Admission Control for Statistical QoS: Theory and Practice

Edward W. Knightly, Rice University
Ness B. Shroff, Purdue University

Abstract

In networks that support quality of service, an admission control algorithm determines whether or not a new traffic flow can be admitted to the network such that all users will receive their required performance. Such an algorithm is a key component of future multiservice networks because it determines the extent to which network resources are utilized and whether the promised QoS parameters are actually delivered. Our goals in this article are threefold. First, we describe and classify a broad set of proposed admission control algorithms. Second, we evaluate the accuracy of these algorithms via experiments using both on-off sources and long traces of compressed video; we compare the admissible regions and QoS parameters predicted by our implementations of the algorithms with those obtained from trace-driven simulations. Finally, we identify the key aspects of an admission control algorithm necessary for achieving a high degree of accuracy and hence a high statistical multiplexing gain.

Provisioning network resources to meet the quality of service (QoS) demands of bursty traffic sources is a key issue for future multiservice networks. Such resource provisioning may be realized by an admission control algorithm, which has the function of limiting the number of traffic flows in a class such that the required QoS constraints can be satisfied. The design of admission control algorithms has important consequences for network performance, since an algorithm that unnecessarily denies access to flows that could have been successfully admitted will underutilize network resources; similarly, an algorithm that incorrectly admits too many flows will induce QoS violations.

Unlike a deterministic service [2], a statistical or soft real-time service associates a small *violation* probability with delay and throughput bounds as needed to obtain a utilization gain over a purely worst-case approach. Developing resource allocation schemes for a statistical service has proven particularly challenging due to both the multiple-timescale characteristics of many multimedia applications [3–5], as well as potential intractabilities arising from complex interactions among traffic flows and the shared multiplexer.

Our goals in this article are threefold. First, we describe a broad set of admission control algorithms from the literature which we divide into the following five classes based on:

- Average and peak rate combinatorics [6, 7]
- Additive effective bandwidths [8–11]
- Engineering the "loss curve" [12–15]
- Maximum variance approaches [16–18]
- Refinements of effective bandwidths using large deviations theory

Second, we perform a large number of experiments to evaluate the accuracy and effectiveness of these admission control algorithms under realistic workloads, namely, 30-minute traces of variable-rate MPEG-compressed video and exponential on-off sources commonly used to model voice traffic. To achieve this, we first implement a number of algorithms from the aforementioned classes, and determine their respective admissible regions for various traffic mixes and QoS parameters. We then simulate a 45 Mb/s multiplexer servicing the same traffic mix, with each flow's arrival sequence given by either a video trace with a random start time, or an on-off source. For each combination of traffic flows and a particular buffer size, we measure the flows' resulting performance parameters. By comparing the measured admissible regions with those predicted by the algorithms, we assess an algorithm's *accuracy* (i.e., its effectiveness in predicting QoS parameters and controlling the admissible region).

Finally, from our experimental results, we identify the components of an admission control test essential to achieving a high degree of accuracy, and find that

- The assumption of a bufferless multiplexer has a significant utilization penalty.
- An algorithm must exhibit economies of scale in the number of multiplexed flows.

Excerpts of this article appear in [1].

- Observed shapes of the "loss curve" can be quite different than the commonly assumed exponential relationship.
- The traffic model, or parameters used to describe the properties of traffic flows to the network, requires more information than is currently standardized.
- An algorithm's accuracy with exponential on-off sources does not ensure accuracy with bursty compressed video sources.

In addition to the above classes of admission control algorithms, several other approaches have been developed, including measurement-based algorithms which control the admissible region based on aggregate traffic measurements [19–22], enforceable statistical services which provision resources based on worst-case statistics of policed traffic flows [23–25], and algorithms for special-purpose systems such as video on demand [26, 27]. While a review of such schemes is beyond the scope of this article, we note that many of these approaches build on the theories and techniques we *do* consider, so our conclusions may provide guidelines for evaluating these schemes as well.

The remainder of this article is organized as follows. First, we overview five classes of admission control algorithms. Next, we describe experimental results obtained from trace-driven simulations and our implementations of the admission control algorithms. Finally, we discuss the aspects of an algorithm most critical for achieving a high degree of accuracy.

Admission Control Tests for Statistical Service

In this section we describe five classes of admission control tests that have been proposed for providing statistical QoS guarantees in multiservice networks. While these classes do not encompass all proposed schemes, they do provide broad coverage of the techniques applied to admission control. For QoS metrics, we consider two variations of what is commonly referred to as *loss probability*. First, we denote the *tail probability* of the queue length distribution $P(Q > B)$, which refers to the fraction of time an infinite buffer queue's occupancy exceeds B. Second, we denote the loss probability P_l, which refers to the fraction of bits dropped by a queue that has finite buffer space B.

Throughout, we denote the link capacity C and the buffer size B, and denote the arrivals of traffic flow j in the interval $[s, t]$ as $A_j[s, t]$.

Average/Peak Rate Combinatorics

In [7], source j is characterized by its peak rate $r_{pk,j}$ and average rate $r_{av,j}$. Assuming an on-off source that either transmits at its peak rate or is idle, the probability that the source is *on* is given by $p_{on,j} = r_{av,j}/r_{pk,j}$ and its rate distribution is given by

$$f_j(x) = \begin{cases} 1 - p_{on,j} & x = 0 \\ p_{on,j} & x = r_{pk,j} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Using this rate distribution, an admission control algorithm is designed that approximates the packet loss probability for a bufferless multiplexer: in a bufferless multiplexer, packet loss occurs whenever the aggregate input arrival rate exceeds the link capacity. Since the distribution of the aggregate arrival rate of the multiplexed sources is given by a convolution of the individual $f_j(x)$s, Lee *et al.* focus on efficient computation of the aggregate arrival rate distribution and subsequently the loss probability [7].

In [6], traffic flow j is also characterized by its peak and average rate. In contrast to [7], in which $r_{av,j}$ represents the long-term average rate, in [6] it refers to the worst-case rate over any interval of length I_j. That is, source j is constrained to send no more than $r_{av,j} \cdot I_j$ packets during any interval of length I_j (changing [6]'s notation for consistency). For Earliest Deadline First schedulers, Ferrari shows how to compute the probability of delay-bound violation by examining combinations of active flows (flows which are *on* with probability $p_{on,j}$) that may cause a delay-bound violation, and by summing their respective probabilities.

In this article we evaluate the test of [7], which we refer to as the *average/peak test*.

Additive Effective Bandwidths

Various *effective bandwidth* admission control tests have been proposed in the literature, including [8–11]. In such schemes, each flow independently reserves a particular bandwidth between its average and peak rates. This bandwidth, the *effective bandwidth*, is a function of the required loss probability P_l and the particular flow's stochastic properties (e.g., autocorrelation function, or peak and average rate together with mean burst duration). Once the effective bandwidth of flow j is determined, which we denote $E_j(P_l)$ (or, equivalently, $E_j(P(Q > B))$), the admission control test requires that

$$\sum_{j=1}^{N} E_j(P_l) < C, \quad (2)$$

where N is the number of multiplexed flows. Effective-bandwidth-type results have been devised using several interrelated techniques, including eigenvalue decomposition of Markovian flows [9], large deviations theory [8, 11], and the theory of envelope processes [28].

For example, [29] shows how the buffer occupancy distribution for Markov modulated fluid sources can be decomposed according to the eigenvalues of the aggregate Markovian arrival process. Since the overall tail probability $P(Q > B)$ or loss probability P_l may be expressed as a sum of exponential terms, the largest eigenvalue, which we denote δ, dominates with asymptotically large buffers. In other words, it has been found that the loss probability (or tail probability) satisfies the following relationship for some constant K:

$$P_l \sim Ke^{-\delta B}, \quad (3)$$

where the similarity relationship $f(x) \sim g(x)$ of two functions $f(\cdot)$ and $g(\cdot)$ indicates that $\lim_{x\to\infty} f(x)/g(x) = 1$. In the additive effective bandwidth literature, the preterm K is approximated by 1,[1] or $P_l \approx e^{-\delta B}$. The admission control test amounts to ensuring that δ (the dominant eigenvalue) is large enough to meet the required loss probability constraint. For example, for on-off sources, the effective bandwidth can be computed using Eq. 2 of [10].

Here, we evaluate the effective bandwidth test of [8], which has

$$E_j(P_l) = r_{av,j} + \frac{\delta\gamma_j}{2B} \quad (4)$$

where

$$\gamma_j = \lim_{t\to\infty} \frac{1}{t} \operatorname{var}(A_j[0,t]) \quad (5)$$

is the arrival sequence's index of dispersion and

$$\delta = -\frac{\log(P_l)}{B}.$$

Note that the effective bandwidth of a flow is independent

[1] *While the preterm* K *is different for tail and loss probability, since effective bandwidth approximates* K *by 1, it does not distinguish between the two.*

of the properties of all other traffic flows as well as the number of sources N and the link capacity C: it is determined only by the stochastic properties of the flow itself, the required loss probability P_l, and the buffer size B.

Engineering the Loss Curve

We refer to a *loss curve* as the relationship between loss probability and buffer size, which for additive effective bandwidths is exponential and given by $P_l \approx e^{-\delta B}$. Since the $e^{-\delta B}$ approximation can be conservative for reasons outlined below, several techniques have been proposed which seek to engineer the shape of the loss curve to better reflect experimentally observed relationships.

The approach in [15] stems from a simple observation based on the authors' work in video modeling [30–32]. When the input traffic is highly correlated, as in the case of JPEG-encoded video, a quasi-stationary approximation called the *histogram model* [32] (or *generalized histogram model* [31, 33]) was found to be quite accurate because P_l does not significantly decrease when B is increased beyond a certain range (referred to as the *cell region*). However, when applying the histogram model to other less correlated sources, it was found that the model could not predict the loss behavior well for large buffer sizes. To overcome this difficulty, in [15] the authors use the histogram model for the cell region (small buffers) and a single exponential approximation for the burst region (large buffers), with the cutoff point found by equating the slopes of the two regions. This approach, which we refer to as the *hybrid scheme*, can be used when the arrival process is modeled as a general Markov modulated arrival process, and in the specific case of the Markov modulated *fluid* source, P_l is expressed by a single exponential, as in Eq. 3. For example, for an aggregate Markov modulated fluid source with mean arrival rate λ, stationary probability π_i of being in state i, and arrival rate λ corresponding to state i, the probability of loss is given by

$$P_l = \frac{1}{\lambda} \sum_{\substack{i \\ \lambda_i > C}} \lambda_i \pi_i \left(1 - \frac{C}{\lambda_i}\right) e^{-\delta B}. \tag{6}$$

In [14], Elwalid *et al.* observed that for Markov modulated fluid sources, the loss curve approximation of Eq. 3 could be improved by approximating the asymptotic constant K by the loss probability in a bufferless multiplexer as estimated by Chernoff's theorem, and using δ as the same dominant eigenvalue for Markovian sources as in the effective bandwidth result. In particular, denoting R_j as a random variable with the steady-state rate distribution of source j (e.g., R_j's distribution is given by Eq. 1 for on-off sources), K is given by[2]

$$K = K' \exp\left(-\sup_{s \ge 0} \{ sC - \sum_j \log E e^{sR_j} \} \right), \tag{7}$$

where K' is a further refining term given by Eq. 55 of [14]. K' is based on Bahadhur Rao asymptotics, and we include it in our implementation of this approach.

Consequently, since K can in practice be substantially less than 1 (its approximated value for additive effective bandwidths), algorithms utilizing this term can have improved accuracy over effective bandwidth tests.

Maximum-Variance-Based Approaches

We classify the next group of admission control algorithms as maximum variance (MV) approaches. Defining X_t as

$$X_t = \sum_j A_j[s-t, s] - Ct, \tag{8}$$

the tail probability is given by [34]

$$P(Q > B) = P\left(\sup_{t \ge 0} X_t > B \right). \tag{9}$$

MV approaches are based on the observation that if X_t is Gaussian, one can derive accurate bounds and approximations to the right side of Eq. 9. Since X_t is composed of the aggregate arrivals from a large number of sources, this may be a reasonable assumption in high-speed networks. For Gaussian X_t, the normalized maximum variance of X_t, given by

$$\sigma_B^2 = \max_t \frac{\mathrm{var}\{X_t\}}{(B - E(X_t))^2}, \tag{10}$$

plays an important role in evaluating the maximum probability $P(\sup_{t \ge 0} X_t > B)$. For example, the time-instant $\hat{t}$ at which

$$\frac{\mathrm{var}\{X_t\}}{(B - E(X_t))^2}$$

achieves its maximum value σ_B^2 is the same time-instant at which $P(X_t > B)$ achieves its maximum value. Hence, one approximation used to estimate Eq. 9 is

$$P\left(\sup_{t \ge 0} X_t > B \right) \approx \max_t P(X_t > B). \tag{11}$$

It can easily be seen that the resulting approximation $\max_t P(X_t > B)$ is a lower bound to the tail probability of the buffer occupancy distribution, and is quite easy to compute. Choe and Shroff [16] have shown through an extensive empirical study that Eq. 11 is quite accurate when the arrival process can be effectively modeled as a Gaussian process. In [16], the authors have also developed an asymptotic (in terms of B) upper bound based on the normalized maximum variance σ_B^2, which they refer to as the maximum variance asymptotic (MVA) upper bound. The authors show that the MVA upper bound in practice behaves like a global upper bound, and together with the lower bound encapsulates the tail probability within a narrow envelope. More recently, the MVA approach has been extended by Kim and Shroff [17] to also estimate P_l by normalizing the MVA upper bound by the exact probability of loss in a bufferless system.

Knightly proposed a related technique using the perspective of stochastic traffic envelopes [18]. A stochastic traffic envelope bounds some statistical properties of $A_j[s, s+t]$ as a function of the interval length t. In particular, traffic is characterized via a *rate-variance envelope* defined by

$$RV_j(t) \ge \mathrm{var}\left(\frac{A_j[s, s+t]}{t} \right), \tag{12}$$

which describes a flow's second moment correlation structure. Based on the flows' $RV_j(t)$ characterizations as well as their mean rates, admission control tests are devised whereby the stochastic envelope of the aggregate traffic is approximated with a Gaussian envelope with variance $\Sigma_j t^2 RV_j(t)$ over intervals of length t. The envelope-based tests then consider the maximal buffer overflow probability in all interval lengths up to the maximal busy period. Consequently, the shape of the loss curve is determined by the properties of the aggregate envelope.

Refinements to Effective Bandwidths and Large Deviations

Two key shortcomings of the additive effective bandwidth approach are that the result is not applicable to traffic sources which exhibit long range dependence; and, by adding the

[2] *The supremum can be replaced by max when the maximum exists; otherwise, it is least upper bound.*

bandwidth requirements of sources, the effects of economies of scale with a large number of sources are not exploited.

An alternative definition of effective bandwidth is given by [35–37]

$$E_j(s,t) = \frac{1}{st} \log E e^{sA_j[0,t]} \quad (13)$$

so that the tail probability of the queue length distribution satisfies

$$\lim_{N \to \infty} \frac{1}{N} \log P(Q > B) = \sup_t \inf_s \left[st \sum_j \rho_j E_j(s,t) - s(b+ct) \right], \quad (14)$$

scaling resources to $C = Nc$, $B = Nb$, and $N\rho_j$ sources of type j.

This result is based on large deviations theory, which is also applied in several other approaches above and has been widely used in providing general results on the asymptotic behavior of $\log P(Q > B)$ [36, 38, 39]. Here, we briefly review related large deviations techniques. In [39], for example, Glynn and Whitt show that for a large class of stochastic processes

$$\log P(Q > B) \sim -\delta B. \quad (15)$$

However, for many important types of processes, such as self-similar or other long-range dependent processes [40, 41], the tail probability may not be exponential; more generally, even Eq. 15 may not hold. To address this problem, Duffield and O'Connell exploit the generality of large deviations techniques [36] and extended the above result through an elegant scaling technique to obtain

$$\log P(Q > B) \sim -g(B) \quad (16)$$

where $g(B)$ is some increasing function of B, which may not be linear in B. However, the significant generality of this result does come at a cost, namely, poor "resolution," since the similarity relation given by Eq. 16 captures only the leading (most rapidly growing) term of $\log P(Q > B)$. For example, if $g(B) = B$ satisfies Eq. 16, then $g(B) = B + \sqrt{B}$ also satisfies Eq. 16, even though it is a very different function of B. Therefore, in general, approximations for $P(Q > B)$ based on Eq. 16 should be used with some caution, since Eq. 16 provides relatively weak theoretical support to the asymptotic behavior of these approximations. (Note that compared to Eq. 3, which shows similarity, the large deviation results only show log-similarity.)

Recent work has focused on the asymptotic behavior of $P(Q > B)$ when the number of sources, queue size, and service rate are all proportionally sent to infinity (e.g., [38]). This limit is quite different from the one in Eq. 16. However, such results have generated approximations such as the one in [38] which, when applied to Gaussian processes, produce the same expression as the MVA upper bound discussed above. This approach also allows for the use of the Bahadhur Rao asymptotics (or Local Central Limit Theorem, also used earlier) to strengthen the log-asymptotic results considerably [42, 43]. In [43] the approximation based on the Bahadhur Rao asymptotics results in the same expression as the MVA lower bound which we consider in this article.

We make the following observations about this class of approaches. First, these effective bandwidths are not "additive" in the same sense as the previous approaches in that the resources required for a particular source depend on the properties of all other sources (the dependencies are through the parameters s and t). Thus, these approaches achieve economies of scale in the number of multiplexed sources and, as we will show in the following section, are considerably more accurate than the previous additive effective bandwidth techniques. Second, we observe that without further assumptions on the traffic flows, admission control tests in this class can be computationally expensive in their calculation of the supremums in Eq. 14. Nevertheless, we show that such approaches can work quite well empirically; moreover, they can address cases of non-Gaussian traffic which the MV approaches discussed above cannot.

Experimental Evaluation of Admission Control Tests

In this section we evaluate the accuracy of the aforementioned admission control algorithms by performing a set of simulation and admission control experiments. We consider two types of traffic:

- Actual traces of MPEG-compressed video
- Markov modulated on–off sources, with on and off times distributed according to standard voice models

We consider various scenarios with different loads, QoS parameters, and so on, and compare the actual admissible regions and QoS values obtained in simulations with those predicted by the admission control tests.

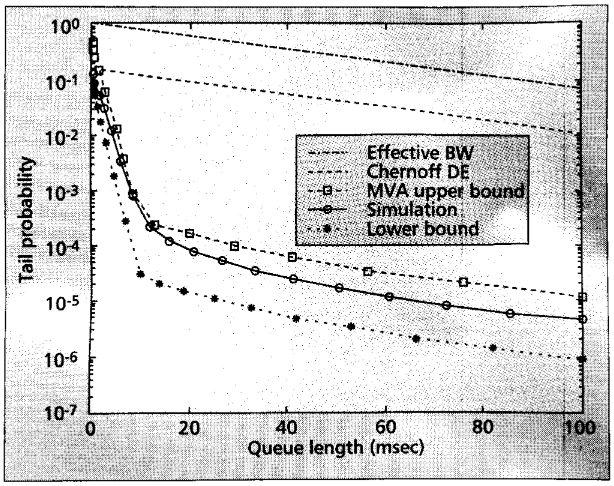
Throughout the experiments, we focus on the following performance metrics. The first is the average utilization of the link, which is the total average rate of all flows divided by the link capacity. For the simulation, this average utilization is also the total number of bits transmitted by the sources in the simulation, divided by the total number of bits that the server can transmit over the duration of the simulation (the link capacity multiplied by the simulation time). Our second performance metric is $P(Q > B)$, the tail of the queue length distribution. For simulations, this corresponds to the fraction of time an infinite buffer queue's occupancy exceeds B. The third performance metric is the empirical fraction of packets that are dropped due to buffer overflow in a finite buffer queue with maximum buffer space B. We denote this measure of loss probability by P_l. In both cases, we consider a range of buffer sizes B which have a corresponding statistically guaranteed delay bound $d = B/C$.

Experimental Evaluation for Video Traces

Here we consider a 30-min trace of MPEG-compressed video which exhibits statistical properties characteristic of long-range-dependent traffic [18].

Scenario — The video is taken from an action movie, digitized to 384 x 288 pels and compressed at 24 frames/s using the MPEG-1 compression algorithm with frame pattern *IBBPBBPBBPBB* [44]. For the simulations, we consider each frame to be transmitted at a constant rate over the frame time 1/24th of a second.

For the trace-driven simulations, a simulation cycle runs until all sources transmit their entire trace twice, with the traces wrapped around to the beginning when they reach the end. The first run through the traces is discarded as a transient, and statistics are collected on the second pass through the traces. In a particular scenario, N flows are multiplexed on a 45 Mb/s link, with each flow's arrival pattern given by the movie trace, and a start time chosen uniformly over the length of the trace. For the purpose of obtaining small confidence intervals at low loss probabilities, 10,000 such simulations were performed, each with independent start times. We calculate 95 percent confidence intervals for each probability estimated via simulation using the method of batch mean [45]. However, since the confidence intervals are so small (the max and min of the error bars virtually overlap each other), we do not show them in the figures.



■ **Figure 1.** *Tail probability comparisons for video sources.*

Tail Probability for Multiplexed Video Flows — The experiments reported in Figs. 1 and 2 (for the tail and loss probabilities, respectively) are based on the above simulation scenario for a fixed utilization of 84 percent, which corresponds to 69 video flows on the 45 Mb/s link.

Figure 1 depicts the tail probability versus the delay that a packet experiences in queue. In this scenario, even if a multiplexer delays a packet beyond its deadline, the packet is not dropped, but held in the buffer in case its delay could be made up downstream.

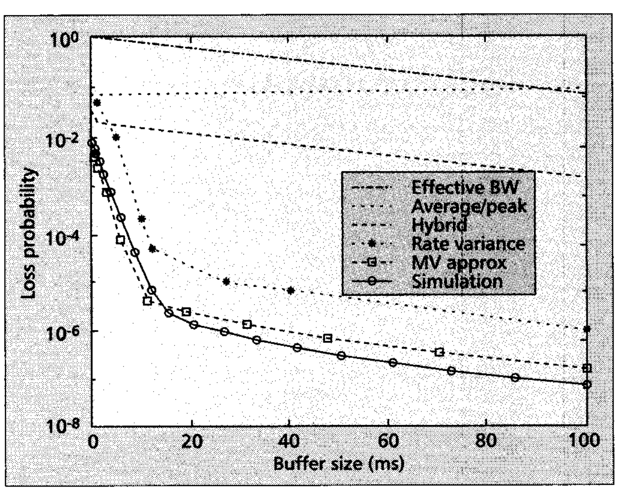
The curve labeled "Simulation" reports the actual fraction of time the buffer exceeds the threshold B or the equivalent delay bound. Notice that this curve drops sharply until buffer sizes reach approximately 10 ms, after which it flattens considerably, indicating significant benefits of adding buffer space to a multiplexer, but in this case only to the extent of a 10 ms delay.

To obtain the upper curve of Fig. 1, labeled "Effective BW," we implemented the admission control algorithm of Eqs. 2 and 4. As described earlier, the effective bandwidth approximation assumes that the loss (or tail) probability decays exponentially with increasing delay or buffer size; hence, the relationship is linear on the figure's semi-log scale. The simulations indicate that the actual loss probabilities for a given buffer size are considerably lower than that predicted by the effective bandwidth scheme; moreover, the measured relationship between loss probability and buffer size is not exponential.

The "Chernoff DE" approach refines the effective bandwidth result by adding a preterm to account for the loss probability in a bufferless multiplexer. We compare the result in this figure to the exact tail probability, since it is usually closer to the tail probability than the corresponding loss curve (e.g., compare this curve to the exact loss curve in Fig. 2). Here, the tail probability is approximated by $Ke^{-\delta B}$ with K given by Eq. 7, and δ calculated the same as for effective bandwidth. In these experiments, the preterm is $K = 0.09$, which correspondingly improves the estimate of the tail probability by that factor. However, as shown in the figure, the estimate of $P(Q > B)$ is still conservative by approximately four orders of magnitude for buffer sizes above 10 ms. Regardless, the asymptotic slopes of the effective bandwidth and Chernoff dominant eigenvalue (DE) curves do match the slope of the "Simulation" curve in the region of 10–50 ms, indicating that [8] does provide a good estimate of δ. Unfortunately, this does not necessarily correspond to a good estimate of $P(Q > B)$.

The curve labeled "MVA upper bound" provides an asymptotic upper bound to the tail probability as described in [16], without assuming a specific shape of the loss curve. From the figure, we note that this admission control curve follows the measured tail probability quite accurately, including emulating its two-segment shape. Moreover, although this approach is an asymptotic upper bound, empirical results suggest that it behaves as a global upper bound, as is the case in Fig. 1. Here, we again note that when the large-deviations-based expression for the tail probability in [38] is applied to the Gaussian arrival case, the resulting tail probability yields the same curve as the MVA upper bound.

The curve labeled "Lower bound" has been theoretically investigated in various papers in the context of both large deviations techniques and Extreme Value Theory [16, 38, 43]. An extensive study in [16] showed that the lower bound provides an accurate estimate of the tail probability and, like the MVA upper bound, closely matches the shape of the tail probability curve. Consequently, the MVA upper and lower bounds envelope the measured tail probability to approximately within an order of magnitude. Finally, we note that the "Rate variance" approach for the loss probability (shown in Fig. 2), while devised using an entirely different technique with stochastic envelopes, is in fact the same curve as the "Lower bound" shown here [18].



■ **Figure 2.** *Loss probability comparisons for video sources.*

Loss Probability for Multiplexed Video Flows — Figure 2 depicts the loss probability versus delay for the case in which a packet is dropped at the multiplexer if it violates its delay requirement d; that is, the actual buffer size of the multiplexer is $B = Cd$.

The curve labeled "Simulation" reports the actual fraction of packets dropped due to buffer overflow. Notice that this curve is quite similar to the simulated tail curve, except that it is one to two orders of magnitude lower than the corresponding curve shown in Fig. 1. The curve labeled "Average/peak" refers to the results of our implementation of [7]. For small buffers, the average/peak admission control test overestimates P_l by only one order of magnitude. However, since this test assumes a bufferless multiplexer, it is increasingly inaccurate for larger buffer sizes.

The "Hybrid" curve refers to the admission control test in [15]. Although the hybrid scheme captures the effect of statistical multiplexing and results in an improvement over the effective bandwidth curve, it too is significantly inaccurate in capturing the exact loss probability. The reason is that the hybrid scheme is based on the "burst region" being of a single exponential type. However, for sources that are correlated at multiple time-scales (such as the MPEG-video example shown here), the loss probability curve does not converge to its asymptotic decay rate quickly (even if there exists an asymptotic decay rate), and hence approximations such as the hybrid scheme perform quite poorly.

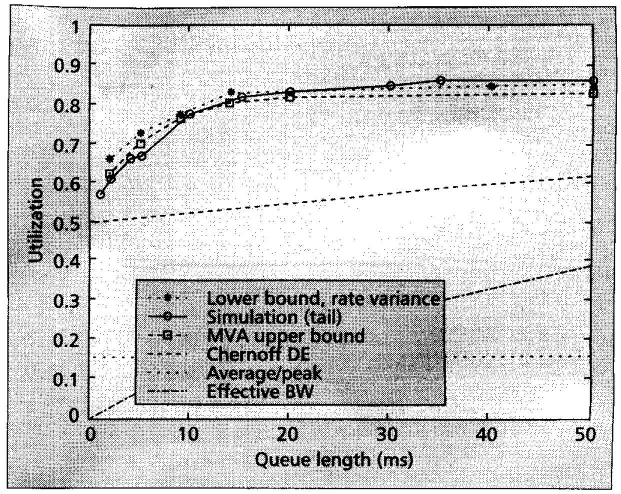
The curve labeled "Rate variance" in Fig. 2 depicts the results of the admission control test in [18]. This admission control test is able to capture the nonexponential relationship between loss and delay, and although theoretically the curve is a lower bound to the tail probability, it empirically behaves like an upper bound to the loss probability.

Finally, the curve labeled "MV approx" is the result of mapping the MVA upper bound for the tail probability to the loss probability in a finite buffer system [17]. This curve follows the simulated tail probability quite closely, capturing the nonexponential nature of the loss curve.

The Admissible Region for Multiplexed Video Flows — An admission control test's effectiveness is ultimately determined by its ability to correctly decide whether or not a new traffic flow can be admitted while still satisfying the QoS constraints of all established flows plus the new one. Figure 3 evaluates a number of admission control tests in such a manner.

For a tail probability of 10^{-6}, the figure shows the maximum number of admissible flows expressed as average utilization versus delay or buffer size. A point on one of the curves indicates the maximum value of N for the corresponding delay and tail probability. The simulation curve depicts the *measured* admissible region, whereas the other curves depict the admissible regions estimated by the corresponding admission control tests. A desirable property of an admission control algorithm is that its admissible region be as close as possible to, but not greater than, the simulation curve. In other words, the goal is to utilize resources as highly as possible without admitting more flows than can actually be supported, which would result in violations of the promised QoS.

From 1 to 50 ms delays, the trace-driven simulation curve of Fig. 3 shows the actual achievable average utilization of the multiplexer is in the range of 58–85 percent. Such high utilizations indicate that these MPEG-compressed video flows are well suited to statistical multiplexing despite their burstiness over multiple timescales. Note also that, as in Fig. 3, buffering has a considerable advantage up to approximately 10–20 ms;



■ Figure 3. *Admissible region comparisons for video sources.*

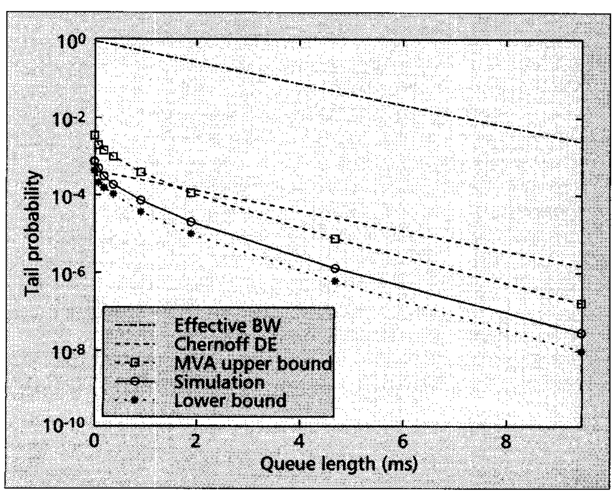
but beyond that, additional buffering does not significantly increase the admissible region.

Of these admission control algorithms, the MVA upper bound (for the tail probability) most closely approximates the measured admissible region, with the MVA loss, lower bound, and rate variance tests also approximately following the true admissible region and capturing trends such as the relative benefits of adding buffer space. The remaining approaches are significantly more conservative. We investigate the reasons for these experimental observation later.

Empirical Evaluation for Voice Sources

In this section we evaluate the admission control tests for input traffic that is modeled via Markov Modulated on-off processes. We use this model for two reasons. First, it provides a baseline for comparison of the admission control tests under a simpler and more widely studied scenario. Second, this model is an accurate and accepted model for voice traffic because it captures the behavior of encoded voice by alternating between "active" (on) and "inactive" (off) states.

Scenario — For our experimental setup, we again consider a 45 Mb/s link serving multiplexed voice sources. With encoded voice's alternation between active and inactive states, Markov modulated on-off processes have frequently been used to model voice traffic [46, 47]. A Markov modulated on-off source is one in which traffic is transmitted only in the on state, and the source spends an exponentially distributed duration of time in the on and off states. In theory, one could solve a series of equations (balance equations) to calculate the exact loss or tail probability for a fluid queue serving on-off sources. However, in this case, since well over 1000 voice sources can be multiplexed on a 45 Mb/s link, such an exact analytical solution becomes computationally infeasible. Hence, we again use simulations to evaluate the various admission control algorithms. For these simulations, we assume a 1 ms slot size and use a discrete-time on-off Markov modulated fluid process as a voice source model. For each source, when in the on state traffic is generated with a fluid rate of 32 kb/s. Furthermore, for each source, let p_{ij} correspond to the transition probability from state i to state j (for i = 0,1 and j = 0,1), where state 0 corresponds to the off state and state 1 to the on state. Then p_{00} = 0.9983, p_{01} = 0.00167, p_{10} = 0.0025, p_{11} = 0.9975. To obtain reliable results at very low loss (or tail) probabilities, we use the *importance sampling* simulation technique described in [28].



■ Figure 4. *Tail probability comparisons for on-off (voice) sources.*

Tail Probability for Voice Sources — The experiments reported in Fig. 4 and 5 (for the tail and loss probabilities, respectively) are based on the above simulation scenario with a utilization of 93 percent, which corresponds to 2900 voice flows on the 45 Mb/s link.

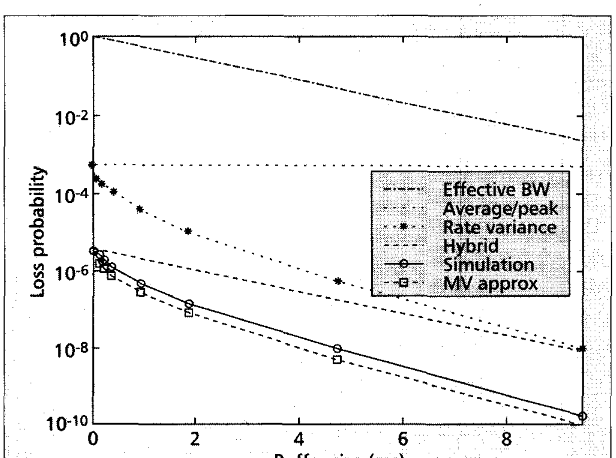
Figure 4 depicts the tail probability versus delay that a packet experiences in the queue. As in the case for MPEG video, in this scenario, even if a packet's deadline is violated at a queue, the packet is not dropped, but held in case its delay could be made up downstream.

Once again, note that the additive effective bandwidth approach results in a curve that is quite conservative, and an almost four orders of magnitude overestimate of the tail probability. However, in this case, the Chernoff DE approach of [14] performs better than in the case of multiplexed video. The reason for this is that the prefactor term used in [14] helps capture the statistical multiplexing gain due to the aggregation of the sources. Since multiplexed on-off sources do not exhibit strong multi-timescale correlation, the single exponential approximation does manage to provide a reasonable match with simulations (even though for larger buffer size the mismatch is almost two orders of magnitude). The MVA upper bound and the lower bound both accurately track the tail probability and encapsulate it within a relatively narrow envelope.

Loss Probability for Voice Sources — Figure 5 depicts the loss probability versus delay for the case in which a packet is dropped at the multiplexer if it violates its delay requirement.

The simulation curve again reports the actual fraction of packets that exceed the buffer level (or, equivalently, the delay requirement). For small buffers, the average/peak admission control test is more accurate than the effective bandwidth test, but again becomes increasingly inaccurate for larger buffer sizes due to the assumption of a bufferless multiplexer.

The "Hybrid" scheme [15] accurately captures the effect of statistical multiplexing, and as can be seen in Fig. 5, accurately captures the loss probability for small buffer values. However, similar to the case of the Chernoff DE approach, for large values of delay, the scheme can overestimate the loss probability by up to two orders of magnitude. The "Rate variance" approach performs worse than the hybrid scheme until larger buffer sizes, and the "MV approx" curve [17] most closely follows the measured loss curve.



■ Figure 5. *Loss probability comparisons for on-off (voice) sources.*

The Admissible Region — In Fig. 6 we compare the simulated admissible region with those of the admission control algorithms. Following the approach taken earlier, for a tail probability of 10^{-6}, the figure shows the maximum number of admissible traffic flows, expressed as average utilization, versus delay or buffer size. The simulation curve of Fig. 6 shows that the actual achievable average utilization of the multiplexer is in the range of 91–94 percent, or 3198–3303 multiplexed flows. These high utilizations are again due to the fact that voice traffic is well suited to taking advantage of a statistical multiplexed system.

Of these admission control algorithms, the rate variance approach (or MVA lower bound) most closely approximates the measured admissible region, but, being a lower bound, slightly exceeds the admissible curve. The MVA upper bound follows the admissible region curve quite well, while being somewhat conservative (a useful feature for admission control). In this case, the Chernoff DE approach and average/peak test also work reasonably well (the latter being increasingly worse for larger allowable delays), while the additive effective bandwidth approach is again quite conservative.

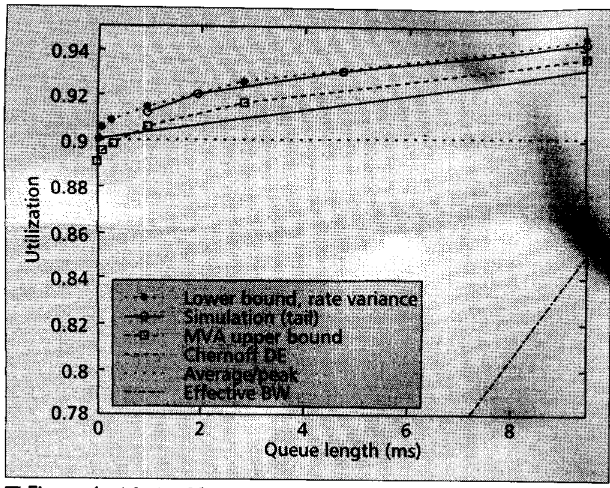
Admission Control Tests: Approximations and Accuracy

The Rationale of Gaussian Modeling

The experiments of the previous section show that modeling the aggregate arrival process as a general Gaussian process results in good estimates of the tail probability even when the number of sources multiplexed is moderate (as in the case of the video sources). As discussed previously, the Gaussian characterization is motivated by the Central Limit Theorem, since a link in a high-speed network is expected to carry many traffic sources. One advantage of a Gaussian characterization versus other types of models is that it requires less detailed knowledge of the source than, for example, Markov arrival models. However, one could then ask the question, why not appeal to a Poisson limit theorem and characterize the aggregate traffic to a multiplexer as a Poisson process? (After all, the Poisson process can be characterized simply by its mean.) The reason is quite simple: the Poisson limit theorem involves a very different type of scaling, and the Poisson traffic characterization destroys the temporal correlation structure of the traffic, in fact making it memoryless (independent and stationary increments!). Since real traffic is highly correlated, a Poisson characterization results in grossly incorrect loss (or tail) probability calculation, and hence a wrong admissible region. The Gaussian characterization, on the other hand, is very useful since it allows for different correlation structures (any function can be a valid autocovariance function), and hence captures the temporal correlation of the traffic. As an illustrative example, using the techniques of [48], one can compute the loss probability for fractional Brownian motion, a self-similar process, as a trivial special case since it merely represents a particular autocovariance function.

The Impact of Buffering

The experiments of the previous section evaluate the impact of buffer-size scaling on both the multiplexer's performance as well as the effectiveness of the different admission control algorithms. The trace-driven simulation experiments indicate that even for highly correlated traffic, some buffering is of substantial benefit. For example, our simulations show that with 66 multiplexed MPEG flows, 30 ms or approximately 170 kbytes worth of buffering decreases the loss probability from .003 to $7 \cdot 10^{-7}$. Furthermore, considering the admissible



■ Figure 6. *Admissable region comparisons for on-off (voice) sources.*

region and a loss probability of 10^{-6}, 30 ms of buffering increases the admissible region from 49 to 66 flows, a 35 percent improvement. As noted earlier, the incremental advantages of an increased buffer size do not extend indefinitely, but rather decay quickly once the multiplexer has 10–20 ms of buffering. In the case of on-off sources, buffering is of less benefit in that it does not result in as significant an increase in the admissible region. However, in real systems one expects a mix of both multi-timescale correlated traffic such as video and short-term correlated traffic such as voice, so the resulting aggregate traffic will indeed exhibit multi-timescale properties. Hence, the conclusions about buffers drawn from the experiments with video traces are more relevant in realistic network design.

Thus, for the video sources with the importance of buffering in the actual system, admission control tests that take into account buffer size scaling are able to significantly outperform those that do not. Indeed, the average/peak admission control test of [7] is one of the more conservative tests partially because of its assumption of a bufferless multiplexer.

We note, however, that the aforementioned 35 percent increase in the admissible region of Fig. 3, or the four-orders-of-magnitude decrease in the loss probability, comes at a cost: the costs of the memory itself; buffer management costs; and potential increased complexity in the admission control tests, since tests incorporating network buffers must also consider the traffic's autocorrelation structure.

Economies of Scale in the Number of Flows

As the number of multiplexed sources N increases, the amount of resources (bandwidth and buffer space) that must be reserved *per source* should decrease as an effect of statistical multiplexing and a simple consequence of the law of large numbers. In other words, we expect to have economies of scale in the number of multiplexed sources. While most admission control tests we considered exploit such economies of scale, additive effective bandwidth tests do not, since such tests determine a flow's resource demands using only the stochastic properties of the source itself (e.g., its index of dispersion), independent of the properties of other sources or the total number of sources being multiplexed. Indeed, independently summarizing the resource requirements of heterogeneous and bursty traffic sources by per-flow bandwidths was exactly the goal of the original work on additive effective bandwidths. However, as is evident from Fig. 3, the approach's lack of N scaling significantly limits the achievable utilization.

The Loss Curve

The relationship between packet loss probability and buffer size, as in Fig. 3, is often referred to as the *loss curve*. While the ultimate goal of an admission control test is to correctly determine the admissible region, many admission control tests have been designed with an intermediary focus on the shape of the loss curve. For example, [14] is motivated by the dominant eigenvalue of Markovian sources to approximate the loss curve with an exponential relationship $P_l \approx Ke^{-\delta B}$. Effective bandwidth schemes also assume an exponential relationship, but with $K \approx 1$.

Our simulation results of Fig. 3 depict a loss curve that is significantly different from exponential. While long-range dependence is a plausible explanation for this [38], it could also be that the multiple-timescale correlation of the sources results in this very slow convergence of the loss (or tail) probability to its asymptotic slope.

We note that the maximum-variance-based admission control tests discussed earlier are able to track this nonexponential loss curve quite well. These tests do not assume that traffic flows are long-range-dependent, but rather use a general second moment traffic description to obtain the required loss (or tail) curves.

In [14], trace drive simulations were also performed and the reported loss curves are nearly exponential. Consequently, the Chernoff DE test of [14] was quite accurate for those experiments. While we found corroborating results for *voice* sources, for video sources we found that the measured tail or loss curve is far from exponential, so the Chernoff DE test is considerably less accurate. This may have been due to a combination of the following facts: reference [14] used a videoconference trace, which likely does not exhibit multi-timescale rate variation, and used JPEG- rather than MPEG-compressed video, with the latter having substantially more rate variation on small timescales as well.

Here, we note that for sources that are correlated over small time scales (e.g., the on-off voice sources), any admission control test that accurately estimates the statistical multiplexing gain via the prefactor K in front of the exponential in Eq. 3 will capture the loss (or tail) probability quite well. However, techniques such as the additive effective bandwidth that are unable to capture this statistical multiplexing gain, or techniques such as the average/peak combinatorics that do not account for nonzero buffers, will still tend to be rather conservative.

Important Traffic Parameters

In addition to assumptions about the shape of the loss curve and network buffer sizes, an admission control test must characterize traffic flows according to a parameterized traffic model.

The on-off traffic model used in the admission control test of [7] is the simplest of the models we considered here (indeed, peak and average rate are likely the minimum amount of information needed to provide a statistical service). While this model is simple and closely related to standard traffic models (which specify peak rate, average rate, and burst length), it is also quite closely tied to the assumption of a bufferless multiplexer, which, as described above, has a considerable utilization penalty for an admission control test. Indeed, to take into account the effects of buffering, more information is needed about the traffic flows such as their autocorrelation structure, their maximum rates over various interval lengths [49], or at least a burst length parameter.

While many admission control tests can work well with on-off sources, when these same admission control tests are applied to the compressed video sources we have considered here, the tests exhibit considerable inaccuracies. For traffic flows that exhibit multiple-timescale rate variation, we argue that refined traffic models (beyond peak and average rate) are needed to extract the full statistical multiplexing gain. For example, the proposed traffic model of [49] characterizes a source by a family of rate-interval pairs where the rate is a bounding rate over the corresponding interval length. Such parameters can also be used to bound or approximate stochastic parameters such as the rate variance envelope considered here [24, 50]. A second possibility is to have users directly convey their second moment characteristics to the network, which our experimental results indicate can accurately estimate the admissible region. Unfortunately, the downside to incorporating more sophisticated traffic models is twofold. First, additional traffic parameters are needed beyond the standard three parameter models, which also means that policing requires at least multilevel leaky buckets rather than a single leaky bucket [49]. Second, the more detailed the traffic parameter, the greater the burden on network clients to accurately characterize their traffic in advance.

It therefore appears that if a refined traffic model is not used, statistical services would yield such low resource utilization for bursty traffic flows that either renegotiated services [5, 51] or measurement-based services [19–22] must be used instead. While such services have their own merits, as discussed in the respective works, unfortunately they cannot provide statistical QoS *guarantees* per se. Moreover, a renegotiated service requires increased signaling overhead, and a measurement-based service must successfully make accurate predictions of future resource requirements using past measurements of aggregate multiple-timescale sources, which, as we have seen here, is a difficult problem even when future arrival statistics are known.

Lastly, we note that admission control tests differ in their computational complexity or the number of instructions that must be executed upon the arrival of a new admission request. While exploration of this issue is beyond the scope of this current work, we do note that all of the schemes we have considered were designed with implementation considerations, with [7] giving this issue the most attention.

Conclusions

From the results of our trace-driven simulations and admission control experiments with a diverse set of admission control algorithms, we make the following observations:

- Assuming a bufferless multiplexer introduces a substantial utilization penalty if the actual multiplexer does contain buffer space.
- Economies of scale in the number of multiplexed flows is a crucial component in achieving a high degree of accuracy.
- Experimentally observed loss curves (loss probability vs. buffer size) for compressed video sources are quite different than the commonly assumed exponential relationship. If, indeed, the curves do asymptotically (with buffer size) become exponential, the convergence rate is usually quite slow, which renders exponential types of approximations fairly inaccurate over any meaningful buffer size or loss probability.
- Refinements of current standard traffic models are required in order to obtain a reasonable statistical multiplexing gain and a statistical QoS guarantee.
- Admission control tests that work demonstrably well with exponential on-off sources can suffer from considerable inaccuracies when applied to multiple-timescale sources such as compressed variable bit rate video.

We found that a number of admission control tests from the literature perform quite well experimentally for both on-off sources and compressed video sources, including algorithms by the authors [16, 18] and others [38, 42, 43].

Acknowledgments

E. Knightly gratefully acknowledges the support of NSF CAREER Award ANI-9733610, NSF Grant ANI-9730104, Nokia Corporation, the Texas Advanced Technology Program, and Texas Instruments.

N. Shroff gratefully acknowledges the support of NSF CAREER Award NCR-9624525, and NSF Grants ANI-9805441, CDA-9422250, and CDA 96-17388.

The authors would like to give many thanks to Jinwoo Choe and Han S. Kim for running the simulations reported in this article, and for preparing the figures for this article. The authors would also like to thank the anonymous reviewers for carefully reading the article, and for their helpful comments.

References

[1] E. Knightly, "On the Accuracy of Admission Control Tests," *Proc. IEEE ICNP '97*, Atlanta, GA, Oct. 1997, pp. 125–33.
[2] D. Wrege *et al.*, "Deterministic Delay Bounds for VBR Video in Packet-Switching Networks: Fundamental Limits and Practical Tradeoffs," *IEEE/ACM Trans. Net.*, vol. 4, June 1996, pp. 352–62.
[3] M. Garret and W. Willinger, "Analysis, Modeling and Generation of Self-Similar VBR Video Traffic," *Proc. ACM SIGCOMM '94*, London, U.K., Aug. 1994, pp. 269–80.
[4] A. Lazar, G. Pacifici, and D. Pendarakis, "Modeling Video Sources for Real Time Scheduling," *ACM Multimedia Sys. J.*, vol. 1, Apr. 1994, pp. 253–66.
[5] H. Zhang and E. Knightly, "RED-VBR: A Renegotiation-Based Approach to Support Delay-Sensitive VBR Video," *ACM Multimedia Sys. J.*, vol. 5, May 1997, pp. 164–76.
[6] D. Ferrari and D. Verma, "A Scheme for Real-Time Channel Establishment in Wide-Area Networks," *IEEE JSAC*, vol. 8, Apr. 1990, pp. 368–79.
[7] T. Lee, K. Lai, and S. Duann, "Design of a real-time call admission controller for ATM networks," *IEEE/ACM Trans. Net.*, vol. 4, Oct. 1996, pp. 758–65.
[8] C. Courcoubetis and R. Weber, "Effective Bandwidths for Stationary Sources," *Probability in Eng. and Info. Sci.*, vol. 9, no. 2, 1995, pp. 285–94.
[9] A. Elwalid and D. Mitra, "Effective bandwidth of general Markovian traffic sources and admission control of high speed networks," *IEEE/ACM Trans. Net.*, vol. 1, June 1993, pp. 329–43.
[10] R. Guérin, H. Ahmadi, and M. Naghshineh, "Equivalent Capacity and Its Application to Bandwidth Allocation in High-Speed Networks," *IEEE JSAC*, vol. 9, Sept. 1991, pp. 968–81.
[11] G. Kesidis, J. Walrand, and C. Chang, "Effective Bandwidths for Multiclass Markov Fluids and Other ATM Sources," *IEEE/ACM Trans. Net*, vol. 1, Aug. 1993, pp. 424–28.
[12] A. Baiocchi *et al.*, "Loss Performance Analysis of an ATM Multiplexer Loaded with High-Speed On-Off Sources," *IEEE JSAC*, vol. 9, Apr. 1991, pp. 388–93.
[13] G. Choudhury, D. Lucantoni, and W. Whitt, "Squeezing the most out of ATM," *IEEE Trans. Commun.*, vol. 44, Feb. 1996, pp. 203–17.
[14] A. Elwalid *et al.*, "Fundamental bounds and approximations for ATM multiplexers with applications to video teleconferencing," *IEEE JSAC*, vol. 13, Aug. 1995, pp. 1004–16.
[15] N. Shroff and M. Schwartz, "Improved Loss Calculations at an ATM Multiplexer," *IEEE/ACM Trans. Net.*, vol. 6, Aug. 1998, pp. 411–22.
[16] J. Choe and N. Shroff, "A Central Limit Theorem Based Approach to Analyze Queue Behavior in ATM Networks," *IEEE/ACM Trans. Net.*, vol. 6, Oct. 1998, pp. 659–71.
[17] H. Kim and N. Shroff, "Loss Probability Calculations in a Finite Buffer Queueing System," Tech. rep., Purdue Univ.
[18] E. Knightly, "Second Moment Resource Allocation in Multi-Service Networks," *Proc. ACM SIGMETRICS '97*, Seattle, WA, June 1997, pp. 181–91.
[19] R. Gibbens, F. Kelly, and P. Key, "A decision-theoretic approach to call admission control in ATM networks," *IEEE JSAC*, vol. 13, Aug. 1995, pp. 1101–14.
[20] S. Jamin *et al.*, "A measurement-based admission control algorithm for integrated services packet networks," *IEEE/ACM Trans. Net.*, vol. 5, Feb. 1997, pp. 56–70.
[21] J. Qiu and E. Knightly, "QoS Control via Robust Envelope-Based MBAC," *IEEE/IFIP IWQoS '98*, Napa, CA, May 1998.
[22] D. Tse and M. Grossglauser, "Measurement-based Call Admission Control: Analysis and Simulation," *Proc. IEEE INFOCOM '97*, Kobe, Japan, Apr. 1997.
[23] A. Elwalid, D. Mitra, and R. Wentworth, "A new approach for allocating buffers and bandwidth to heterogeneous, regulated traffic in an ATM node," *IEEE JSAC*, vol. 13, Aug. 1995, pp. 1115–27.
[24] E. Knightly, "H-BIND: A New Approach to Providing Statistical Performance Guarantees to VBR Traffic," *Proc. IEEE INFOCOM '96*, San Francisco, CA, Mar. 1996, pp. 1091–99.

[25] S. Rajagopal, M. Reisslein, and K. Ross, "Packet Multiplexers with Adversarial Regulated Traffic," *Proc. IEEE INFOCOM '98*, San Francisco, CA, Mar. 1998.
[26] M. Krunz and S. Tripathi, "Impact of video scheduling on bandwidth allocation for multiplexed MPEG streams," *Multimedia Sys. J.*, vol. 5, no. 6, 1997, pp. 347–57.
[27] J. McManus and K. Ross, "Video-on-demand over ATM: constant-rate transmission and transport," *IEEE JSAC*, vol. 14, Aug. 1996, pp. 1087–98.
[28] C. Chang, "Stability, Queue Length, and Delay of Deterministic and Stochastic Queueing Networks," *IEEE Trans. Auto. Control*, vol. 39, May 1994, pp. 913–31.
[29] D. Anick, D. Mitra, and M. Sondhi, "Stochastic Theory of a Data-Handling System with Multiple Sources," *Bell Sys. Tech. J.*, vol. 61, 1982, pp. 1871–94.
[30] N. Shroff and M. Schwartz, "Video Modeling within Networks using Deterministic Smoothing at the Source," *Proc. IEEE INFOCOM '94*, Toronto, Ontario, Canada, June 1994, pp. 342349.
[31] N. Shroff, "Traffic Modeling and Analysis in High Speed ATM Networks," PhD thesis, Graduate School of Arts and Sci., Columbia Univ., 1995.
[32] P. Skelly, M. Schwartz, and S. Dixit, "A Histogram-Based Model for Video Traffic Behavior in an ATM Multiplexer," *IEEE/ACM Trans. Net.*, vol. 1, Aug. 1993, pp. 446–59.
[33] N. Shroff and M. Schwartz, "Modeling VBR Video over Networks End-to-End using Deterministic Smoothing at the Source," *Int'l. J. Commun. Sys.*, Special Issue on Traffic Management in ATM Networks, vol. 7, Dec. 1994, pp. 337–48.
[34] R. Loynes, "The Stability of a Queue with Non-independent Inter-arrival and Service Times," *Proc. Cambridge Phil. Soc.*, vol. 58, 1962, pp. 497–520.
[35] C. Courcoubetis, V. Siris, and G. Stamoulis, "Application and Evaluation of Large Deviation Techniques for Traffic Engineering in Broadband Networks," *Proc. ACM SIGMETRICS '98*, Madison, WI, June 1998.
[36] N. G. Duffield and N. O'Connell, "Large deviations and overflow probabilities for the general single server queue, with application," *Proc. Cambridge Phil Soc.*, vol. 118, 1995, pp. 363–74.
[37] F. Kelly, "Notes on Effective Bandwidths," *Stochastic Networks: Theory and Applications*, Oxford Univ. Press, 1996.
[38] D. Botvich and N. Duffield, "Large deviations, the shape of the loss curve, and economies of scale in large multiplexers," *Queueing Systems*, vol. 20, 1995, pp. 293–320.
[39] P. Glynn and W. Whitt, "Logarithmic asymptotics for steady-state tail probabilities in a single-server queue," *J. Applied Probability*, 1994, pp. 131–55.
[40] J. Beran *et al.*, "Long-range dependence in variable-bit-rate video traffic," *IEEE Trans. Commun.*, vol. 43, Feb.–Apr. 1995, pp. 1566–79.
[41] W. Leland *et al.*, "On the Self-Similar Nature of Ethernet Traffic (Extended Version)," *IEEE/ACM Trans. Net.*, vol. 2, Feb. 1994, pp. 1–15.
[42] N. Likhanov and R. R. Mazumdar, "Cell-loss asymptotics in buffers fed with a large number of independent stationary sources," *Proc. IEEE INFOCOM*, 1998.
[43] M. Montgomery and G. de Veciana, "On the Relevance of Time Scales in Performance Oriented Traffic Characterization," *Proc. IEEE INFOCOM*, 1996, pp. 513–20.
[44] O. Rose, "Statistical Properties of MPEG Video Traffic and their Impact on Traffic Modeling in ATM Systems," *Proc. IEEE Conf. Local Comp. Nets.*, Minneapolis, MN, Oct. 1995, pp. 397–406.
[45] P. Bratley, B. Fox, and L. Schrage, *A Guide to Simulation*, New York: Springer-Verlag, 2nd ed., 1987.
[46] J. Daigle and J. Langford, "Models for Analysis of Packet Voice Communication Systems," *IEEE JSAC*, vol. 4, Sept. 1986, pp. 847–55.
[47] K. Sriram and W. Whitt, "Characterizing Superposition Arrival Processes in Packet Multiplexer for Voice and Data," *IEEE JSAC*, vol. 4, Sept. 1986, pp. 833–46.
[48] J. Choe and N. Shroff, "Queueing Analysis of High-Speed Multiplexers including Long-Range Dependent Arrival Processes," *Proc. IEEE INFOCOM '99*, New York, NY, Mar. 1999.
[49] E. Knightly and H. Zhang, "D-BIND: An Accurate Traffic Model for Providing QoS Guarantees to VBR Traffic," *IEEE/ACM Trans. Net.*, vol. 5, Apr. 1997, pp. 219–31.
[50] E. Knightly, "Enforceable Quality of Service Guarantees for Bursty Traffic Streams," *Proc. IEEE INFOCOM '98*, San Francisco, CA, Mar. 1998.
[51] M. Grossglauser, S. Keshav, and D. Tse, "RCBR: A simple and efficient service for multiple time-scale traffic," *IEEE/ACM Trans. Net.*, vol. 5, Dec. 1997, pp. 741–55.

Biographies

EDWARD KNIGHTLY (knightly@ece.rice.edu) received a B.S. degree from Auburn University in 1991, an M.S. degree from the University of California at Berkeley 1992, and a Ph.D. degree from the University of California at Berkeley in 1996, all in electrical engineering. Since 1996 he has been an assistant professor in the Department of Electrical and Computer Engineering at Rice University. He is currently an editor of *IEEE Network* and served as co-chair for the 1998 IEEE/IFIP International Workshop on Quality of Service. He received the National Science Foundation CAREER Award in 1997. His research interests are in the areas of resource management and quality of service networks.

NESS B. SHROFF (shroff@ecn.purdue.edu) received a B.S. degree from the University of Southern California in 1988, an M.S.E degree from the University of Pennsylvania in 1990, and M. Phil and Ph.D. degrees from Columbia University in 1993 and 1994, respectively, all in electrical engineering. Since 1994, he has been an assistant professor in the School of Electrical and Computer Engineering at Purdue University. He is currently the program chair for the 1999 IEEE Computer Communications Workshop. His research interests are in traffic analysis, network control, resource management, and quality of service in high-speed wired and wireless networks. He has authored or co-authored over 33 conference and journal publications, and one pending patent. He received the National Science Foundation CAREER award in 1996.