Measurement-based Call Admission Control: Analysis and Simulation

David Tse*

Dept. EECS
University of California
Berkeley, CA 94720, USA
dtse@eecs.berkeley.edu

Matthias Grossglauser†

INRIA
Sophia Antipolis
France
matthias.grossglauser@inria.fr

Abstract

We consider the problem of admission control for variable-rate traffic sources sharing a bufferless link, in order to provide a quality-of-service in terms of overload probability. Through analysis and simulations, we study the performance of a scheme which has no prior knowledge of the traffic statistics and makes admission decision based on the current network state only. We analyze the dynamics of the system under this control, and show that in the regime of large link capacity and separation of call and burst time-scales, this scheme performs as well as the optimal scheme which has full knowledge of the statistics. We evaluate the performance of the scheme on real traffic sources.

1 Introduction

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

Traditional approaches to call admission control require an *a priori* traffic specification in terms of the parameters of a deterministic or stochastic model. The admission decision is then based on the specifications of the existing and the new flow. This approach suffers from several drawbacks. First, it is usually difficult for the user to tightly characterize his traffic in advance [1]. This is true even for stored media such as video-on-demand, as the user is expected to be able to exercise interactive control (such as pause, fast-forward etc.) As a result, traffic specifications can be expected to be quite loose. Second, there exists a modeling tradeoff between the ability to police and the statistical multiplexing gain. Deterministic models such as leaky buckets are easy to police, as they specify the *worst-case* behavior of traffic on a single time-scale, but they fail to provide a sufficient characterization to extract a large fraction of the potential statistical multiplexing gain. While a sequence of leaky buckets can approach such a multiple time-scale characterization, the number of model parameters grows accordingly [2]. Stochastic models such as those based on effective bandwidth are better suited to achieve good statistical multiplexing gain, but at the expense of policing [3]. It is not clear how to ensure that a traffic flow correspond to the specified parameters, without which call admission control can easily be "fooled".

In this paper, we focus on a different approach, in which admission control decisions are made based on network measurements alone. Instead of assuming a statistical or worst-case model for the traffic, the behavior of the current calls are monitored, and this information is used to make admission decisions. This measurement-based approach alleviates the burden on the users to provide accurate traffic models, and thus is a more practical approach for achieving statistical multiplexing gain with variable-rate traffic.

In this paper, we concentrate on a single-link bufferless model for the network. The QOS measure is the probability of network overload when the aggregate bandwidth requirements of the calls currently in the system exceeds the link capacity. The goal of admission control is to keep this overload probability below a desired threshold while minimizing the call blocking probability, or equivalently, maximizing the utilization

*This author was supported by AFOSR F49620-96-1-0199 and a grant from Pacific Bell.

†This author was supported in part by a grant from France Telecom/CNET.

of the network. We restrict ourselves to the homogeneous case, in which the statistical behavior of calls are similar.

The specific admission control scheme we are investigating is of the *certainty equivalent* type. The admission controller assumes that the measured statistics are the true statistics of the calls, and uses the information to make decisions in the same way as a controller which has perfect *a priori* knowledge of the call statistics. Certainty-equivalent controllers are generally sub-optimal, as they do not take the *measurement errors* into account. We shall examine the impact of such measurement errors on the performance of the schemes via both theoretical analysis and simulations with synthetic and real traffic.

Our main theoretical result is that in the regime of large link capacities and separation of call and burst time-scale, a memoryless certainty-equivalent control can achieve the performance of an optimal scheme with knowledge of the traffic statistics. The proof of this result yields the important insight that it is essential to consider the *dynamics* of the controlled system to gain a full understanding of its performance. Thus, even though the controller is prone to measurement errors at any single admission decision, it turns out that in the above parameter regime, overload occurs only after a *succession* of admission mistakes, which is an unlikely event.

Past work on measurement-based admission control [4], [5], [6] have either ignored measurement errors or assumed a static situation where calls do not arrive or depart the system and there is arbitrarily long time to make accurate measurements. Recent work by Gibbens et. al. [7] advocates the explicit incorporation of call-level dynamics into the model and provides much inspiration to the present work. However, there are several major differences. First, the separation of time-scale assumption is already built into their model, whereas we deal directly with the interplay between call and burst dynamics. Given that a lot of traffic sources have long-range structure, such a separation of time-scale assumption should not be taken for granted *a priori*. However, due to this complexity in our model, we have to resort to a combination of asymptotic large deviations analysis and simulations. Second, the focus of their work is on the use of prior knowledge about the sources, whereas our scheme uses no such information. Lastly, the performance of their schemes are evaluated on on-off sources, whereas we test on real multi-level traffic sources.

We conclude this section by outlining two reasons for focusing on a bufferless model. First, the dynamics leading to the overload event in a bufferless system is much simpler than that of overflowing in a buffered system, as the event occurs whenever the instantaneous aggregate traffic load exceeds the link capacity. This simplification allows us to focus on the measurement problem that is of central interest in this paper. Second, recent work on multiple time-scale traffic [8] such as compressed VBR video has indicated that a significant bulk of the statistical multiplexing gain can be obtained by a Renegotiated Constant Bit Rate (RCBR) service. In this service model, buffering only occurs at the network edge. Bandwidth renegotiations fail when the current aggregate bandwidth demand exceeds the link capacity, and the renegotiation failure probability is the QOS measure of this service. Thus, our bufferless model is directly applicable to this problem.

2 Analysis

2.1 Basic Model

The network resource is a single bufferless link with capacity C. Calls arrive according to a Poisson process at rate λ and stay for an exponential distributed time with mean T if admitted. The calls have identical bandwidth requirement statistics, described by a continuous-time ergodic Markov fluid process. There are K states in the Markov process, and a call generates fluid at rate of μ_k in state k. The transition rate from state k to state l is r_{kl}. Let π_k be the steady-state probability that the source is at state k. Also, assume that when a call begins to transmit data, the Markov process is at steady state. The fluid processes of different calls are assumed to be independent.

A measurement-based admission control scheme decides whether to accept a new call based on the observed past history of calls that are currently in the system and possibly those that have already departed the system. Note that given any admission control scheme, the entire system is a stationary process. We are interested in two performance measures of a scheme: the steady-state probability of the event that the system overloads, i.e. the instantaneous aggregate fluid rate of calls in the system exceeds C, and the expected fraction of the total bandwidth utilized. For a given arrival rate, maximizing the latter is equivalent to minimizing the blocking probability. The success of the admission control scheme is evaluated by how well it meets the QOS-requirement (in terms of overload probability) and how close its bandwidth utilization is to that of the optimal scheme which knows the bandwidth requirements statistics *a priori*.

Suppose the statistics of each call is known a priori. Then one can estimate the overload probability given the number of calls in the system. For large number of calls, an accurate approximation is the Chernoff's estimate: the probability of overload when there are

m calls in the system is approximately

$$\exp(-mL^*(\frac{C}{m})) \tag{1}$$

where

$$L^*(\mu) = \sup_{r>0}[\mu r - L(\mu)], \qquad L(r) = \log(\sum_{i=1}^{K} \pi_k \exp(\mu_k r))$$

assuming that the average rate $\bar{\mu}$ of each call is less than $\frac{C}{m}$ so that overloading is indeed a rare event. This is the well known approximation for the tail of the distribution of the sum of n i.i.d. random variables having the stationary bandwidth requirement distribution of a call. To satisfy a desired overload probability p_{qos}, one can *a priori* compute the maximum number of calls m^* such that the above approximation is less than p_{qos}. The admission control is simply to accept a new call if there are less than m^* calls in the system, and reject it otherwise. The notion of *effective bandwidth* for bufferless system is developed using this approximation [3]; in this homogeneous case, the effective bandwidth of a call is simply $\frac{C}{m^*}$.

Now we should focus on the problem of interest, when no prior information about the statistics of the calls is available. The idea now is to estimate the statistics of the calls from observing their past empirical behavior. We will study a very simple scheme to focus on the essential issues. Motivated by the recent work of Gibbens et al. [7] on measurement-based admission control, we consider a scheme which is *memoryless*, i.e. every time a new call arrives, the scheme uses only information about the *current* state of the network in making the decision of accepting or rejecting the call. More specifically, the scheme determines the number of calls $n_k(t)$ that is currently generating data at rate c_k, for each k ($k = 1, \ldots, K$). This yields an empirical distribution $\{\hat{\pi}_k\}$ of bandwidth requirements for a typical call, where $\pi_k = \frac{n_k(t)}{x(t)}$ and $x(t)$ is the number of calls currently in the system at time t. The idea is to use $\{\hat{\pi}_k\}$ to estimate the distribution $\{\pi_k\}$ of the bandwidth requirements throughout the entire lifetime of a call. The admission criterion for the new call for a given threshold p_{qos} on the renegotiation failure probability is taken to be:

$$\exp\left(-\hat{L}^*\left(\frac{C}{x(t)+1}\right)\cdot(x(t)+1)\right) \le p_{qos} \tag{2}$$

and

$$\sum_{k=1}^{K} \hat{\pi}_k \mu_k < C \tag{3}$$

where

$$\hat{L}(r) = \log \sum_{k=1}^{K} \hat{\pi}_k \exp(\mu_k r) \qquad \hat{L}^*(\mu) = \max_{r>0}[\mu r - \hat{L}(r)]$$

Note that the left hand side of inequality (2) is the Chernoff approximation of the failure probability for a system with $n+1$ calls each with bandwidth requirements distributed as $\{\hat{\pi}_k\}$. Thus, this admission control scheme is of the certainty-equivalent type: the controller assumes that the measured values are the true parameters and acts like the optimal controller which has perfect knowledge of the values of those parameters.

2.2 Asymptotic Analysis

To get some insights into the dynamics of this scheme and to compare its performance to the scheme with perfect knowledge, we will consider a fluid approximation as well as a large deviations analysis for the above model. Such an analysis is relevant in the asymptotic regime of large link capacity and small loss probabilities.

Let the total capacity be scaled as $C \equiv n$, and the call arrival rate as $\frac{\lambda n}{T}$, where n is a large parameter. Thus, $n\lambda\bar{\mu}$ is the offered load, and $\lambda\bar{\mu}$ is the offered utilization, i.e. offered load normalized by the the link capacity. Keeping the ratio of offered load to system capacity fixed in this scaling means that we are assuming that the link capacity is sized so that it is of the same order of magnitude as the demand. Also, note that the offered load is *independent* of the average call holding time. We also set the quality-of-service requirement $p_{qos} = \exp(-n\delta_{qos})$, for a parameter δ_{qos} which specifies the exponential decay rate. Thus, the larger the system, the more stringent is the QOS requirement.

To explain ideas in the simplest terms, let us focus on two-state on-off sources. When it is on, the source transmits fluid at the peak rate of 1; when it is off, rate 0. Let p be the steady-state probability that a source is on, i.e. its mean rate is p. By an appropriate normalization of the time unit, we can assume that the on-to-off transition rate is $1-p$ and the off-to-on transition rate is p.

2.2.1 Known Statistics

Consider first the case when admission control is done using the Chernoff's estimate eqn. (1) of the overload probability based on perfect prior knowledge of the source statistics. For the n th system, let $G_n^*(t)$ and $H_n^*(t)$ be the total number of calls in the system and the number that are on at time t respectively. (The superscript '*' denotes quantities associated with scheme with perfect prior knowledge.) Define the scaled processes:

$$X_n^*(t) \equiv \frac{G_n^*(t)}{n}, \qquad Z_n^*(t) \equiv \frac{H_n^*(t)}{n}$$

and let $u^*(n,T) \equiv \mathbb{E}[Z_n^*(t)]$ be the average utilization.

(We indicate explicitly the dependence on the mean call holding time T.). Also, let $p^*(n,T)$ be the overload probability.

For the nth system, the maximum number of calls $m^*(n)$ admitted under the admission control is the largest m such that

$$mL^*(\frac{n}{m};p) \geq n\delta_{qos} \qquad \text{and} \qquad \frac{n}{m} \geq p$$

where L^* is large deviations rate function for on-off sources with mean rate p. It can be explicitly calculated as:

$$L^*(\mu;p) = \mu\log(\frac{\mu}{p}) + (1-\mu)\log(\frac{1-\mu}{1-p}) \qquad (4)$$

for $\mu \in [0,1]$ and equals ∞ otherwise. Note that $L^*(1;p) = -\log p$, correspond to the exponent of the probability that all sources are on.

As $n \to \infty$, it can be seen that $\frac{m^*(n)}{n}$ converges to $x^*(\delta_{qos})$ which is the largest $x \in [1, \frac{1}{p}]$ satisfying:

$$xL^*(\frac{1}{x};p) \geq \delta_{qos}$$

or equivalently,

$$-\log(xp) + (x-1)\log\frac{x-1}{x-xp} \geq \delta_{qos}$$

If $\delta_{qos} \leq -\log p$, then $x^*(\delta_{qos})$ is the unique solution to the equation

$$-\log(xp) + (x-1)\log\frac{x-1}{x-xp} = \delta_{qos}$$

On the other hand, if $\delta_{qos} > -\log p$, then $x^*(\delta_{qos}) = 1$. In this case, the QOS requirement is too stringent (relative to the burstiness of the traffic) such that peak-rate admission control has to be done. The system will never overload but also no statistical multiplexing gain is possible.

We now examine the behavior of the system as $n \to \infty$. It can be seen that typical behavior of the scaled system is well approximated by a *fluid limit* in which calls arrive at a deterministic rate of $\frac{\lambda}{T}$ and they transmit data at the mean rate of p once they are admitted into the system. The expected utilization of the system in this fluid limit depends on the offered rate λ:

$$\lim_{n\to\infty} u^*(n,T) \equiv u^*_\infty = \begin{cases} \lambda p & \text{if } \lambda < x^*(\delta_{qos}) \\ x^*(\delta_{qos})p & \text{if } \lambda > x^*(\delta_{qos}) \end{cases}$$

where $x^*(\delta_{qos})$ is computed as above. In the first case, the offered load is sufficiently low such that almost all calls are admitted. In the second case, the offered load is too high so that to maintain the desired QOS, a fraction of the offered calls has to be rejected. In this case, the number of calls in the system most of the time is close to the maximum possible for a given QOS. Moreover, since the number of calls in the system is no greater than $x^*(\delta_{qos})$ at all times, it is clear that by definition of $x^*(\delta_{qos})$, that the QOS is satisfied, i.e. for all T,

$$\lim_{n\to\infty} \frac{1}{n}\log p^*(n,T) \leq -\delta_{qos}\,.$$

In particular, in the case when $x^*(\delta_{qos}) = 1$, $p^*(n,T) = 0$ for all n and T. In the case when $x^*(\delta_{qos}) > 1$ and $\lambda > x^*(\delta_{qos})$, the QOS is satisfied exactly (in an exponential sense) for large call holding time T, i.e.

$$\lim_{T\to\infty}\lim_{n\to\infty} \frac{1}{n}\log p^*(n,T) = -\delta_{qos}\,.$$

2.2.2 Unknown Statistics

We now turn to an asymptotic analysis of the memoryless admission control scheme. In analogous to the previous analysis, for the nth system, let $G_n(t)$ and $H_n(t)$ be the total number of calls in the system and the number that are on at time t respectively, under the memoryless admission control scheme. Note that $(G_n(t), H_n(t))$ is the state of the system at time t. Define the scaled processes:

$$X_n(t) \equiv \frac{G_n(t)}{n}, \qquad Z_n(t) \equiv \frac{H_n(t)}{n}$$

and let $u(n,T) \equiv \mathbb{E}[Z_n(t)]$ be the average utilization. Also, let $p(n,T)$ be the overload probability. Our goal is to compare the average utilization and the overload probability to the corresponding quantities under the scheme with perfect prior knowledge, for n grows large.

The measurement-based scheme estimates the statistics of the sources under their current bandwidth requirements. In the case of on-off sources, the only parameter to estimate is p, the mean bandwidth requirement. The estimate $\hat{p}(t)$ of p at time t is $\hat{p}(t) = \frac{H_n(t)}{G_n(t)}$. Using the certainty-equivalent admission criterion (2), a call is admitted at time t if and only if

$$L^*(\frac{n}{G_n(t)+1};\hat{p}(t)) \geq n\delta_{qos} \qquad \text{and} \qquad H_n(t) < n \qquad (5)$$

where L^* is given by eqn. (4).

The scaled process $(X_n(t), Z_n(t))$ is a jump Markov process on the state space $\mathcal{S}^{(n)} \equiv \{(\frac{i}{n}, \frac{j}{n}) : 0 \leq j \leq i\}$. Define the sets

$$\mathcal{S}^{(n)}_a \equiv \left\{\left(\frac{i}{n}, \frac{j}{n}\right) \in \mathcal{S}^{(n)} :\right.$$

$$L^*\left(\frac{n}{i+1};\frac{j}{i}\right) \geq n\delta_{qos} \qquad \text{and} \qquad \frac{j}{n} < 1\Bigg\}$$

$$\mathcal{S}_r^{(n)} \equiv \mathcal{S}^{(n)} - \mathcal{S}_a^{(n)}$$

New calls are admitted if the state of the system is in $\mathcal{S}_a^{(n)}$ and rejected if in $\mathcal{S}_r^{(n)}$. For large n, the scaled processes $(X_n(t), Z_n(t))$ converge to a fluid limit which take values in a two dimensional continuous state space $\mathcal{S} \equiv \{(x,z) : 0 \leq z \leq x\}$. The regions $\mathcal{S}_a^{(n)}$ and $\mathcal{S}_r^{(n)}$ converge into two regions in $\mathcal{S}$ divided by a boundary. Define f on $\mathcal{S}$ as

$$f(x,z) = \begin{cases} -\log z + (x-1)\log\frac{x-1}{x-z} & \text{if } x \geq 1 \\ \infty & \text{else} \end{cases}$$

Then in the region $\mathcal{S}_a \equiv \{(x,z) \in \mathcal{S} : f(x,z) \geq \delta_{qos}, z < 1\}$, calls are accepted where in the complement $\mathcal{S}_r$, calls are rejected. We shall call $\mathcal{S}_a$ the *acceptance region* and $\mathcal{S}_r$ the *rejection region*. It should be noted that this partitioning depends both on the number of sources x in the system and the number z which are on, whereas in the case with known source statistics, it depends only on the number of sources x which are on (i.e. whether $x \geq x^*(\delta_{qos})$). This is the essence of the difference between a measurement-based scheme and one with known statistics.

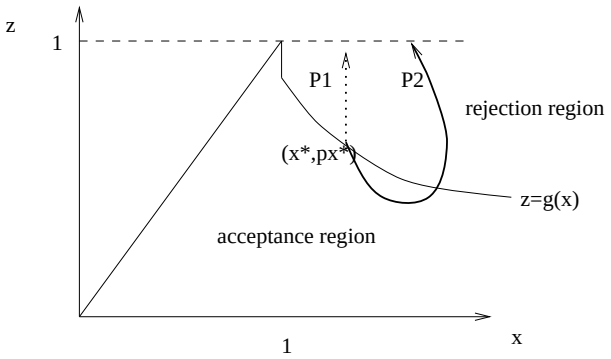

Figure 1: State space of fluid limit.

We shall now gather more information about the boundary. Clearly, $(x,z) \in S_a$ if $x \leq 1$. (No possibility of overload even if all sources in the system transmit at peak rate.) For $x \geq 1$, $(x,z) \in S_a$ if $f(x,z) \geq \delta_{qos}$ and $z < 1$. Further analysis reveals that one can find a function g such that this condition is equivalent to $z \leq g(x)$, where g satisfies $f(x,g(x)) = \delta_{qos}$ and has the properties that:1) $\lim_{x\to 1+} g(x) = \exp(-\delta_{qos})$; 2) $g(x)$ decreases monotonically with x; and 3) $\lim_{x\to\infty} g(x) = z^*$ where $z^* > 0$ satisfies

$$z^* - \log z^* - 1 = \delta_{qos}$$

Note that when there are x sources, $\frac{g(x)}{x}$ is the mean rate of each source such that the overload probability requirement is just satisfied. For the number of sources x close to 1 (link capacity), the only way to have an overload event is for all sources to transmit at peak rate. To satisfy the desired QOS, the mean rate of each source should then be $\exp(-\delta_{qos})$. Properties (2) and (3) imply that as there are more sources in the system, the maximum number of them that can be on and still permits new admissions decrease, and moreover approaches a non-zero limit as x goes to infinity. The situation is depicted in Figure 1. Note that the boundary consists of two segments: $z = g(x)$ for $x \in (1,\infty)$ and $x = 1$ for $z \in (\exp(-\delta_{qos}), 1)$. Also, the boundary partitions the state space into two connected regions.

In the fluid limit, the process follows a deterministic vector field starting from any initial point. In region $\mathcal{S}_a$, calls are accepted into the system at deterministic rate $\frac{\lambda}{T}$, while in region $\mathcal{S}_b$, the acceptance rate is zero. We now find the stable equilibrium points of the vector field and thus compute the average utilization in the fluid limit. First, one can easily see that the stable equilibrium points must lie on the line $z = px$, since p is the (true) mean rate of each source. As a consequence of the law of large numbers, the typical fraction of sources in the system that are on must be p. Moreover, there are no stable equilibrium points strictly inside the rejection region $\mathcal{S}_r$, since in this region, sources leave the system and no new sources enter. Thus, the stable equilibrium points must be strictly inside $\mathcal{S}_a$ or on the intersection of the boundary and the line $z = px$.

Suppose the line $z = px$ intersects the boundary at the point $(x^*(\delta_{qos}), px^*(\delta_{qos}))$. If $\delta_{qos} \leq -\log p$, the intersection is with the segment $z = g(x)$, and $x^*(\delta_{qos}) \geq 1$ satisfies the equation:

$$-\log(xp) + (x-1)\log\frac{x-1}{x-xp} = \delta_{qos}$$

If $\delta_{qos} > -\log p$, the intersection is with the segment $x = 1$, and $x^*(\delta_{qos}) = 1$. Note that $x^*(\delta_{qos})$ is precisely the maximum (scaled) number of calls that the scheme with perfect knowledge of the source statistics would admit. If $\lambda < x^*(\delta_{qos})$, the (unique) stable point $(\bar{x}, \bar{z})$ is $(\lambda, \lambda p)$, which lies strictly inside $\mathcal{S}_a$. In this case, almost all calls are accepted. If $\lambda \geq x^*(\delta_{qos})$, the stable equilibrium point $(\bar{x}, \bar{z})$ is on the boundary: $(x^*(\delta_{qos}), px^*(\delta_{qos}))$. In this case, a fraction of the calls are rejected. Thus, the stable equilibrium for the fluid limit under the measurement-based admission control is *exactly* the same as that under the scheme with perfect knowledge. This implies that the average utilization under the two schemes are asymp-

totically the same as $n \to \infty$, i.e.

$$\lim_{n\to\infty} \frac{u(n,T)}{u^*(n,T)} = 1 \tag{6}$$

We now turn to an analysis of the overload probability $p(n,T)$ under the memoryless admission control scheme. This corresponds to the event $\mathcal{O} = \{(x,z) : z > 1\}$. Since this set does not contain the stable equilibrium, it is a rare event for n large, with probability decaying exponentially with n. We have:

$$p(n,T) \approx \exp[-nI^*(T)] \tag{7}$$

where

$$I^*(T) = \inf_{(\vec{r}(\cdot),\tau)\in F} \int_0^\tau l(\vec{r}(t), \dot{\vec{r}}(t))dt \tag{8}$$

and

$$F = \{(\vec{r}(\cdot),\tau) : \vec{r}(0) = (\bar{x},\bar{z}), \vec{r}(\tau) \in \mathcal{O}\} \tag{9}$$

Thus, the exponent is given by the cost of the cheapest path starting from the equilibrium point and leading to an overload. This is in accordance with *Laplace's* principle: the probability of the most likely way for a rare event to happen is approximately the same as the probability of the rare event. Here, $l(x,z,\vec{v})$ is the *local rate function*, which gives the cost of traveling with velocity $\vec{v}$ at the state (x,z).

At present, we do not have a rigorous large deviations limit theorem justifying the approximation (7); thus, it is only a heuristic. To have such a theorem, we need to prove a large deviations principle for the scaled processes $(X_n(\cdot), Z_n(\cdot))$, which have discontinuous jump rates across a *curved* boundary. Dupuis and Ellis [9] have proved a large deviations principle for discrete-time processes with jump rates across a *straight* boundary, and Alanyali and Hajek [10] has proved one for continuos-time processes. The local rate function we use is a natural generalization of theirs.

Our main result about the overload probability performance is the following.

Theorem 2.1 *If* $x^*(\delta_{qos}) > 1$ *and* $\lambda \geq x^*(\delta_{qos})$ *then*

$$\lim_{T\to\infty} I^*(T) = \delta_{qos}$$

If $x^*(\delta_{qos}) = 1$ *or* $\lambda < x^*(\delta_{qos})$ *then*

$$\lim_{T\to\infty} I^*(T) \geq \delta_{qos}$$

The first case corresponds to the situation when the stable equilibrium is on the segment $z = g(x)$ of the boundary, whereas the second is when the equilibrium is on the segment $x = 1$ or in the interior of $\mathcal{S}_a$.

A large T means that the average call holding time is much longer than the time-scale of the fluctuation of the bit-rate of the sources, and thus implies a regime of separation of call and burst time-scales. Thus, for large T, the memoryless scheme satisfies the QOS requirement in terms of the exponential decay rate of the overload probability, and also performs asymptotically as well as the scheme which have perfect prior knowledge in terms of average utilization (from (6)). We can conclude that for large capacities and a separation of time-scales, the memoryless scheme is *asymptotically optimal*. This is somewhat surprising since the measurement-based scheme has no *a priori* knowledge about the sources and only uses information about the current state of the network to make admission decisions. Due to space limitation, we shall only outline the intuition behind the proof of this result.

If the statistics of the bandwidth requirements of the call were known, then the maximum number of calls allowed in the system, approximately $nx^*(\delta_{qos})$, can be determined, such that the overload probability when there are that many calls in the system satisfy the QOS requirements. The measurement-based scheme essentially tries to estimate $nx^*(\delta_{qos})$ based on observing the current state of the network. Due to statistical fluctuations in the bandwidth requirements of the calls, the estimate may be too high due to atypically large number of calls being off. This may result in the control scheme making a mistake in accepting a new call which it should have rejected. However, the key point is that for a large system, the overload probability will deteriorate significantly only if the number of calls in the system is considerably larger than $nx^*(\delta_{qos})$; thus, making an occasional mistake in admitting a call is not fatal. Rather, making a succession of mistakes in accepting calls will lead to serious degradation in performance. However, because of the separation of the time-scales of call arrivals and fluctuations of bandwidth requirements, it is very unlikely for the state of the network to remain atypical over the period of time for the arrivals of a succession of calls. In fact, it can be shown that the probability of making a succession of mistakes is much smaller than that of overload due to having a typical number $nx^*(\delta_{qos})$ of calls in the system but atypically large number of calls being on.

This intuition can be translated into a proof of Theorem 2.1 via an analysis of the costs of the paths in F (the set of paths leading to overload, as defined in (8)) as T grows large. Two candidate paths are depicted in Fig. 1. Suppose the stable equilibrium is $(x^*(\delta_{qos}), px^*(\delta_{qos}))$ on the segment $z = g(x)$ of the boundary. Consider the straight-line path $P_1 =$

$\{(x^*(\delta_{qos}), z) : z \in [px^*, 1]\}$. This corresponds to the event that the measurement-based scheme estimates the maximum admissible number of calls $nx^*(\delta_{qos})$ correctly, and overload occurs during call rejection phase with $nx^*(\delta_{qos})$ calls in the system but an atypical fraction of them turning on. Also, no call departs the system during this phase. We observe that as the average call holding time increases, the cost of this path approaches δ_{qos}, since the probability of overload when there are $nx^*(\delta_{qos})$ calls *permanently* in the system is approximately $\exp(-n\delta_{qos})$.

Now let us consider alternate paths of which an example is P_2, shown in Fig. 1. These paths consists of two phases. In the first phase, calls are admitted by mistake due to an atypically large fraction of the existing calls in the system turning off. This phase ends at time τ_1, resulting in nx_1 calls in the system, where $x_1 > x^*(\delta_{qos})$. In the second phase, new calls are rejected, but many of the existing calls turn back on, resulting in overload. Let us consider the costs of these paths in the regime of large holding time T. First, observe that the call arrival rate is $\frac{\lambda}{T} \cdot n$; hence, for those paths with $\tau_1 << T$, the cost is high because it is unlikely to have $n(x_1 - x^*(\delta_{qos}))$ arrivals in duration τ_1. On the other hand, for those paths with long duration τ_1, the cost is also high because it is unlikely for the calls in the system to be in an atypical state for so long. Using similar logic, it can be shown that the costs of all paths bounded away from P_1 go to infinity uniformly as average holding time $T \to \infty$. This implies that the cost of the cheapest path approaches δ_{qos} as $T \to \infty$.

We have only considered the case when the equilibria is on the segment $z = g(x)$. For the cases when the stable equilibrium point is on the segment $x = 1$ or in the interior of $\mathcal{S}_a$, one can show that the cost of the cheapest path must in fact be greater than δ_{qos}.

So far, we have focused on only on-off sources for ease and concreteness of exposition. The analysis and the Theorem 2.1 carry over to the general case of Markov fluid sources with general state space.

3 Experiments

In the previous section, we obtained some analytical insights into the performance of a memoryless measurement-based admission control scheme using fluid and large deviations analysis. Such results are valid in the asymptotic regime of large link capacities and for Markov sources. In this section, we will complement the theoretical investigations with simulation results for finite-size systems and on both real and synthetic traffic sources. In particular, we study the following issues: 1) the impact of parameters such as link capacity and offered load on the performance of the memoryless scheme; 2) the appropriateness of using Markov models in studying the performance of measurement-based schemes on real traffic sources.

3.1 Basic Simulation Set-up

Our simulation set-up is based on the Renegotiated CBR (RCBR) service [8]. Each call is represented as an RCBR schedule, i.e., as a sequence of intervals over which the bandwidth consumed by the call is assumed constant. This schedule represents the slow time-scale dynamics of the traffic source. Times at which the bandwidth changes correspond to instants when the source renegotiates for a new CBR rate. In this service, renegotiation fails when there is insufficient capacity in the link to cater for a requested bandwidth. Thus, this corresponds to the overload event in our bufferless model. For a real traffic source, we use a two-hour long MPEG-1 Star Wars trace [11], from which the RCBR schedule is computed by the off-line optimization algorithm presented in [8]. This schedule has a peak-to-mean bandwidth ratio of about 5. One important advantage of using the RCBR set-up is efficiency: we do not simulate on a per video frame basis but rather per renegotiation event. Since the particular schedule we use has about one renegotiation every 20 seconds on the average, this leads to an increased efficiency of 2 to 3 orders of magnitude.

In the simulations, calls arrive according to a Poisson process. We measure both the average network utilization and the renegotiation overload probability. We consider that the system state between time instants separated by more than twice the average call length is independent. Therefore, each interval of that size provides us with one sample for these metrics. We collect samples until the 95%-confidence interval for both metrics is sufficiently small with respect to the estimated value (within $\pm$ 30% of the estimated value.) For the overload probability, we also stop if the target overload probability of 10^{-5} lies to the right of the confidence interval, i.e., if we are confident that the actual failure probability is lower than the target. This is necessary in order to terminate simulations within reasonable time when the observed overload probability is very low (e.g. 10^{-8}).

3.2 Performance of Memoryless Scheme

The first set of experiments evaluates the impact of the link capacity and of the offered load on the performance of the memoryless scheme as defined by (2), where the desired overload probability p_{qos} is chosen to be 10^{-5}.

Arriving calls are randomly phase-shifted versions of the *entire* RCBR schedule for the two-hour Star Wars trace. Fig. 2 and 3 show the overload probability and the average network utilization respectively, plotted as a function of link capacity, expressed as a multiple of the call average rate, and the offered utilization, which is the offered load normalized by the

link capacity. The network utilization is normalized to the utilization that is achieved with the same parameter values when call admission is performed based on the Chernoff approximation (1) and perfect knowledge of the call's marginal distribution. Thus, for example, a value of 1 means that the memoryless scheme does equally well as the scheme with perfect knowledge.

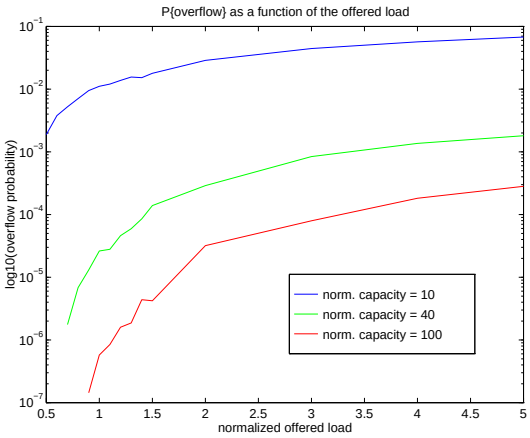

Figure 2: Overload probability for call duration equals length of entire trace

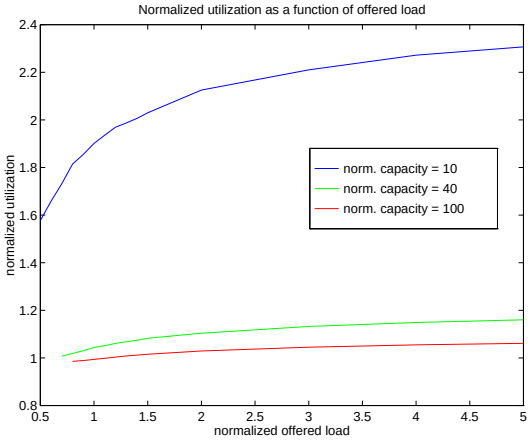

Figure 3: Utilization for call duration equals length of entire trace

It can be seen from Fig. 2 that the memoryless scheme performs poorly for small link capacities, say less than 60. The overload probability is much larger than the target of 10^{-5}.

3.3 Equivalent Markov Model

Our analysis in the previous section is based on a Markov model of the traffic source. It is therefore of interest to see how well a Markov model of a real traffic source can predict the performance of a measurement-based admission control scheme on the actual traffic source. In particular, we would like to investigate if the *long-range dependence* in the correlation structure, postulated recently to be present in many types of traffic, has any impact on the performance.

For a real traffic source, such as the Star Wars RCBR schedule, we obtain an *Equivalent Markov* model in the following way. We match both the marginal distribution π_i of the bandwidth, as well as the average residence time per bandwidth level τ_i. The bandwidth is described by a random process that stays at one of the bandwidth levels μ_i for an exponentially distributed time with mean τ_i. It then jumps to a new bandwidth level μ_j with probability f_j. The lengths of the intervals and the bandwidth levels within intervals are individually and mutually independent. To match the marginal distribution, we want

$$\pi_i = \frac{f_i \tau_i}{\sum_{j=1}^{M} f_j \tau_j}. \tag{10}$$

Thus,

$$f_i = \left(\sum_{j=1}^{M} \frac{\pi_j}{\tau_j} \right) \frac{\pi_i}{\tau_i}. \tag{11}$$

Note that the actual expected residence time per bandwidth level is slightly higher than τ_i, as with probability f_i, the model selects the same bandwidth in two consecutive intervals[1]. We thus have a traffic model that allows us to match the first-order characteristics of an RCBR trace, namely the marginal distribution of the bandwidth, and the mean residence time per bandwidth level, but which exhibits an exponentially decreasing correlation function, and therefore no long-range dependence (LRD).

We now evaluate the performance of the memoryless scheme when the Equivalent Markov model of the Star Wars RCBR scheme is used. Figs. 4 and 5 show the overload probability and utilization respectively in the scenario when the bandwidth processes of calls are generated independently and according to the Equivalent Markov model of the Star-Wars schedule, and the calls stay for the same duration as the entire trace (approx. 2 hours). Comparing them to the corresponding Figs. 2 and 3 when the actual RCBR schedule is used, we see that the performance is quite similar in the two scenarios.

It is interesting that [11] concluded that this traffic exhibits long-range dependence, based on a statistical analysis of the Star Wars video trace. Our results indicate that such long-range dependence has little impact on the performance of the measurement-based admission control scheme considered here.

[1]The actual expected residence time is $\tau_i/(1-f_i)$.

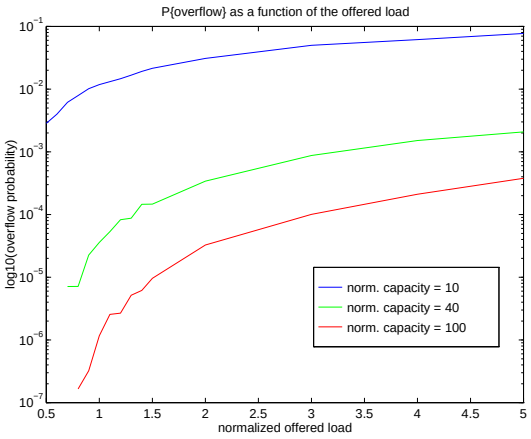

Figure 4: Overload probability for equivalent Markov model

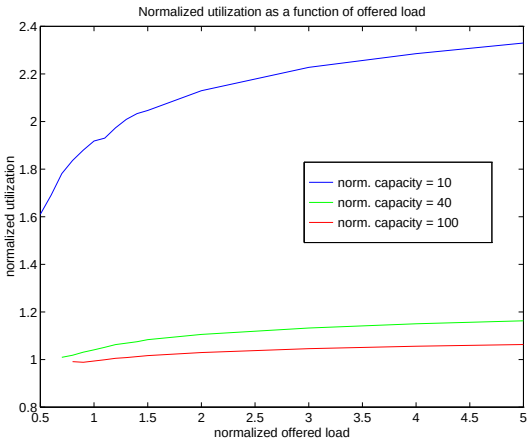

Figure 5: Utilization for equivalent Markov model

4 Conclusions

To obtain insights on the problem of measurement-based admission control, we have focused on a simple scheme, which has the main characteristics of being memoryless (admission decision based only on the current network state) and certainty-equivalent (measured statistics taken to be true values of the unknown parameters). We have studied this scheme through a combination of asymptotic analysis and simulations on real and synthetic traffic sources. The main theoretical result is that the scheme is asymptotically optimal in the regime of large link capacity and a separation of call and burst time-scales, in the sense of attaining the performance of the optimal scheme which has perfect knowledge of the traffic statistics. The performance is measured in terms of the ability to maximize the utilization of the network while maintaining the overload probability below a desired threshold. Our simulation results on the Star Wars trace indicate that the scheme only works well for large link capacities (> 100 times the mean rate of a call) and not too high offered load (no more than 2 to 3 times the link capacity.) For small link capacities, it makes too many admission mistakes due to measurement errors. Wars trace serves as a good predictor of the performance of the admission control scheme on the actual traffic, despite the claimed presence of long-range dependence in the latter. Our current work focuses on finding schemes which have better performance under high offered load.

References

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

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

[3] J. Hui, "Resource allocation for broadband networks," *IEEE Journal on Selected Areas of Communications*, Dec. 1988.

[4] C. Courcoubetis, G. Kesidis, A. Ridder, J. Walrand, and R. Weber, "Admission control and routing in ATM networks using inferences from measured buffer occupancy," *to be published.*

[5] H. Saito and K. Shiomoto, "Dynamic call admission control in ATM networks," *IEEE Journal on Selected Areas of Communications*, vol. 9, pp. 982–989, 1991.

[6] I. Hsu and J. Walrand, "Dynamic bandwidth allocation for ATM switches," *submitted to J. of Applied Probability.*

[7] R. Gibbens, F. Kelly, and P. Key, "A decision-theoretic approach to call admission control in ATM networks," *IEEE Journal on Selected Areas of Communications*, pp. 1101–1114, Aug. 1995.

[8] M. Grossglauser, S. Keshav, and D. Tse, "RCBR: A Simple and Efficient Service for Multiple Time-Scale Traffic," in *Proc. ACM Sigcomm '95*, (Boston, Mass.), pp. 219–230, August 1995.

[9] P. Depuis and R. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations.* to be published.

[10] M. Alanyali and B. Hajek, "On large deviations of Markov processes with discontinuous statistics," *submitted to* Annals of Applied Probability, 1996.

[11] M. W. Garrett and W. Willinger, "Analysis, Modeling and Generation of Self-Similar VBR Video Traffic," in *ACM Sigcomm '94*, (London, UK), pp. 269–280, August 1994.