqrt{\pi(K-1)}\,\Gamma\left(\frac{K-1}{2}\right)}\left(1 + \frac{x^2}{K-1}\right)^{-(K/2)} \quad (16)$$

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.

[6]Note that this holds even though M_0 and the $X_i(t)$s are dependent.

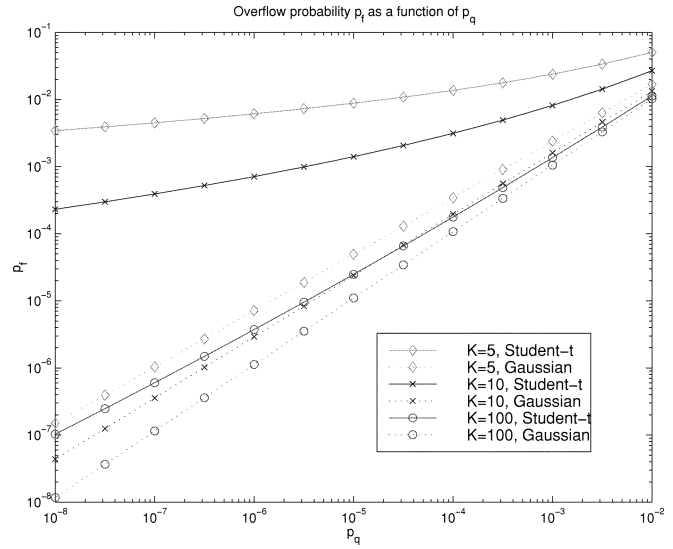

Fig. 1. Overflow probability p_f as a function of the target overflow probability p_q, for various K [`Student-t` corresponds to aggregate measurements according to (14), `Gaussian` to per-flow measurements according to (15)].

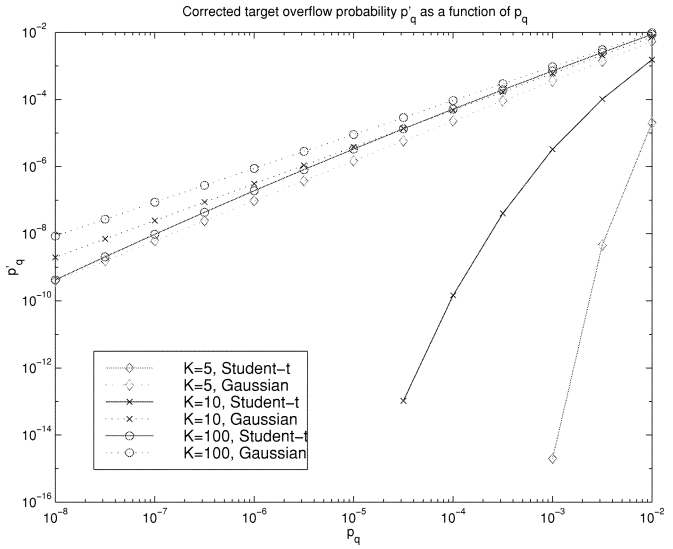

Fig. 2. Corrected overflow probability p'_q as a function of p_q.

Thus, for small K, the target overflow probability is missed significantly more in the case when only aggregate measurements are available; see Fig. 1. 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 $K \to \infty$, p_f approaches p_q under both aggregate and individual flow measurements.

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

To compensate for the measurement uncertainty for a fixed K, one way is to choose a more conservative value p'_q instead of p_q in the admission rule (3) so that we can meet the desired target p_q. The appropriate value of p'_q can be calculated according to the expression on the right-hand side of (15), i.e., choose p'_q to satisfy

$$F_{K-1}\left(\sqrt{\frac{K}{K+1}}\,Q^{-1}(p'_q)\right) = p_q$$

for a given QoS requirement. Fig. 2 compares the adjusted values of p'_q needed in the aggregate and individual flow measurement cases. We see that much more compensation is needed in the former case, especially for small K. From (12), we see that this conservative choice translates directly to a loss in average utilization $E[S_t]$ of

$$\left[Q^{-1}(p'_q) - Q^{-1}(p_q)\right] \mathrm{E}\left[\hat{\sigma}_K\right] \sqrt{n}.$$

It is interesting to note that the difference between estimation using individual flows and aggregate flow measurements is analogous to that between estimating the mean of a Gaussian distribution with and without knowing the variance. Without knowing the variance, it has to be estimated from the data as well and the resulting confidence intervals are much larger than when the variance is known.

IV. Critical Time Scale $\widetilde{T}_h$

In the previous section, it was assumed that flows stay in the system for infinite duration, and the goal of the MBAC is to determine the appropriate number of flows to admit on the basis of measurements of the *long-term* mean and variance of their stationary bandwidth distribution. If flow departure and arrival dynamics are now taken into account, then a more basic question is: What are the right statistics to measure? To address this question, we now take a step back and look more carefully at the interplay between flow dynamics, traffic fluctuation dynamics, and the admission controller. We argue that one should still measure the mean and variance statistics of the traffic fluctuations, but on a certain *critical time scale* dictated by how fast flows depart from the system.

As before, let S_t^n be the aggregate bandwidth when there are n flows in the system, and suppose that the flows are i.i.d. random processes, with stationary mean μ and variance σ^2. As in (5), the Central Limit Theorem implies that for large n

$$S_t^n = n\mu + Y_t\sqrt{n} + o\left(\sqrt{n}\right) \tag{17}$$

with the fluctuation of S_t^n around $n\mu$ on the order of $\sqrt{n}$.

Suppose now at time t, there are N_t flows in the system. This is random as a result of both the admission control and the flow departure processes. Let S_t denote the aggregate bandwidth of these N_t flows. As in (11), the fluctuation of S_t around its mean has two components, one due to the fluctuation of the number of flows in the system, and one due to the bandwidth fluctuation:

$$\begin{aligned} S_t &= N_t\mu + Y_t\sqrt{n} + o\left(\sqrt{n}\right) \\ &= n\mu + Y_t\sqrt{n} - (n - N_t)\mu + o\left(\sqrt{n}\right). \end{aligned} \tag{18}$$

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 while rejecting new ones. The aggregate 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 is the rate at which N_t can decrease if no new flows are admitted, and corresponds to a "bandwidth departure rate" of $n\mu/T_h$.

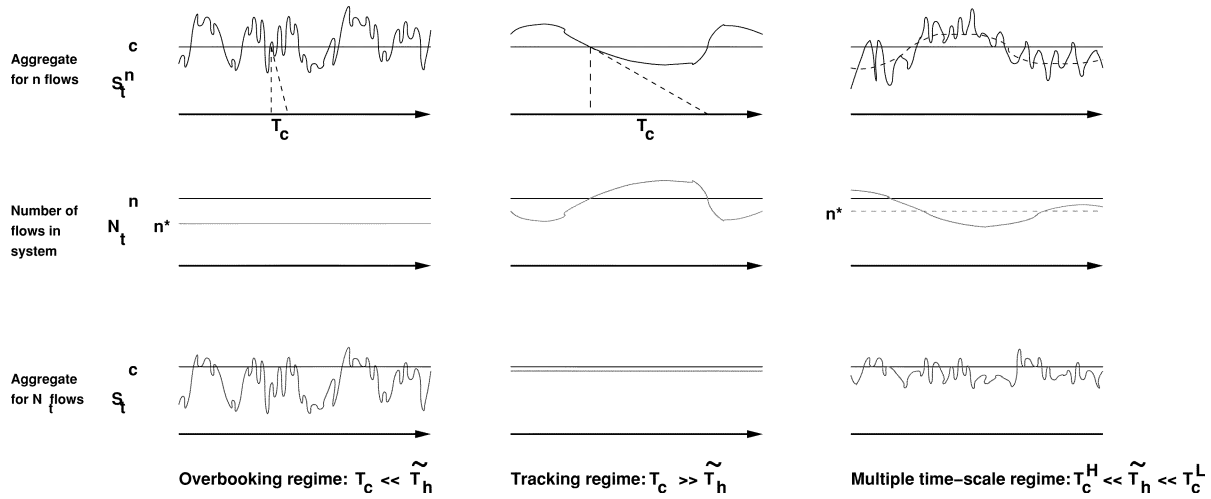

Fig. 3. Overbooking and tracking regimes. In the overbooking regime, bandwidth fluctuation is absorbed by overbooking resources, i.e., setting spare bandwidth aside to accommodate the fluctuation of the aggregate load. In the tracking regime, bandwidth fluctuation is absorbed by a corresponding fluctuation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

First, assume that the aggregate bandwidth S_t^n fluctuates over a single time scale T_c.[7] This means the rate of bandwidth fluctuation is of the order of $\sigma\sqrt{n}/T_c$. If $n\mu/T_h \ll \sigma\sqrt{n}/T_c$, or $T_c \ll (\sigma/\mu)(T_h/\sqrt{n})$, the rate of bandwidth fluctuation is much faster than the flow bandwidth departure rate. As a result, spare bandwidth has to be set aside by the MBAC to cater for the burstiness of the traffic, and full link utilization cannot be achieved. The amount of spare bandwidth is given by $\mu(n - E[N_t]) = \mu(n - n^*)$. (See the first column of Fig. 3.) Let us call this the *overbooking* regime.

Consider the other extreme, when $T_c \gg (\sigma/\mu)(T_h/\sqrt{n})$, i.e., the bandwidth fluctuation rate is much slower than the flow departure rate. In this case, there is actually no need to set aside spare bandwidth to cater for the fluctuations. Instead, the fluctuations can simply be compensated for by controlling the number of flows in the system. This is possible because flows are departing fast enough. When S_t^n happens to be larger than $n\mu$, i.e., exceeding the link capacity, the number of flows can be lowered to $N_t < n$ such that the aggregate bandwidth does not exceed the link capacity. This can be called the *tracking* regime. Provided that there are enough flows requesting admission, full utilization can be achieved. (See the second column of Fig. 3.) The time scale

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

can now be thought of as a *critical time scale* separating the tracking and the overbooking regimes.

More generally, aggregate bandwidth fluctuates over multiple time scales. The components having time scale $T_c^L \gg \widetilde{T}_h$ can be compensated for through flow admissions and departures, while the components having time scale $T_c^H \ll \widetilde{T}_h$ have to be absorbed through allocation of spare bandwidth in the link. (See the last column of Fig. 3.) The answer to the question of "what to measure" is now obvious: The slow time-scale fluctuations should be *tracked* to allow for compensation, while the variance of the fast time-scale fluctuations should be measured so that the appropriate amount of spare bandwidth can be set aside. Note that the slow time-scale fluctuation is essentially the aggregate bandwidth time-averaged over a sliding window of length $\widetilde{T}_h$. Hence, this reasoning suggests that, as in the previous section, we should be measuring the mean and variance of traffic fluctuations, but now over $\widetilde{T}_h$ rather than over the infinite horizon.

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 that 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. MBAC Design

A. Basic Architecture

Fig. 4 shows the basic architecture of the proposed MBAC design that realizes the conceptual ideas developed in the last section. By means of a pair of low-pass and high-pass filters, the aggregate bandwidth process S_t is decomposed 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 $1/\widetilde{T}_h$. 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. Together, these two estimates determine 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 the main insight we gained from Section III:

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

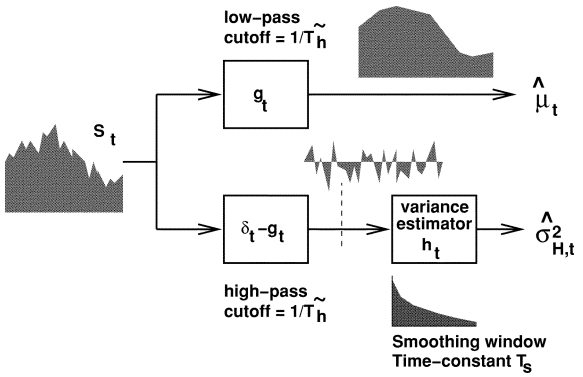

Fig. 4. Decomposition of the measured aggregate bandwidth into a high-frequency component for the variance estimator and a low-frequency component for the mean estimator.

- With only aggregate measurements, the performance of an MBAC can be quite poor if there are only a small 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 σ_H^2 spaced at $\widetilde{T}_h$ apart. The above observation thus translates into the need of a measurement window with length $K\widetilde{T}_h$, $K \gg 1$.

With this choice of measurement window size, a natural question is the robustness to nonstationarities, 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. Description of the Proposed MBAC

We now give a specific algorithm to make admission decisions based on the architecture just described. We first specify the filters. For simplicity, the filters will be defined in continuous time, although in practice they will be implemented in discrete time via sampling of the traffic. While many low-pass filters can be used, for concreteness let us consider a simple first-order AR filter with impulse response given by

$$g_t := \frac{1}{\widetilde{T}_h} \exp\left(-\frac{t}{\widetilde{T}_h}\right) u_t \tag{19}$$

where u_t is the unit step function. Let

$$h_t := \frac{1}{T_s} \exp\left(-\frac{t}{T_s}\right) u_t$$

be the low-pass filter for estimating the variance, where $T_s = K\widetilde{T}_h$ is the window length for the variance estimator. If S_t is the aggregate load at time t, the estimated mean is then

$$\hat{\mu}_t = \int_0^\infty \frac{S_{t-\tau}}{N_{t-\tau}}\, g_\tau\, d\tau \tag{20}$$

where N_t is the number of flows in the system at time t. One can think of S_t/N_t as the *instantaneous average per-flow bandwidth*. The high-pass component of the aggregate load is

$$S_t^H := S_t - N_t\hat{\mu}_t$$

which corresponds to filtering S_t through a filter with impulse response $\delta_t - g_t$. The estimate of the high-pass variance is given by

$$\hat{\sigma}_t^H = \left[\int_0^\infty \left[\frac{S_{t-\tau}^H}{N_{t-\tau}} - \int_0^\infty \frac{S_{t-u}^H}{N_{t-u}}\, h_u\, du\right]^2 h_\tau\, d\tau\right]^{1/2}. \tag{21}$$

The number of flows M_t admissible by the MBAC at time t is given by the solution to the equation

$$Q\left(\frac{c - M_t\hat{\mu}_t}{\sqrt{M_t}\hat{\sigma}_t^H}\right) = p_q. \tag{22}$$

The MBAC, therefore, admits a new flow if $M_t \geq N_t + 1$, i.e., if

$$c - (N_t + 1)\hat{\mu}_t > \alpha_q\hat{\sigma}_t^H\sqrt{N_t + 1} \tag{23}$$

and rejects it otherwise.

The left-hand side of (23) can be interpreted as the estimated available spare bandwidth (after acceptance of the new flow), and the right-hand side as the estimated *required* spare bandwidth to accommodate the fast time-scale fluctuations.

One observation is that although the algorithm uses aggregate rather than individual load measurements, it still needs to keep track of the number of flows in the system (N_t). In Section VII, we discuss a relaxed version of the above admission criterion that does not even require this knowledge. This is beneficial for distributed admission control.

VI. Performance of MBAC Scheme

We now analyze the performance of the MBAC scheme proposed above in a fully dynamical model with flow arrival and departures. We assume that 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 QoS performance (overflow probability) experienced by admitted flows of any admission control algorithm under finite arrival rate will be no worse than its performance in this model.[8] Another advantage of this model is that we need not worry about the specific flow arrival process, which may be difficult to model in practice. From the analysis point of view, this model is convenient, as the link is always filled with at least the number of flows deemed admissible by the controller. The drawback of this arrival model is that it only yields an upper bound on the utilization achieved under finite arrival loads.

We first analyze the performance of the MBAC when the traffic is homogeneous. Then we will extend the analysis to a heterogeneous traffic model. The main new ingredient here is that flow heterogeneity leads to a time-varying flow mix in the system. Under a natural heterogeneous traffic model, we show that the time constants of the filters in the proposed MBAC

[8]However, the utilization would be slightly lower when the flow arrival rate is finite. The infinite arrival model introduced in [10] reflects our belief that robustness to heavy offered load is more important than maximizing utilization during periods of modest load.

scheme are scaled appropriately to track and compensate for this time variation.

Compared with the analysis in Section III, the performance analysis in this section is heuristic in nature. Rigorous justifications will invoke the theory of weak convergence of random processes. This was done in the related analysis in [10], and we expect that a similar treatment can be done for this paper as well.

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 $\rho(t) := \mathrm{E}[(X_i(0)-\mu)(X_i(t)-\mu)]$. The capacity c is scaled as $n\mu$.

Our analysis is in the asymptotic regime where n is large, i.e., $n \to \infty$. As we scale up the system, we keep the critical time scale $\widetilde{T}_h$ fixed, such that the average flow holding time scales as $T_h = \sqrt{n}\widetilde{T}_h$. The earlier discussion on the fundamental nature of $\widetilde{T}_h$ suggests why this scaling makes sense, as it allows us to focus on the time scale "where the action is." The same scaling is used in our earlier paper [11].

The key quantities to be analyzed are M_t, which is the number of flows the MBAC determines that *should be* admissible at time t, and N_t, which is the number of flows that are actually in the system at time t. In the asymptotic regime of large capacity (equivalent to large n), both of these quantities are of order n, with random fluctuation of order $\sqrt{n}$. This is due to the Central Limit Theorem. The goal is to analyze the fluctuation to enable us to approximate the overflow probability.

We can analyze the distribution of the process $\{M_t\}$ in a similar way as in Section III. First, let us focus on the aggregate load S_t. Write

$$S_t = N_t\mu + \sum_{i\in\mathcal{F}_t}[X_i(t)-\mu].$$

Recall that $\mathcal{F}_t$ is the set of flows that are in the system at time t. By the Central Limit Theorem for the sum of a random number of random variables [as in (11)]

$$\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{F}_t}[X_i(t)-\mu] \to Y_t \tag{24}$$

where $\{Y_t\}$ is a zero-mean Gaussian process. To compute the covariance function of $\{Y_t\}$, consider for $s \le t$

$$\mathrm{E}\left[\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{F}_s}[X_i(s)-\mu]\cdot\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{F}_t}[X_i(t)-\mu]\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i\in\mathcal{F}_s\cap\mathcal{F}_t}[X_i(s)-\mu][X_i(t)-\mu]\right].$$

Now, both N_s and N_t are random variables of order n. Because the flow holding time is scaled as $\sqrt{n}\widetilde{T}_h$, with $\widetilde{T}_h$ fixed, the number of flows that depart during the time interval $[s, t]$ are of the order of $\sqrt{n}$. Hence, $|\mathcal{F}_s \cap \mathcal{F}_t|$ is of the order of n. This implies that

$$\mathrm{E}\left[\frac{1}{n}\sum_{i\in\mathcal{F}_s\cap\mathcal{F}_t}[X_i(s)-\mu][X_i(t)-\mu]\right] \to \rho(t-s)$$

where $\rho(\tau)$ is the covariance function of an individual flow:

$$\rho(\tau) := \mathrm{E}\left[[X_i(0)-\mu][X_i(\tau)-\mu]\right].$$

Hence, one can take the covariance function of the approximating Gaussian process $\{Y_t\}$ to be $\rho(\tau)$.

We now have the approximation

$$S_t \approx N_t\mu + \sqrt{n}Y_t. \tag{25}$$

Using (20), the low-pass mean estimator is asymptotically given by

$$\hat{\mu}_t \approx \mu + \frac{1}{\sqrt{n}}Z_t$$

where $Z_t = (g*Y)_t$ is the Gaussian process after filtering $\{Y_t\}$ by the low-pass filter g defined in (19).

By a linearization of the defining (22) for M_t, it can be shown that, analogous to Proposition III.1

$$M_t \approx n - \frac{\sqrt{n}}{\mu}\left(Z_t + \alpha_q\hat{\sigma}_t^H\right). \tag{26}$$

Hence, the number of admissible flows at time t is a random quantity with fluctuations 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 compensation for 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.

If the measurement window size $T_s = K\widetilde{T}_h$ is chosen such that $K \gg 1$, we observe that $\hat{\sigma}_t^H$ is approximately a constant σ_H for any t, 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. This observation 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. If $K \gg 1$, 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. Substituting this into (26), we obtain the following:

$$M_t \approx n - \frac{\sqrt{n}}{\mu}\left(Z_t + \alpha_q\sigma_H\right). \tag{27}$$

The actual number of flows N_t in the system at time t is no less than M_t because 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 follow the fluctuations of M_t. 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 the 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

$$N_t = N_{s^*} - D[s^*, t] = M_{s^*} - D[s^*, t]. \tag{28}$$

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

$$N_t = N_s + A[s,t] - D[s,t] \ge N_s - D[s,t] \ge M_s - D[s,t] \tag{29}$$

where $A[s, t]$ is the number of flows *admitted* during $[s, t]$. Thus, we conclude from (28) and (29) that

$$N_t = \sup_{s\le t}\{M_s - D[s, t]\}. \tag{30}$$

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.

Under the scaling of $T_h = \sqrt{n}\widetilde{T}_h$ for fixed $\widetilde{T}_h$, the number of flows departed in $[s, t]$ can be calculated to be

$$D[s, t] \approx \frac{t-s}{\widetilde{T}_h}\sqrt{n}. \tag{31}$$

Substituting (27) and (31) into (30), we obtain the following asymptotics for N_t in the regime of large n:

$$N_t \approx n + \sqrt{n}\cdot\frac{1}{\mu}\sup_{s\le t}\left\{-Z_s - \frac{\mu(t-s)}{\widetilde{T}_h} - \alpha_q\sigma_H\right\}.$$

Thus, the actual number of flows in the network is a random process which fluctuates on the order of $\sqrt{n}$. Under the proposed MBAC, the randomness is due only to the randomness in the low-pass mean bandwidth estimator $\hat{\mu}_t$ and not that of the variance estimator. This is because the measurement window chosen has a much longer time scale than that of the high-frequency fluctuations we want to measure.

Once we obtain an approximation for N_t, we can immediately deduce an approximation for the aggregate load S_t via (25) and, hence, the steady-state overflow probability $p_f = \Pr\{S_t > c\}$

$$S_t \approx n\mu + \sqrt{n}\cdot\sup_{s\le t}\left\{Y_t - Z_s - \frac{\mu}{\widetilde{T}_h}(t-s) - \alpha_q\sigma_H\right\}. \tag{32}$$

Hence, the overflow probability p_f converges to

$$\Pr\left\{\sup_{s\le 0}\left\{Y_0 - Z_s + \frac{\mu}{\widetilde{T}_h}s\right\} > \alpha_q\sigma_H\right\}. \tag{33}$$

To repeat, $\{Y_t\}$ is a zero-mean Gaussian process with covariance function $\rho(\cdot)$, and $Z_t = (g * Y)_t$ is the low-pass filtered version of $\{Y_t\}$.

Expression (33) 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 is given by [12], [13]

$$\frac{1}{2}\int_0^\infty v^+(0)\,\frac{\alpha_q\sigma_H + \frac{\mu}{\widetilde{T}_h}t}{\sigma^3(t)}\,\phi\left(\frac{\alpha_q\sigma_H + \frac{\mu}{\widetilde{T}_h}t}{\sigma(t)}\right)dt + Q\left(\frac{\alpha_q\sigma_H}{\sigma(0)}\right) \tag{34}$$

where $\sigma^2(t) := \mathrm{E}[(Z_{-t} - Y_0)^2]$, $v^+(0)$ is the right derivative of the function $\sigma^2(t)$ at $t = 0$, and $\phi(\cdot)$ is the $N(0, 1)$ probability density function. This expression can be numerically computed given the covariance function $\rho(\cdot)$ of the individual flow process. It is an approximation in the sense that as $p_q \to 0$, the ratio of the expression and the probability (33) approaches 1.

Let us apply the above results on two specific examples to obtain a better intuitive understanding of how the MBAC scheme functions. Under a separation of time-scale assumption, we will see that the choice of the low-pass filter time scale as $\widetilde{T}_h$ is the appropriate one. More generally, (34) can be used to assess the impact of using a different low-pass filter time scale on the performance of the MBAC scheme.[9]

[9]Expression (33) depends on the low-pass filter time scale through the second-order properties of $\{Z_t\}$.

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

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

with correlation at a single time scale T_c. By straightforward calculations, the covariance function of $\{Z_t\}$ is

$$\rho_Z(t) = \frac{\sigma^2}{2\left[\left(\frac{\widetilde{T}_h}{T_c}\right)^2 - 1\right]}\left[\frac{\widetilde{T}_h}{T_c}\exp\left(-\frac{|t|}{\widetilde{T}_h}\right) - \exp\left(-\frac{|t|}{T_c}\right)\right]$$

and the variance of the high-frequency component is

$$\sigma_H^2 = \frac{\widetilde{T}_h}{T_c + \widetilde{T}_h}\,\sigma^2.$$

Consider the regime when $T_c \ll \widetilde{T}_h$; this can be considered as a separation between the burst and flow time scales, and corresponds to the overbooking regime discussed in Section IV.

$$\rho_Z(t) \approx \frac{1}{2}\left(\frac{T_c}{\widetilde{T}_h}\right)\sigma^2\exp\left(-\frac{|t|}{\widetilde{T}_h}\right) \approx 0$$

and

$$\sigma_H^2 \approx \sigma^2$$

so that

$$\begin{aligned}\Pr&\left\{\sup_{s\le 0}\left\{Y_0 - Z_s + \frac{\mu}{\widetilde{T}_h}s\right\} > \alpha_q\sigma_H\right\}\\ &\approx \Pr\left\{\sup_{s\le 0}\left\{Y_0 + \frac{\mu}{\widetilde{T}_h}s\right\} > \alpha_q\sigma\right\}\\ &= \Pr\{Y_0 > \alpha_q\sigma\} = p_q.\end{aligned}$$

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 have 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 will not be met.

For general T_c, we can use (34) to compute the performance. This is plotted in Fig. 5 for two values of p_q. We see that the actually achieved 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 [11], which always reserves spare bandwidth proportional to σ, where σ^2 is the total variance. The per-flow scheme is effectively using a low-pass filter with a time scale much larger than T_c. 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_1 and T_2. More concretely, suppose

$$X(t) = \mu + X^{(1)}(t) + X^{(2)}(t)$$

where $\{X^{(1)}(\cdot)\}$ and $\{X^{(2)}(\cdot)\}$ are zero-mean independent stationary processes with covariance functions

$$\rho^{(1)}(t) = \sigma_1^2\exp\left(-\frac{|t|}{T_1}\right) \quad \rho^{(2)}(t) = \sigma_2^2\exp\left(-\frac{|t|}{T_2}\right).$$

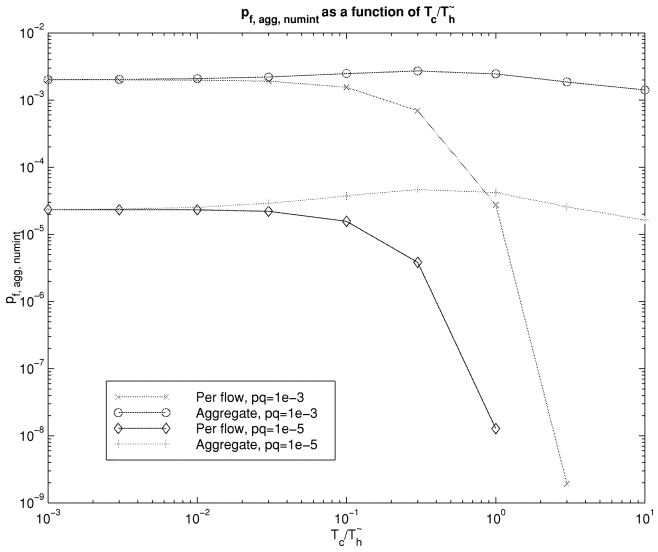

Fig. 5. Overflow probability of the proposed aggregate scheme and a per-flow scheme which always overbooks at σ.

Then we can decompose the scaled aggregate fluctuation $Y_t = Y_t^{(1)} + Y_t^{(2)}$ and the low-pass output $Z_t = Z_t^{(1)} + Z_t^{(2)}$ accordingly. The covariance functions of $\{Z^{(j)}(\cdot)\}$ is given by

$$\rho_Z^{(j)}(t) = \frac{\sigma_j^2}{\left(\frac{\widetilde{T}_h}{T_j}\right)^2 - 1}\left[\frac{\widetilde{T}_h}{T_j}\exp\left(-\frac{|t|}{\widetilde{T}_h}\right) - \exp\left(-\frac{|t|}{T_j}\right)\right], \quad j = 1,\, 2$$

and for $s,\, t \le 0$

$$\mathrm{E}\left[\left(Y_0^{(j)} - Z_t^{(j)}\right)\left(Y_0^{(j)} - Z_s^{(j)}\right)\right] = \sigma_j^2 - \frac{\sigma_j^2}{1 + \frac{\widetilde{T}_h}{T_j}}\left[\exp\left(\frac{t}{T_j}\right) + \exp\left(\frac{s}{T_j}\right)\right] + \rho_Z^{(j)}(t-s). \quad (35)$$

Now, consider the regime when $T_1 \ll \widetilde{T}_h$ and $T_2 \gg \widetilde{T}_h$. By performing a rescaling of time, we can write the overflow probability (33) as

$$\Pr\left\{\sup_{s\le 0}\left\{Y_0 - Z_{\widetilde{T}_h s} + \mu s\right\} > \alpha_q \sigma_H\right\} = \Pr\left\{\sup_{s\le 0}\left\{Y_0^{(1)} - Z_{\widetilde{T}_h s}^{(1)} + Y_0^{(2)} - Z_{\widetilde{T}_h s}^{(2)} + \mu s\right\} > \alpha_q \sigma_H\right\}. \quad (36)$$

Using (35), we see that for $s,\, t < 0$

$$\mathrm{E}\left[\left(Y_0^{(1)} - Z_{\widetilde{T}_h t}^{(1)}\right)\left(Y_0^{(1)} - Z_{\widetilde{T}_h s}^{(1)}\right)\right] \approx \mathrm{E}\left[\left(Y_0^{(1)}\right)^2\right] = \sigma_1^2 \quad (37)$$

and

$$\mathrm{E}\left[\left(Y_0^{(2)} - Z_t^{(2)}\right)\left(Y_0^{(2)} - Z_s^{(2)}\right)\right] \approx 0. \quad (38)$$

This means that the process $\{Y_0^{(1)} - Z_{\widetilde{T}_h t}^{(1)}\}$ has variance σ_1^2 and is highly correlated over time (with correlation coefficient close to 1 for all $s,\, t$), while the process $\{Y_0^{(2)} - Z_{\widetilde{T}_h t}^{(2)}\}$ is close to zero. Thinking of $Y_t^{(1)}$ as the fast time-scale fluctuation and $Y_t^{(2)}$ as the slow time-scale fluctuation of the traffic, this means that the low-pass filter tracks the latter almost perfectly (at the time scale defined by $\widetilde{T}_h$) but leaves the former essentially unchanged. It can also be verified that $\sigma_H \approx \sigma_1$. Using (36), the overflow probability can be seen to be close to the target.

The choice of $\widetilde{T}_h$ as the memory time scale of the low-pass filter is important to keep the utilization high. Using (32) and again rescaling time, the average utilization is given by

$$\mathrm{E}[S_0] \approx n\mu + \sqrt{n}\,\mathrm{E}\left[\sup_{s\le 0}\left\{-Z_{\widetilde{T}_h s} + \mu s\right\}\right] - \sigma_1 \alpha_q \sqrt{n}.$$

Now, for the regime considered above and the memory time scale equal to $\widetilde{T}_h$, one can calculate that $\rho_Z(\widetilde{T}_h t) \approx \sigma_2^2$, i.e., over the critical time scale, Z_t is highly correlated and, hence, remains essentially constant. Thus, the supremum above is achieved at $s = 0$, and

$$\mathrm{E}[S_0] \approx n\mu + \sqrt{n}\,\mathrm{E}[Z_0] - \sigma_1 \alpha_q \sqrt{n} = n\mu - \sigma_1 \alpha_q \sqrt{n}.$$

The spare bandwidth $\sigma_1 \alpha_q \sqrt{n}$ is precisely left for catering for the fast time-scale fluctuation.

Suppose now that the memory time scale of the low-pass filter is chosen to be larger than $\widetilde{T}_h$. As the memory time scale approaches T_2, some of the slow fluctuations (at the time scale T_2) are filtered into the high-frequency component, resulting in a larger than necessary spare bandwidth. In the extreme case when $T_m \gg T_2$, $\sigma_H^2 = \sigma_1^2 + \sigma_2^2$, resulting in a utilization of

$$n\mu - \sqrt{\sigma_1^2 + \sigma_2^2}\,\alpha_q \sqrt{n}.$$

Compared with the case when $T_m = \widetilde{T}_h$, this represents a loss of utilization of

$$\left(\sqrt{\sigma_1^2 + \sigma_2^2} - \sigma_1\right)\alpha_q \sqrt{n}.$$

This calculation serves as a validation of the design choice of $\widetilde{T}_h$ as the low-pass filter time scale, and confirms our informal discussions on the importance of the critical time scale $\widetilde{T}_h$.

B. Heterogeneous Flows

To study the robustness of the MBAC scheme to flow heterogeneity, 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 i.i.d. with zero mean and unit variance, and are stationary and ergodic with covariance function $\rho_U(t)$; they are also independent of μ_is and σ_is. 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_is and σ_is for simplicity. The random variables μ_i and σ_i^2 have the following statistics:

$$\mathrm{E}[\mu_i] = \mu \quad \mathrm{Var}[\mu_i] = v^2 \quad \mathrm{E}[\sigma_i^2] = \sigma^2.$$

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. Also, we only model heterogeneity in first- and second-order traffic statistics because our asymptotic analysis only depends on them.

As in the homogeneous case, we are here interested in the regime where the capacity $c = n\mu$ is large, $\widetilde{T}_h$ is fixed, and the

average flow duration T_h scales as $\sqrt{n}\widetilde{T}_h$. The aggregate load in the system is given by

$$S_t = N_t\mu + \sum_{i\in\mathcal{F}_t}(\mu_i - \mu) + \sum_{i\in\mathcal{F}_t}[X_i(t) - \mu_i].$$

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) $\sum_{i\in\mathcal{F}_t}(\mu_i-\mu)$, 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) $\sum_{i\in\mathcal{F}_t}[X_i(t)-\mu_i]$, which is the fluctuation of the flows from their long-term average rates.

Similar to (24) in the homogeneous case, we can approximate the third term by $\sqrt{n}\,V_t$, where V_t is a zero-mean Gaussian process. The covariance function $\rho_V(\cdot)$ of V_t can be calculated as follows:

$$\begin{aligned}
&\mathrm{E}\left[\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{F}_s}[X_i(s)-\mu_i]\cdot\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{F}_t}[X_i(t)-\mu_i]\right]\\
&\quad\mathrm{E}\left[\frac{1}{n}\sum_{i\in\mathcal{F}_s\cap\mathcal{F}_t}[X_i(s)-\mu_i][X_i(t)-\mu_i]\right]\\
&\quad\to E\left[(X_i(s)-\mu_i)(X_i(t)-\mu_i)\right]\\
&\quad= E\left[\sigma_i^2U_i(s)U_i(t)\right]\\
&\quad= \sigma^2\rho_U(t-s).
\end{aligned}$$

Hence, $\rho_V(\tau)=\sigma^2\rho_U(t)$.

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 $\{L_\tau\}$ is a zero-mean Gaussian process. To compute the covariance function of $\{L_\tau\}$, consider

$$\begin{aligned}
&\mathrm{E}\left[\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{F}_0}[\mu_i-\mu]\cdot\frac{1}{\sqrt{n}}\sum_{j\in\mathcal{F}_{T_h\tau}}[\mu_j-\mu]\right]\\
&\quad=\mathrm{E}\left[\frac{1}{n}\sum_{i\in\mathcal{F}_0\cap\mathcal{F}_{T_h\tau}}[\mu_i-\mu]^2\right]\\
&\quad\to v^2e^{-\tau}
\end{aligned}$$

where $v^2=\mathrm{E}[[\mu_i-\mu]^2]$. The convergence in the last line follows from the fact that $N_0/n\to 1$ and that out of the N_0 flows in the system at time 0, the expected number of flows still remaining in the system at time $T_h\tau$ is $N_0e^{-\tau}$. Hence, we have

$$\rho_L(\tau)=v^2e^{-\tau}.$$

The process L_{t/T_h} 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.

Summarizing the above, we have

$$S_t \approx N_t\mu + \sqrt{n}\,L_{t/T_h} + \sqrt{n}\,V_t. \tag{39}$$

We note that the time fluctuation in the *flow variances* due to heterogeneity has disappeared in the approximation (39) 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. This observation will have important ramifications in the estimation of the high-pass variance.

Continuing on the performance analysis, the low-pass mean estimator is given by [via (20)]

$$\hat{\mu}_t \approx \mu + \frac{1}{\sqrt{n}}\,L_{t/T_h} + \frac{1}{\sqrt{n}}\,Z_t$$

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

$$M_t \approx n - \frac{\sqrt{n}}{\mu}\left(L_{t/T_h} + Z_t + \alpha_q\hat{\sigma}_t^H\right) \tag{40}$$

where $\hat{\sigma}_t^H$ is the high-pass variance estimator given by (21). If the variance measurement window $T_s = K\widetilde{T}_h$ is chosen to be much larger than $\widetilde{T}_h$, then it can be shown that

$$\begin{aligned}
\hat{\sigma}_t^H &\approx \sigma_H^2 := \mathrm{Var}\left[V_0 - (g*V)_0\right]\\
&= \sigma^2\mathrm{Var}[U(0)-(g*U)(0)].
\end{aligned} \tag{41}$$

Although this statement is identical to the corresponding one 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 much larger than $\widetilde{T}_h$. 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.

We can now compute the asymptotic distribution of N_t, the number of flows in the system.

$$\begin{aligned}
N_t &= \sup_{s\le t}\{M_s - D[s,t]\}\\
&\approx \sup_{s\le t}\left\{n - \frac{\sqrt{n}}{\mu}\left(L_{s/T_h} + Z_s + \alpha_q\sigma^H\right) - \frac{t-s}{\widetilde{T}_h}\sqrt{n}\right\}\\
&\approx n - \frac{\sqrt{n}}{\mu}\,L_{t/T_h} + \frac{\sqrt{n}}{\mu}\sup_{s\le t}\left\{-Z_s - \frac{\mu(t-s)}{\widetilde{T}_h} - \alpha_q\sigma_H\right\}
\end{aligned}$$

where the first equality follows from (30), the second equality from (31), (40), and (41), and the third equality from the fact that $T_h \gg 1$ so that L_{t/T_h} remains essentially constant in the maximization.

The aggregate load and the overflow probability can be similarly obtained, as follows:

$$\begin{aligned}
S_t &\approx N_t\mu + \sqrt{n}\,L_{t/T_h} + \sqrt{n}V_t\\
&\approx n\mu + \sqrt{n}\sup_{s\le t}\left\{V_t - Z_s - \frac{\mu(t-s)}{\widetilde{T}_h} - \alpha_q\sigma_H\right\}
\end{aligned}$$

and

$$\Pr\{S_0 > n\mu\} = \Pr\left\{\sup_{s\le 0}\left\{V_0 - Z_s + \frac{\mu s}{\widetilde{T}_h}\right\} > \alpha_q\sigma_H\right\}.$$

Comparing these results with (32) and (33), we observe that the (asymptotic) utilization and overflow probability 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^2. There are two reasons. 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 tracking. Second, the fluctuation due to change in flow variances σ_i has an insignificant

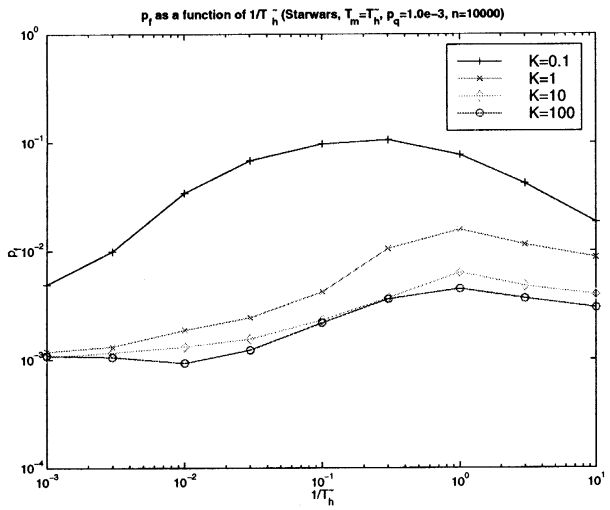

Fig. 6. Simulated overflow probability p_f for different values of K and $n = 10\,000$, based on the *Star Wars* trace; target overflow probability $p_q = 10^{-3}$.

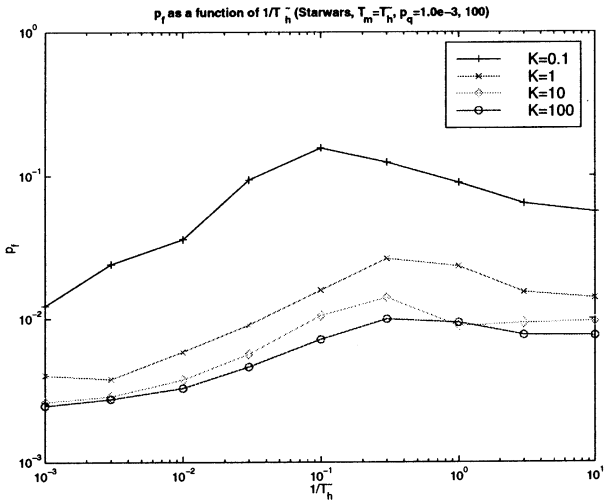

Fig. 7. Simulated overflow probability p_f for different values of K and $n = 100$, based on the *Star Wars* trace; target overflow probability $p_q = 10^{-3}$.

impact on the aggregate load and the overflow probability. This ensures that although the memory time scale for estimating the high-pass variance is much longer than $\widetilde{T}_h$, 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 a much longer averaging interval $T_s = K\widetilde{T}_h$ to estimate the high-pass variance σ_H^2. 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.

C. Simulation Results

We have performed trace-based simulations of our MBAC to verify the results of the performance analysis. The goal is to evaluate how well the MBAC performs with real network traffic, in particular traffic that exhibits fluctuations on a wide range of time scales.

The simulation is based on a compressed video trace of the movie *Star Wars*, which has been extensively studied in the literature and has been shown to exhibit fluctuations on all time scales [8].[10] In Figs. 6 and 7, we plot the measured overflow probability p_f as a function of the critical time scale, for different values of the high-pass variance estimation window $T_s = K\widetilde{T}_h$.[11]

In Fig. 6, we can clearly see that $K \gg 1$ is necessary to obtain a sufficiently reliable variance estimation. Small values of K ($K \leq 1$) lead to system overload, with the overflow probability p_f exceeding the target by up to two orders of magnitude. If K is chosen large enough, we observe that p_f is within about a half order of magnitude of the target, depending on the flow holding time T_h. Note that in the numerical results of Fig. 5, we have observed a similar "bump" in the overflow probability for the regime where the correlation and the critical time scales are close.

The results in Fig. 7 look qualitatively very similar, but they are offset toward higher overflow probabilities by about a half order of magnitude. The reason for this is as follows. The amount of spare bandwidth is on the order of $\sqrt{n}$ flows, i.e., ten flows in this simulation. The discretization effect due to the fact that bandwidth does not depart the system continuously, but in discrete steps at flow termination is not negligible, and increases the overflow probability. However, this effect is $O(1)$ and can easily be compensated for.

VII. Implementation of MBAC Within the Differentiated Services Architecture

The Internet research community has been grappling for a long time with the conflicting goals of scalability and guaranteeing QoS. Currently, the differentiated services architecture (diffserv) being defined within the IETF is the most promising solution to this problem [2]. Diffserv is capable of offering a reasonable set of QoS guarantees while maintaining one of the basic tenets of the Internet architecture: a stateless core. In the diffserv framework, only edge routers are aware of and manipulate individual flows; core routers only handle traffic aggregates through a small set of per-hop behaviors (PHBs).

In this section, we discuss a variant of our MBAC that can be implemented in such an architecture. For this, we have to relax the assumptions about what information an MBAC has available to make admission decisions. We have assumed throughout this work that the MBAC does not collect any per-flow measurements or maintain any per-flow state, but we *did* assume that the MBAC knows the exact number of flows in the system. Specifically, in order to compute (23), the MBAC has to precisely know the current number of flows N_t in the system at time t. In the diffserv architecture, this is undesirable due to the distributed nature of flow admission. Core routers can only be expected to collect and process aggregate traffic measurements. An ingress

[10]The details of the simulation setup are described in [11].

[11]In contrast to the numerical results in Fig. 5, there is no notion of a time-scale separation parameter $T_c/\widetilde{T}_h$ here, as we make no assumptions about the correlation structure of the trace. We plot the x axis of Figs. 6 and 7 as a function of $1/\widetilde{T}_h$ in analogy with Fig. 5.

router would use either in-band (e.g., using probe packets) or out-of-band signaling (e.g., RSVP [20]) to query core routers along the path of the new flow.[12] A core router then performs a local admission test, and updates the in-band probe packet or the out-of-band signaling message accordingly. However, the core router does not keep track of ongoing flows and, therefore, the admission test cannot rely on knowledge of N_t.[13]

We now discuss a conservative approximation to the admission criterion (23) that does not rely on the number of flows. In Section V-C, exact knowledge of N_t is necessary in the computation of the estimated per-flow statistics $\hat{\mu}_t$ and $\hat{\sigma}_t^H$. Let us try to forgo computing these per-flow statistics and instead use the corresponding *aggregate* variables. The goal is to develop an admission criterion analogous to (23), but that does not explicitly depend on N_t. It is convenient to repeat (23) here in a slightly modified form:

$$c - N_t\hat{\mu}_t - \hat{\mu}_t > \alpha_q\hat{\sigma}_t^H\sqrt{N_t+1}. \tag{42}$$

Recall that the left-hand side estimates the spare capacity *after* admitting the new flow; $\hat{\mu}_t$ estimates the average rate of the new flow, i.e., we have implicitly assumed that the new flow is statistically identical to the existing flows.

Without knowledge of the number of flows N_t, we cannot compute an estimator of the per-flow mean rate $\widehat{\mu_t}$. This has two consequences. First, we have to replace the estimate $\widehat{\mu_t}$ for the new flow, e.g., with a peak-rate constraint r. Fortunately, for large n, replacing $\hat{\mu}_t$ with r in (42) does not affect performance. The reason for this is that even though the peak rate assumption is conservative, as soon as a flow is admitted, the dependence of future admission decisions on this flow is *only* through its contribution to the *measured* aggregate bandwidth. Therefore, a conservative choice of r only affects one flow at a time. This effect is $O(1)$, and therefore negligible in a large system.

Second, we have to define approximations to the high and low-pass filtered aggregate bandwidth, S_t^H and S_t^L, that are independent of the number of flows N_t.

$$A_t^L := \int_0^\infty S_{t-\tau}g_\tau\,d\tau \approx n\hat{\mu}_t \tag{43}$$

$$A_t^H := S_t - A_t^L \approx S_t^H. \tag{44}$$

The estimated variance $\hat{\sigma}_t^{AH}$ of the *aggregate* fast time-scale component A_t^H can be computed from A_t^H as follows:

$$\hat{\sigma}_t^{AH} = \left[\int_0^\infty\left[A_{t-\tau}^H - \int_0^\infty A_{t-u}^H h_u\,du\right]^2 h_\tau\,d\tau\right]^{1/2} \approx \sqrt{n}\,\sigma_H. \tag{45}$$

It is tempting to use the following admission criterion:

$$c - A_t^L - r > \alpha_q\hat{\sigma}_t^{AH} \tag{46}$$

where, analogous to (23), the left-hand side of (46) is an estimation of the available spare bandwidth after admission of the new flow, and the right-hand side is the estimated required spare bandwidth to accommodate fast time-scale fluctuations of the aggregate bandwidth.

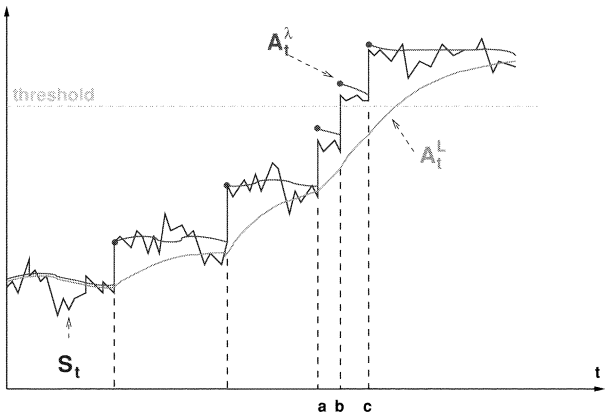

Fig. 8. Illustration of the uncorrected low-pass filtered aggregate bandwidth process A_t^L and the corrected process A_t^λ. A new flow should be admitted only if A_t^λ is below the threshold given by $c - r - \alpha_q\hat{\sigma}_t^{AH}$; A_t^L reacts too slowly to new flows, leading to possible overload (by admitting flows b and c).

However, this does not work. The problem is that the bandwidth of a new admitted flow is not immediately reflected in A_t^L. This is in contrast to the term $N_t\hat{\mu}_t$ in (23), which reacts to a flow admission immediately through the increment of N_t. Now, suppose that the flow arrival rate is very high. Using (46), the MBAC could admit a potentially large number of flows in a very short period of time and overload the system. This point is illustrated in Fig. 8: The three flows a, b, and c to the right arriving in rapid succession increase the aggregate bandwidth S_t in discontinuous steps; the low-pass filtered process A_t^L in (46) increases too slowly to avoid admission of too many flows; in this example, only flow a should have been admitted.

Specifically, suppose a new flow is admitted at time t_0. Then the mean of the low-pass filtered aggregate bandwidth $S_t^L = N_t\widehat{\mu_t}$ increases instantaneously by the mean rate μ of the new flow. However, the mean of A_t^L only converges to S_t^L exponentially with time-constant $\widetilde{T}_h$, because the increase in the aggregate bandwidth due to the new flow is low-pass filtered. The difference between the means of the two processes after t_0 is therefore approximately $\mathrm{E}[S_t^L - A_t^L] \approx \mu\exp((t_0-t)/\widetilde{T}_h)$. To correct for new flows, we have to add this term to the low-pass estimate A_t^L for each admitted flow. We obtain a corrected low-pass estimate A_t^λ, which is given by

$$A_t^\lambda = (A_t^L + \lambda_t) * g_t \tag{47}$$

where $*$ is the convolution operation. The function λ_t contains a Dirac pulse for each arriving flow

$$\lambda_t = \sum_i r\delta(t - t_i) \tag{48}$$

where t_i is the arrival time of flow i, and r is its peak rate.

Suppose that the aggregate bandwidth only changes at discrete time instants. Then A_t^λ can be computed recursively as follows. Let t_i denote the time instants where the aggregate bandwidth S_t changes or where a new flow is admitted.

$$A_{t_i}^\lambda := \phi_i A_{t_{i-1}}^\lambda + (1-\phi_i)S_{t_{i-1}} + r\cdot\mathbf{1}_{\{\text{flow admission at } t_i\}} \tag{49}$$

where

$$\phi_i = \exp\left(\frac{t_{i-1} - t_i}{\widetilde{T}_h}\right). \tag{50}$$

[12]The admission decision could also be made by a centralized *bandwidth broker*, which collects traffic measurements from the entire network. However, for scalability, the bandwidth broker is unlikely to maintain records for individual flows as well.

[13]In principle, a core router could keep track of the number of flows without identifying individual flows by tracking flow admissions and departures; however, it is obvious that this is not robust to problems such as loss of a signaling message or to a node failure.

The admission criterion with the correction term becomes

$$c - A_t^{\lambda} - r > \alpha_q \hat{\sigma}_t^{AH}. \tag{51}$$

We have seen that replacing $\hat{\mu}_t$ with r in (42) did not in itself affect the performance, because it affects *only* the new flow demanding admission. Nevertheless, for the criterion (51) without knowledge of N_t above, the peak-rate assumption does incur a penalty in utilization proportional to $(r-\mu)\sqrt{n}$. The reason is that the correction terms added to A_t^{λ} in (49) *persist* over a time scale of $\widetilde{T}_h$. This effect is $O(\sqrt{n})$, because the MBAC admits on the order of $O(\sqrt{n})$ new flows per critical time scale $\widetilde{T}_h$, each of which results in a conservative correction. Thus, A_t^{λ} overestimates $N_t\hat{\mu}_t$ on average by an amount proportional to $(r-\mu)\sqrt{n}$. This utilization penalty is the price for the limited information available to the MBAC to make robust admission decisions.

VIII. Conclusion

Previous approaches to the admission control problem 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, and a measurement-based scheme estimates these statistics to compute the number of admissible flows. For real-world traffic exhibiting multiple time-scale dynamics and flow heterogeneity, the time-scale separation assumption is questionable and the notion of "burst time-scale" ill defined. By explicitly incorporating flow dynamics into the picture, we have shown that there is a critical time-scale $\widetilde{T}_h = T_h/\sqrt{n}$ on which the traffic statistics is relevant for admission control purposes. This time scale depends only on flow dynamics and is decoupled from the traffic statistics of the flows, thus allowing us to bypass the difficult question of defining a burst time scale. The MBAC scheme proposed in this paper tracks the mean and estimates the variance of the traffic fluctuations at this time scale and makes admission control decisions accordingly. The measurement windows for these statistics are sized to make the estimation errors negligible. We have shown that the scheme is robust with respect to flow heterogeneity and bandwidth fluctuations on multiple time scales, achieves high resource utilization, and is amenable to an efficient implementation using only aggregate bandwidth measurements 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. Commun.*, vol. 43, pp. 1566–1579, Feb. 1995.

[2] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss. (1998, Dec.) An architecture for differentiated services. IETF, RFC 2475. [Online]. Available: http://www.ietf.org/rfc

[3] P. J. Bickel and K. A. Doksum, *Mathematical Statistics:Basic Ideas and Selected Topics*. San Francisco, CA: Holden-Day, 1977.

[4] P. Billingsley, *Probability and Measure*, 3rd ed. New York: Wiley, 1995.

[5] L. Breslau, S. Jamin, and S. Shenker, "Comments on the performance of measurement-based admission control algorithms," in *Proc. IEEE INFOCOM*, Tel Aviv, Israel, Mar. 2000, pp. 1233–1242.

[6] M. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic: Evidence and possible causes," in *Proc. ACM Sigmetrics*, Philadelphia, PA, May 1996, pp. 160–169.

[7] A. Elwalid and D. Mitra, "Effective bandwidth of general Markovian traffic sources and admission control of high-speed networks," *IEEE/ACM Trans. Networking*, vol. 1, pp. 329–343, June 1993.

[8] M. W. Garrett and W. Willinger, "Analysis, modeling and generation of self-similar VBR video traffic," in *Proc. ACM SIGCOMM*, London, U.K., Aug. 1994, pp. 269–280.

[9] R. J. Gibbens, F. P. Kelly, and P. B. Key, "A decision-theoretic approach to call admission control in ATM networks," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1101–1114, Aug. 1995.

[10] M. Grossglauser and D. Tse, "A framework for robust measurement-based admission control," in *Proc. ACM SIGCOMM*, Cannes, France, Sept. 1997, pp. 237–248.

[11] M. Grossglauser and D. N. C. Tse, "A framework for robust measurement-based admission control," *IEEE/ACM Trans. Networking*, vol. 7, pp. 293–309, June 1999.

[12] H. U. Bräker, "High boundary excursions of locally stationary Gaussian processes," in *Proc. Conf. Extreme Value Theory and Applications*, Gaithersburg, MD, May 1993.

[13] ——, "High boundary excursions of locally stationary Gaussian processes," Ph.D. dissertation, Univ. Bern, Bern, Switzerland, 1993.

[14] S. Jamin, P. B. Danzig, S. Shenker, and L. Zhang, "A measurement-based admission control algorithm for integrated services packet networks," *IEEE/ACM Trans. Networking*, vol. 5, pp. 56–70, Feb. 1997.

[15] G. Kesidis, J. Walrand, and C. S. Chang, "Effective bandwidths for multiclass Markov fluids and other ATM sources," *IEEE/ACM Trans. Networking*, vol. 1, pp. 424–428, Aug. 1993.

[16] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of ethernet traffic (extended version)," *IEEE/ACM Trans. Networking*, vol. 2, pp. 1–15, Feb. 1994.

[17] V. Paxson and S. Floyd, "Wide area traffic: The failure of poisson modeling," *IEEE/ACM Trans. Networking*, vol. 3, pp. 226–244, June 1995.

[18] E. P. Rathgeb, "Modeling and performance comparison of policing mechanisms for ATM network," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 325–334, Apr. 1991.

[19] J. Roberts, U. Mocci, and O. Virtamo, Eds., *Broadband Network Teletraffic*. ser. Lecture Notes in Computer Science 1155. New York: Springer, 1996.

[20] L. Zhang, S. Deering, D. Estrin, S. Shenker, and D. Zappala, "RSVP: A new Resource ReSerVation Protocol," *IEEE Network*, vol. 7, pp. 8–18, Sept. 1993.

Matthias Grossglauser (S'92–M'99) received the Diplôme d'Ingénieur en Systèmes de Communication from the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, and the M.Sc. degree from the Georgia Institute of Technology, Atlanta, both in 1994, and the Ph.D. degree from the University of Paris 6, Paris, France, in 1998.

He did most of his thesis work at INRIA, Sophia Antipolis, France. From 1998 to 2002, he was a member of the Networking and Distributed Systems Laboratory of AT&T Labs—Research, Florham Park, NJ. He is currently an Assistant Professor with the School of Computer and Communication Sciences, EPFL. His research interests are in mobile ad hoc and sensor networking, and in network traffic measurement and modeling.

Dr. Grossglauser received the 1998 Cor Baayen Award from the European Research Consortium for Informatics and Mathematics (ERCIM) and the Best Paper Award from the IEEE INFOCOM 2001 conference. He serves on the editorial board of the IEEE/ACM Transactions on Networking.

David N. C. Tse (S'91–M'96) received the B.A.Sc. degree in systems design engineering from University of Waterloo, Waterloo, ON, Canada in 1989 and the M.S. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1991 and 1994, respectively.

From 1994 to 1995, he was a Postdoctoral Member of Technical Staff with AT&T Bell Laboratories. Since 1995, he has been with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, where he is currently a Professor. His research interests are in information theory, wireless communications, and networking.

Dr. Tse received an NSERC four-year graduate fellowship from the government of Canada in 1989, a National Science Foundation CAREER award in 1998, the Best Paper Awards at the IEEE INFOCOM 1998 and INFOCOM 2001 conferences, the Erlang Prize in 2000 from the INFORMS Applied Probability Society, and the IEEE Communications and Information Theory Society Joint Paper Award in 2001. He is currently an Associate Editor for the IEEE Transactions on Information Theory.