On the Accuracy of Admission Control Tests

Edward W. Knightly
ECE Department
Rice University
knightly@ece.rice.edu

Abstract

In guaranteed-performance networks, a Connection Admission Control (CAC) algorithm determines whether or not a new connection can be admitted to the network such that all connections will obtain their required Quality of Service (QoS) guarantees. However, inaccuracies *in a CAC algorithm lead to either under-utilization of resources or violations of the promised QoS guarantees. In this paper, we evaluate the accuracy of a broad range of CAC algorithms that have been proposed in the literature. Our approach is to perform experiments with long traces of MPEG-compressed video, considering both heterogeneous traffic streams and priority schedulers. We compare the admissible regions and QoS parameters predicted by our implementations of the CAC algorithms with those obtained from trace-driven simulations. We then identify the key aspects of a CAC test for achieving a high degree of accuracy and hence a high statistical multiplexing gain.*

1 Introduction

Bursty traffic sources that require Quality of Service (QoS) guarantees in terms of loss and delay are emerging as one of the most important types of traffic of future integrated services networks. Because of the variable bit rate nature and the multiple-time-scale characteristics of many realistic sources, e.g., [14, 22, 31], it is difficult to determine the amount of resources that need to be allocated to individual connections such that each connection obtains the performance it requires. This problem is exacerbated when different connections require different throughputs, delay bounds, and loss probabilities, since resources must then be allocated in networks that use *prioritized* service disciplines such as Static Priority (SP).

A statistical network service guarantees performance parameters such as the probability of packet loss and delay-bound violation. To provide such guarantees, Connection Admission Control (CAC) tests are used to limit the number of streams to the maximum number such that all connections obtain their required QoS. A CAC test strives to *a priori* determine this maximum number of admissible streams as accurately as possible. Indeed, if the CAC test incorrectly admits too many streams, then violations of the guaranteed QoS will occur; alternatively, if it denies access to streams that could have been successfully admitted, then under-utilization of network resources will occur.

In the literature, many CAC tests have been proposed for statistical services including (1) tests based on average and peak rate combinatorics [11, 23], (2) tests based on effective bandwidth theory [7, 9, 16, 17], (3) tests based on refinements of effective bandwidth theory [6, 8, 10, 26, 28], (4) tests based on stochastic envelope processes [3, 19, 21], and (5) algorithms designed explicitly for long-range-dependent traffic streams [2, 10, 31]. However, there has been little work towards evaluating the accuracy and effectiveness of these schemes under realistic workloads. Indeed, many CAC tests have been evaluated only with exponential on-off sources that have properties quite different from those observed by measurement studies of many actual traffic streams [14, 22, 31]. Moreover, these tests have not yet been comparatively evaluated to understand the components of a test necessary for accurate resource allocation.

In this paper, we experimentally evaluate the accuracy of a diverse set of proposed CAC tests using tests from each of the five types enumerated above. To evaluate these CAC tests, we utilize two 30-minute traces of bursty MPEG-compressed video, with one trace of an action movie and the other of a newscast. We first implement five CAC tests from the literature and determine their respective admissible regions at SP and FCFS schedulers for numerous combinations of video connections and QoS parameters. We then evaluate the *accuracy* of the CAC tests with trace-driven simulations which allow us to examine how well the CAC tests make their respective *a priori* determinations of the streams' performance parameters. Specifically, we simulate a 45 Mbps SP multiplexer servicing various mixes of traffic streams, with each stream's arrival sequence given by a video trace with a random start time. For each combination of traffic streams and multiplexer buffer size, we measure the streams' resulting performance parameters. We evaluate a CAC test's accuracy by examining how well it determines the actual admissible regions of the trace-driven simulations.

0-8186-8061-X/97 $10.00

Our study is the first to study these CAC tests using trace-driven simulations with heterogeneous traffic streams and priority service disciplines. From our experimental results, we identify the components of a CAC test essential to achieving a high degree of accuracy and show that (1) the assumption of a bufferless multiplexer has a significant utilization penalty, (2) a CAC test must exhibit economies of scale in the number of multiplexed connections, (3) observed shapes of the "loss curve" are quite different than the commonly assumed exponential relationship, (4) the traffic model, or parameters used to describe the properties of traffic streams to the network, requires more information than is currently standardized, and (5) a CAC test's accuracy with exponential on-off sources does not assure accuracy with bursty compressed-video sources.

Finally, we note that we consider a single multiplexer. Studies of end-to-end statistical QoS guarantees can be found in [4, 30, 32] and elsewhere.

2 CAC Tests for Statistical Service

In this section, we describe five classes of CAC tests that have been proposed for providing statistical QoS guarantees in multi-service networks. While these classes do not encompass all proposed schemes, they do provide broad coverage of the techniques applied to the CAC problem.

2.1 Average/Peak Rate Combinatorics

In [23], source j is characterized by its peak rate MAX_j and its average rate AVE_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 $p_{on,j} = \text{AVE}_j/\text{MAX}_j$, and its rate distribution is given by

$$f_j(x) = \begin{cases} p_{on,j} & x = 0 \\ 1 - p_{on,j} & x = MAX_j \\ 0 & 0 < x < MAX_j \end{cases} \quad (1)$$

Using this rate distribution, a CAC test 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, [23] focuses on efficient computation of the aggregate arrival rate distribution and subsequently the loss probability.

In [11], traffic stream j is also characterized by its peak rate MAX_j and average rate AVE_j. In contrast to [23] in which AVE_j represents the long-term average rate, in [11] 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 $\text{AVE}_j \cdot I_j$ packets during any interval of length I_j (changing [11]'s notation for consistency). For Earliest Deadline First schedulers, the probability of delay-bound violation is calculated by examining combinations of active connections (connections which are *on* with probability $\text{AVE}_j/\text{MAX}_j$) that may cause a delay-bound violation, and by summing their respective probabilities.

In this paper, we evaluate the test of [23], which we often refer to as the "Avg/Peak" test.

2.2 Effective Bandwidths

Various *effective bandwidth* CAC tests have been proposed in the literature including [7, 9, 16, 17]. In such CAC tests, a bandwidth is reserved for each stream according to its stochastic properties and the required loss probability p. Once the effective bandwidth of stream j is determined, which we denote by $E_j(p)$, the CAC test checks that

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

where C is the link capacity and N is the number of multiplexed connections. Effective-bandwidth type results have been derived using several inter-related techniques, including eigenvalue decomposition of Markovian streams [9], large deviations theory [7, 17], and theory of envelope processes (stochastic bounds of a process' moment generating function) [3].

For example, [1] shows how the buffer occupancy distribution for Markov Modulated Fluid Sources can be decomposed according to the eigenvalues of the aggregate Markovian arrival process. With the overall buffer distribution found to be a sum of exponential terms, a single eigenvalue dominates with asymptotically large buffers. Hence, the buffer overflow probability can be approximated as $p \approx e^{-\delta B}$ where B is the buffer size, δ is the dominant eigenvalue, and the preterm is approximated by 1. Under such conditions, a CAC test of the form of Equation (2) is derived in [9] based on each stream's transition matrix.

Note that the effective bandwidth of a stream is independent of the properties of all other traffic streams as well as the number of sources N and the link capacity C: it is determined only by the stochastic properties of the traffic stream itself and the required loss probability p. For example, in [7, 17], a stationary stream's effective bandwidth is given by

$$E_j(p) = \frac{1}{\delta} \lim_{t \to \infty} \frac{1}{t} log E e^{\delta A_j[0,t]} \quad (3)$$

where $A_j[0,t]$ denotes stream j's arrivals in $[0,t]$. Here, we evaluate the effective bandwidth CAC test of [7] which is based on large deviations theory.

2.3 Refinements of Effective Bandwidth Theory

Since effective bandwidth schemes can be overly conservative (for reasons discussed in Sections 3 and 4), a number of refinements have been proposed [6, 8, 19, 26, 28]. Here, we describe two recent approaches: the burst-region/cell-region scheme of [28] and the Chernoff Dominant Eigenvalue approach of [8]. Both of these works focus on the shape of the "loss curve" or the loss probability as a function of the buffer size B, which for effective bandwidth is approximated by $p \approx e^{-\delta B}$.

In [29], each video source's frame-size distribution (and hence rate-distribution) is modeled with an 8-bin histogram. The distribution of the aggregate rate of the multiplexed streams is calculated in [29] via convolutions of the individual histograms, and an approximate queueing analysis is performed by mapping the aggregate arrival rate distribution to that of an MMPP. Noting that [29] is only applicable in the small buffer regime, [28] refined the large-buffer asymptotics of the effective bandwidth approach with the small-buffer results of the histogram approach to determine the loss probability in both the cell region (small buffers) and burst region (large buffers). Roughly, [28] refines the effective bandwidth estimate of p by using a second exponential term.

In [8], Elwalid et al. observed that effective bandwidth's loss curve approximation of $p \approx e^{-\delta B}$ could be significantly improved with the addition of a preterm. Thus, [8] proposed the approximation

$$p \approx Le^{-\delta B} \tag{4}$$

where L is the loss probability in a bufferless multiplexer as estimated by Chernoff's theorem and δ is the same dominant eigenvalue for Markovian sources as in the effective bandwidth result of [9]. Since L can in practice be substantially less than 1, CAC tests utilizing this term can have improved accuracy as compared to effective bandwidth tests.

Note that these approaches should not be termed effective bandwidth schemes in that there is no notion of an additive bandwidth requirement per source as in Equation (2), though they do use the same asymptotic exponential loss rate δ. Here, we evaluate the approach of [8], with the minor modification that we calculate δ using large deviations theory rather than eigenvalue decomposition, since the former approach applies to more general classes of traffic streams than Markovian.

2.4 Stochastic Envelope Based Resource Allocation

A stochastic traffic envelope bounds some statistical properties of $A_j[s, s+t]$ as a function of the interval length t, e.g., in [21] the *distribution* of $A_j[s, s+t]$ is bounded, and in [3], the moment generating function. In [19], traffic is characterized via a *rate-variance envelope* defined by:

$$RV_j(t) = Var\left(\frac{A_j[s, s+t]}{t}\right) \tag{5}$$

which describes a stream's second moment correlation structure. Based on the streams' $RV_j(t)$ characterizations as well as their mean rate m_j, CAC tests are derived in [19] for SP and FCFS schedulers. In the tests, the stochastic envelope of the aggregate traffic is approximated with a Gaussian envelope with variance $\sum_j t^2 RV_j(t)$ 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 aggregate envelope $\sum_j t^2 RV_j(t)$.

In this paper, we evaluate the CAC test of [19]. Other CAC tests based on second moment characterizations of traffic streams can be found in [5, 25] and elsewhere.

2.5 Algorithms for Long-Range-Dependent Traffic

Since many realistic traffic streams have been shown to exhibit rate variations over multiple time scales [14, 22, 31], a number of resource allocation algorithms have been designed explicitly to deal with such traffic streams, e.g., [2, 10, 31]. Roughly, a long-range-dependent source is one with an autocorrelation function that decays hyperbolically rather than exponentially.

As an illustrative example, [10] uses the Bahadur-Rao theorem from large deviations theory to derive an expression for loss probability for fractional Brownian motion processes. While restriction to the homogeneous case precludes use of this technique as a general admission control algorithm, we include it here for illustrative purposes. With a Hurst parameter H, the loss probability is approximated in [10] by:

$$p \approx \frac{k_1}{\sqrt{2\pi N}} exp(-k_2 N^{2H-1} B^{2-2H}) \tag{6}$$

where k_1 and k_2 are given in [10] and are functions of the link capacity, mean rate, etc.. The important points about [10] for our purposes here are (1) the loss curve is *sub-exponential* in B and (2) the required bandwidth per source is sub-additive, unlike an effective bandwidth scheme.

3 Experimental Evaluation of CAC Tests

In this section, we evaluate the accuracy of the aforementioned CAC tests by performing a set of experiments using

trace-driven simulations with 30-minute traces of MPEG-compressed video. These traces were shown in [19] to exhibit second order statistical properties characteristic of long-range-dependent traffic streams. We consider various scenarios with different loads, QoS parameters, etc., and compare the QoS actually obtained by the streams in these trace-driven simulations with that predicted by the CAC tests.

3.1 Workload and Performance Metrics

The first trace consists of a newscast and the second of an action movie, with both traces taken from [27]. The traces are digitized to 384x288 pels and compressed at 24 frames per second using the MPEG compression algorithm with frame pattern *IBBPBBPBBPBB*. For the simulations, we segment video frames into ATM cells with the cells transmitted at equally-spaced intervals over the frame time, $\frac{1}{24}^{th}$ of a second.

We consider two scenarios in our experiments. The first scenario investigates trace-driven simulations with multiple randomly-offset copies of a single traffic trace and all traces obtaining the same QoS parameters. The second scenario investigates the heterogeneous case, with various combinations of randomly-offset copies of two different traces, and with the different streams obtaining different QoS parameters in a static priority scheduler.

Throughout the experiments, we focus on the following two performance metrics. The first is average utilization of the link, which is the total average rate of the streams 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 during the duration of the simulation (the link capacity multiplied by the simulation time).

Our second performance metric is the empirical fraction of packets that are dropped due to buffer overflow which we denote by p . We consider a range of buffer sizes B which have a corresponding statistically-guaranteed delay bound $d = B/C$.

In each experiment, the simulation runs until all sources have transmitted 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; the recorded simulated time of the multiplexer is therefore approximately 30 minutes. For each scenario (i.e., a certain number of connections, buffer size, etc.) thirty simulations are performed, each with independent start times; average results are reported.

3.2 Homogeneous Connections and QoS Parameters

In the first set of experiments, we consider the simulation scenario of Figure 1. For each simulation, N streams are multiplexed on a 45 Mbps link, with each stream's arrival pattern given by the movie trace, and its start time τ_j chosen uniformly over the length of the trace. We consider a single QoS pair for all connections represented by the pair (d, p) where d is the delay bound and p is the probability that a packet violates its delay bound or is dropped due to buffer overflows. The buffer size of the multiplexer for each simulation is set to be equal to Cd bits. Hence, the outputs of the experiments consist of three-tuples (N, d, p).

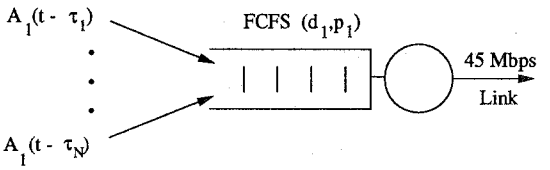


Figure 1: Scenario for Homogeneous Trace-Driven Simulation

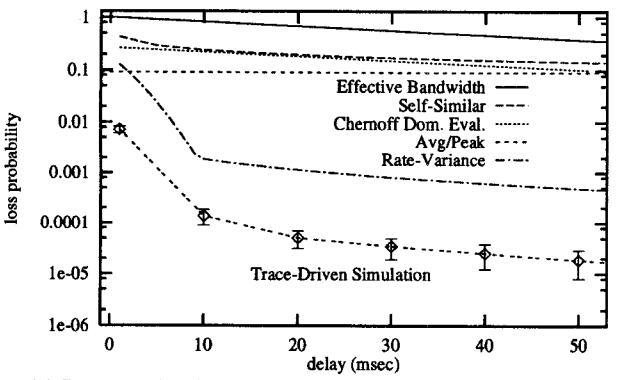
We compare the empirical performance results obtained by multiplexing randomly offset traces with the performance guarantees predicted by the CAC tests. For a given (d, p) QoS pair, a CAC test determines the maximum number of connections N that can be admitted so that all connections obtain their required QoS. Thus, the outputs of the CAC tests also consist of three-tuples (N, d, p). A test's accuracy is evaluated by comparing the actual (N, d, p) combinations obtained from the trace-driven simulations with those of the CAC tests.

3.2.1 Loss Curve

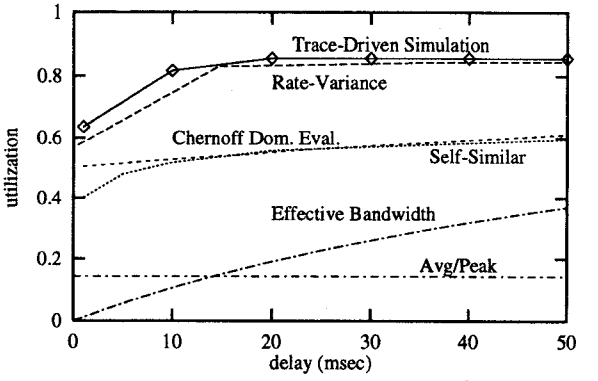
Figure 2(a) depicts the loss probability (or delay bound violation probability) versus delay, which is proportional to the multiplexer's total buffer space. The experiments reported in Figure 2(a) are for a fixed utilization of 88%, which corresponds to 69 MPEG-compressed movie connections on the 45 Mbps link.

The curve labeled "Trace-Driven Simulation" (the lowest curve) reports the actual fraction of packets dropped for the corresponding buffer size or delay bound. Notice that this curve drops sharply until buffer sizes of approximately 10 msec, after which it flattens considerably, indicating significant benefits for adding buffer space to a multiplexer, but in this case, only to the extent of a 10 msec delay.

To obtain the upper curve of Figure 2(a) labeled "Effective Bandwidth", we implemented the CAC algorithm in Theorem 2 of [7]. As described in Section 2, effective



(a) Loss Probability vs. Buffer Size at 88% Utilization



(b) Admissible Region for $p = 10^{-6}$

Figure 2: Trace-Driven Simulation and CAC Tests

bandwidth theory states that the loss 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 curve labeled "Chernoff Dominant Eigenvalue" refines the effective bandwidth result by adding a preterm to account for the loss probability in a bufferless multiplexer. The loss curve is approximated by $p \approx Le^{-\delta B}$ with L given by Equation (55) of [8]. In this case, the preterm is 0.14 which therefore correspondingly improves the estimate of p. However, as shown in the figure, the estimate of p is still conservative by approximately 5 orders of magnitude for buffer sizes above 10 msec. However, the asymptotic slopes of the effective bandwidth and Chernoff Dominant Eigenvalue curves do match the slope of the "Trace Driven Simulation" curve in the region of 10-50 msec, indicating that [7] does provide a good estimate of δ. However, this does not necessarily correspond to a good estimate of p.

To obtain the "Self-Similar" curve, we calculated a Hurst parameter of 0.86 using a least-squares fit for the slope of the trace's variance-time plot [24], and applied the test of [10]. Unlike the previous curves, this curve depicts a sub-exponential relationship, but is also quite conservative.

The curve labeled "Avg/Peak" refers to the results of our implementation of the CAC test in [23]. For small buffers, the Avg/Peak CAC test over-estimates p by only one order of magnitude. However, since the this test assumes a bufferless multiplexer, it is increasingly inaccurate for larger buffer sizes.

Finally, the curve labeled "Rate-Variance" in Figure 2(a) depicts the results of the CAC test in [19]. The test does not assume a specific shape of the loss curve (it is determined by the properties of the traffic streams) nor does it assume a bufferless or infinite buffer multiplexer. The rate-variance envelope based CAC test is able to characterize the non-exponential relationship between p and d, though the test is still somewhat conservative.

3.2.2 Admissible Region

A CAC test's effectiveness is ultimately determined by its ability to correctly decide whether or not a new connection can be admitted while still satisfying the QoS constraints of all established connections plus the new connection. Figure 2(b) evaluates five CAC tests in such a manner.

For a loss probability of $p = 10^{-6}$, the figure shows the maximum number of admissible connections, 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 (d, p) QoS combination. The curve labeled "Trace-Driven Simulation" depicts the *actual* admissible region, whereas the other five curves depict the admissible regions estimated by the corresponding CAC tests. A desirable property of a CAC algorithm is that its admissible region be as close as possible to, but not greater than, the "Trace-Driven Simulation" curve. That is, the goal is to utilize resources as highly as possible without admitting more connections than can actually be supported, which would result in violations of the promised QoS.

From 1 msec to 50 msec delays, the trace-driven simulation curve of Figure 2(b) shows the actual achievable average utilization of the multiplexer is in the range of 64% to 88% or 49 to 68 connections multiplexed connections. Such high utilizations indicate that these MPEG-compressed video streams are well suited to statistical multiplexing despite their burstiness over multiple time-scales. Note also that, as in Figure 2(a), buffering has a considerable advantage up to approximately 10 to 20 msec; but beyond that, additional buffering does not significantly in-

crease the admissible region.

Of the five CAC tests, the "Rate-Variance" test most closely approximates the measured admissible region, including the trends capturing the relative benefits of adding buffer space. Its admissible region ranges from 45 to 65 connections (58% to 84% average utilization), indicating that the algorithm admits 92% to 96% of the allowable connections (100% being the actual admissible region). In this delay range, the "Chernoff Dominant Eigenvalue" CAC test admits 39 to 50 connections (51% to 65% utilization) which corresponds to 61% to 74% of the actual admissible region. The resource allocation algorithm for homogeneous self-similar streams admits 31 to 46 connections, the "Effective Bandwidth" test 0 to 37, and the "Avg/Peak" CAC test admits 11 connections independent of the buffer size.

Thus, the five CAC tests differ dramatically in how well they determine the admissible region in various environments. We investigate the reasons for this experimental observation in Section 4.

3.3 Heterogeneous Connections and QoS Parameters

In the second set of experiments, we consider the simulation scenario shown in Figure 3. The figure shows a two-priority Static Priority scheduler with a link capacity of 45 Mbps. Each queue of the SP scheduler has an associated quality of service with a delay bound d_1 or d_2 and a probability of delay-bound violation or loss p_1 or p_2. At the high priority queue (level 1), N_1 news traces are multiplexed, each with a uniformly independent random offset. At the lower priority queue (level 2), N_2 movie traces are multiplexed, again each having uniform and statistically independent phases.

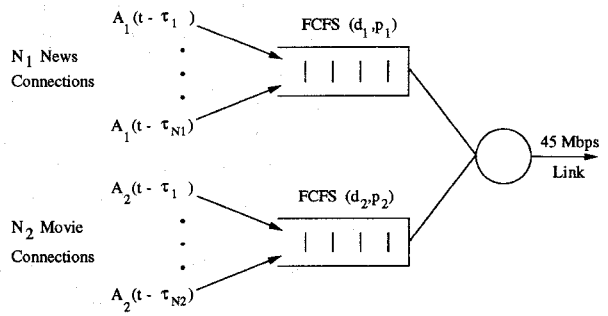


Figure 3: Scenario for Heterogeneous Trace-Driven Simulation

As in the experiments of Section 3.2, in order to provide a given QoS and (d_1, p_1) pair, N_1 can only be so large. Similarly, at the second level, there is a maximum number of connections N_2 that can be multiplexed such that all of the level 2 connections obtain their (d_2, p_2) QoS. The difference with level 2 connections is that they are also affected by the number of level 1 connections (and their respective traffic characteristics) since level 1 connections have priority over level 2 connections. Thus, we are interested in the maximum number of connections N_1 and N_2 that are schedulable such that all level 1 connections obtain the (d_1, p_1) QoS and all level 2 connections obtain the (d_2, p_2) QoS. These experiments therefore consider both heterogeneous traffic mixes and heterogeneous QoS requirements.

As of this writing, the CAC tests of [7, 8, 10, 23] have not been extended to SP schedulers. For the Rate-Variance CAC test, we use the CAC algorithm of Equation (7) in [19]. To compare roughly with effective bandwidth CAC, we linearly interpolate the admissible region from the homogeneous cases.

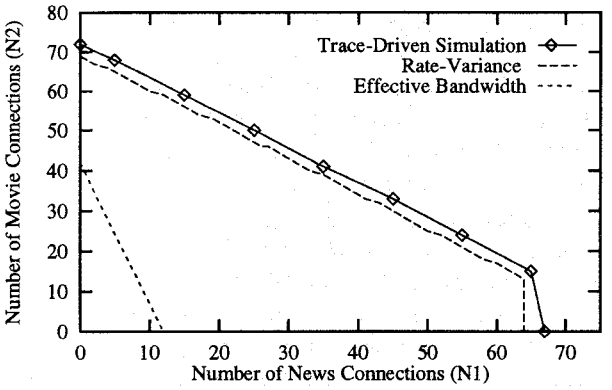


Figure 4: Performance of CAC Tests for Heterogeneous Traffic and SP Scheduler

Figure 4 investigates the case of $d_1 = 10$ msec and $p_1 = 10^{-6}$ for the news connections, and $d_2 = 50$ msec and $p_2 = 10^{-3}$ for the movie connections. The curve labeled "Trace-Driven Simulation" shows the measured schedulable region, or the maximum (N_1, N_2) combinations such that the respective QoS requirements are met. As was the case in the homogeneous experiments, the figure shows that the Rate-Variance CAC test admits most of the connections that can be multiplexed for the required QoS constraints, while the effective bandwidth test is quite conservative.

4 CAC Tests: Approximations and Accuracy

4.1 B Scaling

The experiments of Section 3 evaluate the impact of buffer-size scaling on both the multiplexer's performance as well

as the effectiveness of the different CAC algorithms. The trace-driven simulation experiments indicate that some buffering is of substantial benefit. For example, our simulations show that with 66 multiplexed MPEG connections, 30 msec or approximately 3200 cells worth of buffering decreases the loss probability from .003 to $7 \cdot 10^{-7}$. As well, considering the admissible region and a loss probability of 10^{-6}, Figure 30 msec of buffering increases the admissible region from 49 to 66 connections for a 35% improvement. As noted in Section 3, the incremental advantages of an increased buffer size do not extend indefinitely, but rather decay quickly once the multiplexer has 10 to 20 msec of buffering.

With the importance of buffering in the actual system, CAC tests that take into account buffer size scaling are able to significantly out-perform those that do not. Indeed, the Avg/Peak CAC test of [23] is the most conservative of the tests partially because of its assumption of a bufferless multiplexer. In contrast to [23], the remaining tests consider *buffered* multiplexers, with [7, 8, 10] considering asymptotically large buffers.

We note however, that the aforementioned 35% increase in the admissible region or 4 order-of-magnitude decrease in the loss probability comes at a cost, including the costs of the memory itself, buffer management costs, and potential increased complexity in the CAC tests, since the buffered system is strongly influenced by the traffic stream's autocorrelation structure.

4.2 N Scaling

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 the CAC tests of [8, 10, 19, 23] have this property, effective bandwidth tests such as [7] do not since a stream's effective bandwidth is determined only by the stochastic properties of the source itself, namely its asymptotic log moment generating function, independent of the properties of other sources or the total number of sources being multiplexed. While this property is closely related to the "effective bandwidth" concept, (that a bursty traffic source's resource requirements may be summarized by a single bandwidth), as is evident from Figure 2(b) this lack of N scaling limits the achievable utilization.

4.3 Loss Curve

The relationship between the packet loss probability and the buffer size as in Figure 2(a) is often referred to as the loss curve. While the ultimate goal of a CAC test is to correctly determine the admissible region, many CAC tests have been designed with an intermediary focus on the shape of the loss curve. For example, [8] is motivated by the dominant eigenvalue of Markovian sources to approximate the loss curve with an exponential relationship $p \approx Le^{-\delta B}$. Effective bandwidth schemes also assume an exponential relationship but with $L = 1$.

Our simulation results of Figure 2(a) depict a loss curve that is significantly different than exponential. While long range dependence is a plausible explanation for this [2], the one CAC test that explicitly addresses this issue with a designed sub-exponential loss curve [10] was still noticeably conservative.

We note that the rate-variance CAC test of [19] is able to track this non-exponential loss curve quite well. [19] does not assume traffic streams are long range dependent, but rather uses a general second moment traffic envelope to allocate network resources in accordance with the time scales of the underlying traffic streams.

In [8], trace drive simulations were also performed and the reported loss curve in that work is nearly exponential. Consequently, the Chernoff Dominant Eigenvalue test of [8] was quite accurate for those experiments. As noted above, in our experiments, the loss curve is far from exponential so that the Chernoff Dominant Eigenvalue test is considerably less accurate. We conjecture that this is due to two factors (1) [8] used a video conference trace which likely does not exhibit long-time-scale rate variation, whereas we considered movies and newscasts which do have long-time-scale rate variations due to scene changes; (2) [8] used JPEG- rather than MPEG-compressed video, with the latter having substantially more rate variation on small time scales as well.

4.4 Traffic Model

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

The on-off traffic model used in the CAC test of [23] 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 [13], it is also quite closely tied to the assumption of a bufferless multiplexer, which as described above, has a considerable utilization penalty for a CAC test. Indeed, to take into account the effects of buffering, more information is needed about the traffic streams such as their autocorrelation structure, their maximum rates over various interval lengths [20], or at least a burst length parameter.

While effective bandwidth [16] or Avg/Peak [23] CAC

tests were shown to work well with on-off sources, when these same CAC tests are applied to the compressed video sources we have considered here, the tests exhibit considerable inaccuracies. For traffic streams that exhibit multiple time scale 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 recently proposed traffic model of [20] 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 characterization considered here (see also [18]). For the statistical service evaluated here, a second possibility is to have users directly convey their second moment characteristics to the network. Such models allow for accurate characterization of the admissible region as depicted in Figure 2(b). The downside to such approaches is that more traffic parameters are needed than the three parameters of [13] which also means that policing requires at least multi-level leaky buckets rather than a single leaky bucket (see [20]). Others have suggested the use of the Hurst parameter to specify traffic parameters. However, such a parameter is extremely difficult to police; moreover, the compressed video streams we have studied are quite different than exactly self similar processes (see [19]).

If a refined traffic model is not used, then statistical services will likely yield such low resource utilization for bursty traffic streams that either renegotiated services [15, 31] or measurement based services [12] must be used instead. While such services have their own merits as discussed in the respective works, they unfortunately cannot provides a statistical QoS *guarantee* 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 time scale sources, which, as we have seen here, is a difficult problem even when future arrival statistics *are* known.

Lastly, we note that CAC tests differ in their computational complexity or the number of instructions that must be executed upon the arrival of a new connection 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 [23] giving this issue the most attention.

5 Conclusions

From the results of our trace-driven simulations and CAC experiments with five diverse admission control tests, 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 connections is a crucial component to achieving a high degree of accuracy.
- Observed loss curves (loss probability vs. buffer size) for compressed video sources are quite different than the commonly assumed exponential relationship.
- Refinements of current standard traffic models are required in order to obtain a reasonable statistical multiplexing gain and a statistical QoS guarantee.
- CAC tests that work demonstrably well with exponential on-off sources can suffer from considerable inaccuracies when applied to multiple time scale sources such as compressed VBR video.

We conclude with an illustrative example that reflects the relative accuracies of the evaluated CAC tests. For a multiplexer with a 45 Mbps link capacity and 30 msec of buffering, the trace driven simulations indicate that at most 66 connections can be multiplexed while obtaining a loss probability of 10^{-6}. The Rate-Variance CAC test of [19] admits 64 connections (97% of the possible), both the Chernoff Dominant Eigenvalue CAC test of [8] and the algorithm for homogeneous self-similar streams [10] admit 44 connections (69% of the possible), the Effective Bandwidth CAC test of [7] admits 20 connections (30% of the possible), and the Avg/Peak CAC test of [23] admits 11 connections (17% of the possible).

References

[1] D. Anick, D. Mitra, and M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *Bell System Technical Journal*, 61:1871–1894, 1982.

[2] D. Botvich and N. Duffield. Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. *Queueing Systems, Theory and Applications*, 20(3-4):293–320, 1995.

[3] C. Chang. Stability, queue length, and delay of deterministic and stochastic queueing networks. *IEEE Transactions on Automatic Control*, 39(5):913–931, May 1994.

[4] C. Chang. Sample path large deviations and intree networks. *Queueing Systems, Theory and Applications*, 20(1-2):7–36, 1995.

[5] J. Choe and N. Shroff. A new method to determine the queue length distribution at an ATM multiplexer. In *Proceedings of IEEE INFOCOM '97*, Kobe, Japan, April 1997.

[6] G. Choudhury, D. Lucantoni, and W. Whitt. Squeezing the most out of ATM. *IEEE Transactions on Communications*, 44(2):203–217, February 1996.

[7] C. Courcoubetis and R. Weber. Effective bandwidths for stationary sources. *Probability in Engineering and Informational Sciences*, 9(2):285–294, 1995.

[8] A. Elwalid, D. Heyman, T. Lakshman, D. Mitra, and A. Weiss. Fundamental bounds and approximations for ATM multiplexers with applications to video teleconferencing. *IEEE Journal on Selected Areas in Communications*, 13(6):1004–1016, August 1995.

[9] A. Elwalid and D. Mitra. Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. *IEEE/ACM Transactions on Networking*, 1(3):329–43, June 1993.

[10] Z. Fan and P. Mars. Accurate approximation of cell loss probability for self-similar traffic in ATM networks. *Electronics Letters*, 32(19):1749–1751, September 1996.

[11] D. Ferrari and D. Verma. A scheme for real-time channel establishment in wide-area networks. *IEEE Journal on Selected Areas in Communications*, 8(3):368–379, April 1990.

[12] S. Floyd. Comments on measurement-based admissions control for controlled-load services, July 1996. Lawrence Berkeley Laboratory Technical Report.

[13] ATM Forum. ATM user-network interface specification. version 3.1, September 1994. ATM Forum document.

[14] M. Garret and W. Willinger. Analysis, modeling and generation of self-similar VBR video traffic. In *Proceedings of ACM SIGCOMM'94*, pages 269–280, London, UK, August 1994.

[15] M. Grossglauser, S. Keshav, and D. Tse. RCBR: A simple and efficient service for multiple time-scale traffic. To appear in *IEEE/ACM Transactions on Networking*.

[16] R. Guérin, H. Ahmadi, and M. Naghshineh. Equivalent capacity and its application to bandwidth allocation in high-speed networks. *IEEE Journal on Selected Areas in Communications*, 9(7):968–981, September 1991.

[17] G. Kesidis, J. Walrand, and C. Chang. Effective bandwidths for multiclass Markov fluids and other ATM sources. *IEEE/ACM Transactions on Networking*, 1(4):424–428, August 1993.

[18] E. Knightly. H-BIND: A new approach to providing statistical performance guarantees to VBR traffic. In *Proceedings of IEEE INFOCOM '96*, pages 1091–1099, San Francisco, CA, March 1996.

[19] E. Knightly. Second moment resource allocation in multi-service networks. In *Proceedings of ACM SIGMETRICS '97*, pages 181–191, Seattle, WA, June 1997.

[20] E. Knightly and H. Zhang. D-BIND: An accurate traffic model for providing QoS guarantees to VBR traffic. *IEEE/ACM Transactions on Networking*, 5(2):219–231, April 1997.

[21] J. Kurose. On computing per-session performance bounds in high-speed multi-hop computer networks. In *Proceedings of ACM SIGMETRICS '92*, pages 128–139, Newport, RI, June 1992.

[22] A. Lazar, G. Pacifici, and D. Pendarakis. Modeling video sources for real time scheduling. *ACM Multimedia Systems Journal*, 1(6):253–266, April 1994.

[23] T. Lee, K. Lai, and S. Duann. Design of a real-time call admission controller for ATM networks. *IEEE/ACM Transactions on Networking*, 4(5):758–765, October 1996.

[24] W. Leland, M. Taqqu, W. Willinger, and D. Wilson. On the self-similar nature of Ethernet traffic. *IEEE/ACM Transactions on Networking*, 2(1):1–15, February 1994.

[25] S. Li and C. Hwang. Queue response to input correlation functions: Discrete spectral analysis. *IEEE/ACM Transactions on Networking*, 1(5):552–533, October 1993.

[26] M. Reisslein and K. Ross. Call admission for prerecorded sources with packet loss. *IEEE Journal on Selected Areas in Communications*, 15(6):1167–1180, August 1997.

[27] O. Rose. Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems. In *Proceedings of IEEE Conference on Local Computer Networks*, pages 397–406, Minneapolis, MN, October 1995.

[28] N. Shroff and M. Schwartz. Improved loss calculations at an ATM multiplexer. In *Proceedings of IEEE INFOCOM '96*, pages 561–568, San Francisco, CA, March 1996.

[29] P. Skelly, M. Schwartz, and S. Dixit. A histogram-based model for video traffic behavior in an ATM multiplexer. *IEEE/ACM Transactions on Networking*, 1(4):446–459, August 1993.

[30] H. Zhang and E. Knightly. Providing end-to-end statistical performance guarantees with bounding interval dependent stochastic models. In *Proceedings of ACM SIGMETRICS'94*, pages 211–220, Nashville, TN, May 1994.

[31] H. Zhang and E. Knightly. RED-VBR: A renegotiation-based approach to support delay-sensitive VBR video. *ACM Multimedia Systems Journal*, 5(3):164–176, May 1997.

[32] Z. Zhang, D. Towsley, and J. Kurose. Statistical analysis of generalized processor sharing scheduling discipline. *IEEE Journal on Selected Areas in Communications*, 13(6):368–379, August 1995.