Determining Priority Queue Performance from Second Moment Traffic Characterizations

Edward W. Knightly

TR-96-022

June 1996

Abstract

A crucial problem to the efficient design and management of integrated services networks is how to best allocate and reserve network resources for heterogeneous and bursty traffic streams in multiplexers that support prioritized service disciplines. In this paper, we introduce a new approach for determining per-connection QoS parameters such as delay-bound violation probability and loss probability in multi-service networks. The approach utilizes a traffic characterization that consists of the variances of a stream's rate distribution over multiple interval lengths, which captures its burstiness properties and autocorrelation structure. The resource allocation scheme is based on application of the Central Limit Theorem over intervals, together with use of stochastic delay-bounding techniques; it results in simple and efficient algorithms for determining QoS parameters. We perform experiments with long traces of MPEG-compressed video and show that the new scheme is accurate enough to capture most of the inherent statistical multiplexing gain, achieving average network utilizations of up to 90% for these traces.

1 Introduction

Bursty traffic sources that require a Quality of Service (QoS) guarantee in terms of loss and delay are emerging as one of the most important types of traffic in future integrated services networks. Because of the variable bit rate nature and the multiple-time-scale correlation characteristics of many realistic sources (e.g., [8],[19]), it is a difficult problem to decide how much resources need to be allocated to individual connections such that each connection obtains the performance that it requires. This problem is exacerbated when different connections require different services, such as 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) or Earliest Deadline First (EDF).

In this paper, we introduce a new scheme for allocating resources to heterogeneous, bursty traffic sources that are multiplexed with prioritized service disciplines. The scheme is based on a simple traffic characterization that consists of the variances of the streams' rates over multiple interval lengths. This traffic characterization is general enough to capture the full range of second order statistics for bursty correlated traffic streams. We show how this characterization of the variances of the arrival-rate distribution over intervals of different length captures the intuitive property that the variance tends to decrease with increasing interval length, indicating that over longer and longer interval lengths, the rate distribution becomes more and more concentrated at the mean rate. We show empirical curves of the rate's variance vs. interval length for several long traces of VBR compressed video in order to show how this characterization portrays the burstiness properties of realistic network traffic sources.

Based on the second moment characterization of traffic streams, we introduce new resource allocation algorithms that provide simple-to-compute approximations for network performance metrics such as packet-loss probability and delay-bound-violation probability. Specifically, we build on the stochastic delay bounds of [16] and [18] and introduce techniques for application of the Central Limit Theorem over intervals. The techniques allow for *direct* calculations of QoS from the variance characterization without requiring that the traffic characterization be mapped to other processes such as the MMPP. Moreover, we derive these performance bounds in an environment that allows for both heterogeneous traffic streams and heterogeneous performance requirements, considering the case of priority service disciplines that are well suited for providing integrated network services.

Lastly, we investigate the effectiveness of the new scheme via experiments with trace-driven simulation. We utilize 30-minute traces of MPEG-compressed and simulate multiple connections aggregating at a queue. For a given set of connections, we compare the QoS actually obtained by the streams in the trace-driven simulation with the QoS that our algorithms predict from the streams' second moment characteristics. Our results indicate that for these bursty VBR video sources, the new variance based resource allocation scheme is accurate enough to capture most of the achievable statistical multiplexing gain.

2 Related Work

In the literature, second-moment characterizations of traffic streams have been used previously for performance analysis of networks. Here, we highlight several examples, including the use of indices of dispersion, characterization of traffic by its power spectrum, and use of the Central Limit Theorem.

In [6] and [11], traffic is characterized by Indices of Dispersion for Intervals (IDI) with an aim towards performance analysis of voice and data multiplexers. Particularly, in [11], the IDI characterization is mapped to MMPP parameters for approximations of multiplexer performance approximations. A similar approach was taken in [12] which approximates a general second moment autocorrelation function by that of a two-state Markov-Modulated Poisson Process and determines queue performance based on the MMPP characterization.

In [13], [21], [22], [23], second-order traffic statistics are used to calculate queue performance metrics by considering the power spectrums of the individual input streams. For example, in [23], the impact of each individual component of a stream's power spectrum on queue performance was analyzed. Specifically, a stream's power spectrum is mapped to a Markov Chain transition matrix, from which the queue performance is obtained and related back to the properties of the original power spectrum.

Our work differs from the above in that we do not map the arrival process' variances over intervals to a specific random process such as the MMPP. Rather, we determine the multiplexer's performance metrics directly from the second moment traffic characterization. Indeed, as discussed below, our techniques for determining queue performance are quite different from those above. Moreover, we consider queue performance for *priority* service disciplines which were not considered in these previous works.

The Stop-and-Go service discipline is proposed and analyzed in [10]. With Stop-and-Go's framing structure (time is slotted into frames), specific properties of the traffic streams are maintained even as connections traverse multiple network hops. This allows provisioning of *end-to-end* performance bounds in terms of loss and delay, and delivers a bounded delay-jitter service as well. To provide statistical QoS bounds, an algorithm is presented in [9] which utilizes the Central Limit Theorem. Specifically, the scheme is termed "duration-limited" statistical multiplexing because the mean and variance of arrivals over a single frame-time (the Stop-and-Go frame time) are used to calculate the loss probability. However, if traffic streams have multiple-time-scale characteristics, then considering a single time scale for statistical multiplexing may not capture enough of the achievable statistical multiplexing gain. Hence, our approach exploits variances over *multiple* interval lengths to calculate statistical performance bounds, and is able to take into account the correlation structure of the arrival process.

While we share a second-order traffic characterization with the above works, our focus here is quite different. In particular, we use new analysis techniques that are analogous to the *deterministic* delay bounding techniques of [3], which are extended in [24] to provide necessary and sufficient deterministic delay bounds for prioritized service disciplines such as Static Priority (SP) and Earliest Deadline First (EDF). Extending

performance analysis techniques to such service disciplines is crucial to providing integrated services, i.e., to providing connections with a variety of delay bounds, loss probabilities, and so on. While our work has some analogies to [3] and [24], it differs considerably in that we are considering *statistical* performance guarantees rather than deterministic, and hence we are exploiting a statistical multiplexing gain. However, we view our work as complimentary to other second moment works such as [13] and [11], in that our work reinforces the notion that second moment traffic characterizations are quite sufficient for capacity allocation to heterogeneous bursty traffic streams.

Hence, our focus here is to make *a priori*, simple-to-compute determinations of a connection's QoS based directly on the variances of the connection's rate over multiple interval lengths. This QoS will also depend on the service discipline at the network node and the traffic characterizations of other streams traversing the node.

The remainder of this paper is organized as follows. In Section 3, we describe the rate-variance over intervals traffic characterization and show how it provides a general second order description of traffic streams. We derive performance bounds in priority schedulers based on stochastic bounding techniques and show how application of the Central Limit Theorem over intervals can approximate these bounds. In Section 4, we perform a set of experiments based on 30-minute traces of MPEG compressed video. We compare the QoS obtained in the trace-driven simulations with those predicted by the algorithms of Section 3. Finally, in Section 5, we conclude.

3 Performance Bounds from 2^{nd} Moment Traffic Characterizations

In this section, we describe the analytical foundations for using a second moment traffic characterization to calculate queue performance metrics such as probability of delay-bound violation and probability of loss. The techniques apply to priority service disciplines (i.e., service disciplines that serve packets in non-FCFS order) that are well suited to providing network clients with heterogeneous QoS requirements. The traffic characterization is also quite general in that it allows for an arbitrary autocorrelation structure of individual connections and heterogeneous characteristics among connections.

3.1 Variance-Interval Traffic Characterization

For simplicity, we consider time to be slotted and consider a stationary random process that is described by a sequence of random variables $\{X_1, X_2, X_3, \cdots\}$. To relate the analysis below to the empirical investigations of Section 4, the X_i's may be considered, without loss of generality, to represent the frame sizes of a compressed-

video stream. Otherwise, the X_i's may represent the distribution of arrivals over the smallest time scale of relevance to the stream.

Denote the first and second moments of the frame-size distribution by EX and EX^2 respectively so that the variance of the frame size distribution, $Var(X)$, is $EX^2 - (EX)^2$. The stream's mean rate is then $m = EX/T$ where T is the duration of the time slot or frame time.

We allow the sequence $\{X_1, X_2, X_3, \cdots\}$ to have an arbitrary autocorrelation structure and denote the *variance* of the distribution of the total number of arrivals over n consecutive frames by

$$Var(X_i + X_{i+1} + \cdots + X_{i+n-1}). \tag{1}$$

This variance can be normalized to the length of the interval to get the variance of the *rate* distribution over the respective interval length. We denote this rate variance by

$$RV(n) = Var(\frac{X_i + X_{i+1} + \cdots + X_{i+n-1}}{Tn}),\ n \geq 1. \tag{2}$$

With abuse of notation, we will refer to $RV(t_n = nT)$ and $RV(n)$ interchangeably, with the distinction made by the respective arguments, t_n or n.

$RV(n)$ thus captures the second moment correlation structure of the process $\{X_1, X_2, X_3, \cdots\}$ in the same way as the autocorrelation function EX_iX_{i+n} or the power spectrum as in [23]. We use the variance of the rate over intervals rather than such other second moment traffic characterizations simply because it relates more directly to our resource allocation algorithms presented in Section 3.2. Moreover, as shown below, characterizing a stream by the variance of its rate as a function of interval length provides an intuitive representation of the stream.

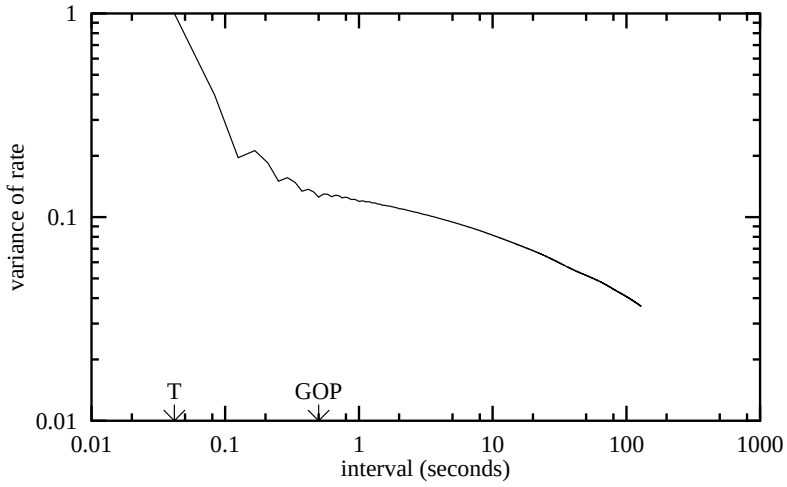


Figure 1: Variance of Rate over Multiple Interval Lengths for Movie Trace

Figure 1 shows the rate's variance vs. interval length for a 30-minute trace of an MPEG-compressed movie sequence. The trace is of an action movie (a "James Bond" film) taken from [25]. The movie is digitized to 384x288 pels and compressed at 24 frames per second using the MPEG compression algorithm [7] with frame pattern *IBBPBBPBBPBB*. Further details of the trace can be found in [25].

Thus, for the 30-minute sequence of frame sizes, the figure shows the empirical variance of the rate as given by Equation (2) as a function of the interval length on a log-log scale. The rate-variance depicted on the vertical axis is normalized to the frame-size variance so that $RV(t_1 = 1/24) = 1$. The horizontal axis depicts the interval length in seconds, which is nT, or the interval length in frame-times, n (as in Equation (2)), multiplied by the frame time T, $\frac{1}{24}^{th}$ of a second in this case. The figure depicts intervals of up to 100 seconds rather than up to the length of the trace so that there are an adequate number of sample points to determine the sample variance.

We make several observations about the figure. First, note that the rate's variance tends to decrease with increasing interval lengths. In other words, over wider and wider interval-lengths, the distribution of the stream's rate becomes more and more concentrated at the mean rate. Indeed, the rate's variance has fallen considerably even by the MPEG Group Of Pictures time of 0.5 seconds. Such a characteristic can be viewed as a stochastic analogue of the deterministic characterization in [17]. That deterministic model characterizes traffic with bounding or worst-case rates over multiple interval lengths. It captures the property that over longer interval lengths, sources can be upper-bounded by a rate that is lower than their peak rate, and closer to their long-term average rate. Here, we are interested in stochastic rather than worst-case properties of streams so that we can achieve a statistical multiplexing gain, and we use *second moment* traffic characterizations rather than deterministic characterizations.

Lastly, we note that the shape of the curve of Figure 1 has implications on the relevant time-scales of the stream's correlation structure and whether the stream exhibits long-range dependencies. We defer discussion of this issue to Section 3.4 and note here that we will directly use the streams' variances in the resource allocation algorithm discussed below, regardless of the specific shape of the variance-interval curve.

3.2 Bounding Delay in a Static Priority Scheduler

As shown in Figure 2, a static priority scheduler consists of a number of prioritized FCFS queues, where, as in [28], each queue has an associated delay bound and probability of delay bound violation. At connection setup time, each connection is assigned a priority level that is based on the connection's requested QoS, including its requested end-to-end delay bound and probability of delay bound violation. During data transmission, each packet of the connection is then serviced at its pre-established priority level. Thus, a static priority service discipline has an advantage as compared to many other service disciplines in its simplicity of implementation. For example, in the Earliest Deadline First (EDF) and Generalized Processor

Sharing (GPS) service disciplines, when the scheduler determines which packet to service next, it must sort the packets according to either due dates or some other priority index. Alternatively, for static priority, the packet at the head of the highest-priority non-empty queue is always serviced next.

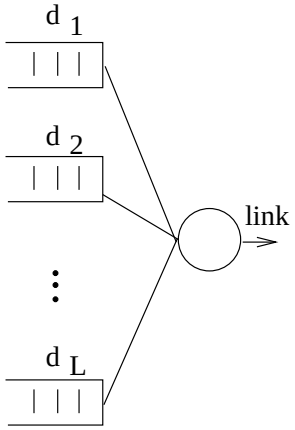


Figure 2: Static Priority Scheduler

In [18], Kurose introduced a framework for providing statistical network performance bounds based on characterizing sources with a family of bounding random variables. In that work, stream j is described by a family of random variables $\{(B_j(t_1), B_j(t_2), ...\}$ that stochastically bound the stream's arrivals over the respective interval lengths t_k. A random variable B is said to be stochastically larger than a random variable S (denoted $B \succeq_{st} S$) if and only if $Prob(B > x) \geq Prob(S > x)$ for all x.

A stochastic bound on delay for a FCFS multiplexer with link speed l serving N connections characterized by their respective families of bounding random variables is shown in [18] to be:

$$Prob\{D > d\} \leq \max_{0 \leq t_k \leq \beta} Prob\{\sum_{j=1}^{N} B_j(t_k) - lt_k \geq ld\} \tag{3}$$

where D is a random variable representing the delay of a randomly selected cell and β is an upper bound on the busy period (calculation of β is discussed in [3],[18]).

In order to apply the framework of [18] to priority service disciplines that are better suited than FCFS for providing integrated services, we use the necessary and sufficient *deterministic* schedulability tests of [24] as a foundation for providing *stochastic* bounds on delay and loss. For a static priority scheduler, the following theorem provides a stochastic bound on the probability that a random cell of connection j in priority level p is delayed beyond its bound d_p.

Theorem 1 *Assume a Static Priority scheduler has L priority levels, where priority level p has an associated delay bound d_p. Let C_q be the set of connections at level q, and let the distribution of the number of cells*

transmitted by the j^{th} connection in C_q over intervals of length t be stochastically bounded by the random variable $B_{q,j}(t)$. For a link speed l, the delay violation probability for a random cell in level p is bounded by:

$$Prob\{D_p > d_p\} \leq \max_{0 \leq t_k \leq \beta_p} Prob\{\sum_{j \in C_p} B_{p,j}(t_k) + \sum_{q=1}^{p-1} \sum_{j \in C_q} B_{q,j}(t_k + d_p) - lt_k \geq ld_p\} \quad (4)$$

where β_p is a bound on the priority-p busy period.

Proof: Consider a random cell from level p that arrives in a given busy period. Both cells from priority level-p that arrive before the random cell, and cells that arrive in a higher priority level will contribute to the cell's delay. For the previous t_k seconds, the distribution of the number of level-p cells that arrive (and hence contribute to the random cell's delay) is stochastically bounded by

$$\sum_{j \in C_p} B_{p,j}(t_k), \quad (5)$$

even in the worst case that all of these cells arrive within the random cell's busy period. This was demonstrated in [18] for the FCFS result of Equation (3). For the previous t_k seconds, cells from higher priority levels may also arrive and contribute to the random cell's delay. Moreover, higher priority cells may contribute to a delay bound violation for the random cell, even if they arrive *after* the random cell by up to d_p seconds. The distribution of this total number of higher-priority cells is bounded by

$$\sum_{q=1}^{p-1} \sum_{j \in C_q} B_{q,j}(t_k + d_p). \quad (6)$$

Since lt_k cells are served in this interval, that many fewer are queued. Thus, a stochastic bound on the total number of cells that arrived no more than t_k seconds ago, have priority over the random cell, and are queued is given by

$$\sum_{j \in C_p} B_{p,j}(t_k) + \sum_{q=1}^{p-1} \sum_{j \in C_q} B_{q,j}(t_k + d_p) - lt_k. \quad (7)$$

Since cells that arrive in other busy periods will not contribute to the random cell's delay, maximizing over all interval lengths less than the maximal busy period yields the result of Equation (4). □

The computational complexity of Equation (4) may limit its usefulness in an on-line resource allocation algorithm. In particular, the second moment approximation proposed in the next section is motivated by the following difficulties with Equation (4).

- Equation (4) requires knowledge of a family of bounding distributions for each traffic stream. Such a complex traffic characterization will likely be difficult to obtain in practice.
- The sum of random variables in Equation (4) must be calculated via a convolution or a Fast Fourier Transform, which may be too costly for on-line resource allocation systems such as connection admission control, especially since these convolutions must be performed over multiple interval lengths in order to maximize Equation (4).

3.3 Variance-Based Resource Allocation

We make the following two observations about Theorem 1 in order to allocate resources according to second-moment characterizations of traffic. First, the sums of the random variables in Equation (4) that constitute the calculation of the tail probability of the delay are made up of statistically *independent* random variables. This property is likely to be true in practice for many types of workloads. Moreover, this property is essential to any scheme that extracts a statistical multiplexing gain. Indeed, if this property is not satisfied and, for example, streams tend to synchronize, resources will need to be allocated using worst-case schemes such as in [17].

Second, within a single stream, an arbitrary correlation structure is allowed, so that no restrictions are made on the burstiness properties of streams or the time-scales of their autocorrelation. The same is true of the aggregate process, which also does not have restrictions on its correlation structure.

Thus we have a large number of statistically independent streams and Theorem 1 allows us to express the probability of delay-bound violation (or loss) as a function of sums of independent random variables over intervals. Our approach is therefore to apply the Central Limit Theorem (CLT) over intervals in order to efficiently approximate Equation (4). The CLT states that if the random variable Y_i has mean μ_i and variance σ_i^2, and the Y_i's are mutually independent, then the distribution of the sum $Y_1+Y_2+\cdots Y_n$ converges to a normal distribution with mean $\sum_{i=1}^{n}\mu_i$ and variance $\sum_{i=1}^{n}\sigma_i^2$. The theorem can be shown to hold under fairly general conditions on the Y_i's, including for example, the Lindeberg sufficient condition [2].

Proposition 1 *Assume a Static Priority scheduler has link speed l and has L priority levels, where priority level p has an associated delay bound d_p. Let C_q be the set of connections at level q, and let the j^{th} connection in C_q be characterized by its mean rate $m_{q,j}$ and its rate's variance over intervals $RV_{q,j}(t_k)$ as in Equation (2). With application of the Central Limit Theorem over intervals, the delay-bound violation probability for a random cell in level p is approximated by:*

$$Prob\{D_p > d_p\} \approx \max_{0\le t_k\le \beta_p} Prob\{\ \hat{B}_p(t_k) + \hat{B}_{q<p}(t_k+d_p) - lt_k \ge ld_p\} \tag{8}$$

Where according to the CLT, the distributions of the random variables in Equation (8) are given by

$$\hat{B}_p(t_k) \sim N(\sum_{j\in C_p} t_k m_{p,j},\ \sum_{j\in C_p} t_k^2 RV_{p,j}(t_k)) \tag{9}$$

$$\hat{B}_{q<p}(t_k+d_p) \sim N(\sum_{q=1}^{p-1}\sum_{j\in C_q}(t_k+d_p)m_{q,j},\ \sum_{q=1}^{p-1}\sum_{j\in C_q}(t_k+d_p)^2 RV_{q,j}(t_k+d_p)) \tag{10}$$

with $N(\mu,\sigma^2)$ denoting a normal distribution with mean μ and variance σ^2.

Proposition 1 therefore applies the CLT over multiple interval lengths in order to approximate the delay-violation probability. The importance of characterizing traffic by properties of arrivals over intervals can be

seen not only from Equation (4), but also from direct calculation of the total queue length as a function of the arrivals and link speed:

$$q(s) = \max_{\tau \leq s} \{ \sum_{j=1}^{N} A_j[\tau, s] - l(s - \tau) \} \tag{11}$$

where $q(s)$ denotes the queue length at time s and $A_j[\tau, s]$ represents the arrivals of connection j in the interval $[\tau, s]$. Hence, queue lengths are determined by the number of arrivals over intervals in relation to the link speed, and Proposition 1 approximates the tail probability of the delay by using the first two moments of each streams arrivals over intervals.

3.4 Implications for Traffic Models and CAC

One application of the above resource allocation scheme is in network planning and design such as buffer sizing for switches and routers. In such a case, the $RV(n)$ characterization of the expected workloads can be used directly in the calculation of the terms in Proposition 1.

For an on-line resource allocation system such as a Connection Admission Control (CAC) algorithm, or for an adaptive resource management algorithm, Equation (8) has the advantage that it is a simple computation that does not require, for example, convolutions or large matrix computations. However, in order to integrate the scheme with a network signaling protocol, streams must be able to specify their traffic characteristics more concisely than the $(m, RV(n))$ characterization. Several possibilities are described below.

First, the $(m, RV(n))$ characterization can be *inferred* or bounded based on the streams' specified parameters, which may be worst case parameters such as the (σ, ρ) leaky bucket model [3] or the D-BIND model's rate-interval pairs [17]. For example, in [16], a scheme is presented to upper bound a stream's variance over intervals based on its worst-case parameters as given by the D-BIND model. Such a scheme has the advantage that the stochastic properties of the stream (its variances over intervals) can then be policed by using the appropriate deterministic filters at the network edge.

Second, if a specific shape of the $RV(n)$ rate's variance-interval curve is assumed, then the curve can be specified to the network more concisely, either by specifying several points on the curve and interpolating, or by mapping parameters from other processes (e.g., a MMFS) to the $RV(t_k)$ characterization. For example, consider Figure 3 which depicts empirical $RV(n)$ curves for various traces of compressed video.

The figure depicts $RV(t_k)$ as given by Equation (2) on a log-log scale, just as in Figure 1 for the MPEG trace of the action movie. All of the traces in the figure are of MPEG-compressed video except for the one labeled "JPEG" which uses only intraframe compression in a manner similar to motion JPEG. The Star Wars traces are from Bellcore [8], the action movie (Bond) and news traces are from the University of Wuerzburg [25], and the advertisements and lecture traces are from Berkeley [17]. Further details of the traces may be found in the respective references. Based on the shape of the curves in Figure 3, we observe

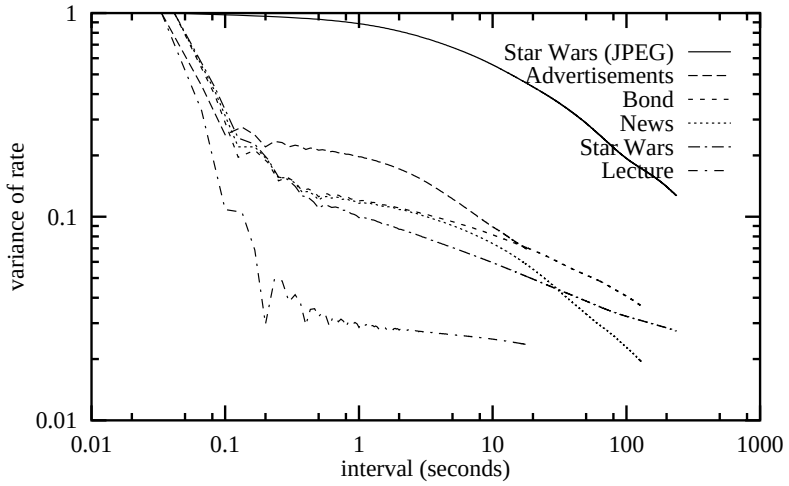


Figure 3: Variance of Rate over Multiple Interval Lengths for Various Traces

that using two or three piece-wise linear segments to bound or approximate the $RV(t_k)$ curve (on a log-log scale) would closely characterize the second moment properties of these streams.

Lastly, we note that the slope at which the rate's variance decreases with interval length, as depicted on Figure 3's log-log scale, has implications on the time-scales of the source's correlation structure and whether or not the stream exhibits long-range dependence. While it is well-known that streams' autocorrelation structure has an impact on queue performance as compared to a renewal process, it has been the subject of ongoing debate which correlation time-scales are of interest for determination of the queue performance (see [5] and [26] for opposing viewpoints).

For example, as noted in [8] and [20], an uncorrelated process with $EX_m X_{m+n} = 0$ for $n \neq 0$ has $RV(n) = Var(X)/n$. As well, short-range-dependent processes have an $RV(n)$ that asymptotically decreases with $1/n$. However for long range dependent processes, $RV(n)$ asymptotically decreases as $n^{-\beta}Var(X)$ for $0 < \beta < 1$. In other words, on a log-log scale, a long-range dependent process has an $RV(n)$ curve with slope greater than -1. While none of the curves in Figure 3 behave as strongly in this manner as the Ethernet trace of [20], whether or not long-range dependence is present can impact the type of parameters that one chooses to concisely describe $RV(n)$.

Our point here is not to speculate on the existence of long-range dependence in these streams, but rather to show the relationship of the $RV(n)$ traffic characterization to this topic. Hence, our scheme allocates resources according to the second moment traffic characterization, regardless of its specific shape and time-scales or the properties of the underlying random process. Typical shapes of $RV(n)$ will however have an impact in mapping $RV(n)$ to concise traffic specifications.

4 Empirical Investigations

In this section, we evaluate the resource allocation scheme described in the previous section with a 30-minute trace of MPEG-compressed video. We perform a set of experiments using trace-driven simulations considering various scenarios with different loads, QoS parameters, and so on. We compare the queue performance obtained in these trace-driven simulations with those predicted by the analytical results in the previous section.

4.1 Trace-driven Simulation Scenario

The trace is of the action movie sequence described in Section 3.1. For the simulations, we fragment each video frame into ATM cells and transmit the cells in equally-spaced intervals over the frame time, $\frac{1}{24}^{th}$ of a second.

Throughout the experiments, we focus on the following two performance metrics. The first is average utilization of the link. For a given video trace that consists of F frames, we define its average rate as

$$\gamma = \frac{\sum_{i=1}^{F} x_i}{TF} \tag{12}$$

where x_i is the size of the i^{th} frame. In other words, γ is the total number of bits transmitted by the source divided by the duration of the transmission. The average utilization of the link is therefore

$$\frac{\sum_{n=1}^{N} \gamma_n}{l} \tag{13}$$

for N multiplexed connections. For the simulation, this average utilization is also the total number of bits transmitted by all of the sources for the duration of the simulation, divided by the total number of bits that the server can transmit during the duration of the simulation (the link speed multiplied by the simulation time).

Our second performance metric is the total fraction of packets that either violate their delay bound or are dropped due to buffer overflow. We set the buffer size to be equal to the delay bound multiplied by the link speed, and drop packets that arrive to a full buffer. Note that if the buffer size was larger, these packets would violate their delay bounds rather than being dropped. We consider a range of delay bounds d and corresponding buffer sizes, and report the empirical $Prob\{Delay > d\}$ or $Prob\{loss\}$ as the measured fraction of packets that are dropped due to buffer overflow. We refer to this probability as p.

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 time through the traces is used to ensure that the multiplexer is in steady state, and the second time through the traces is used to collect results. Multiple simulations are performed with independent start times, and average results are reported.

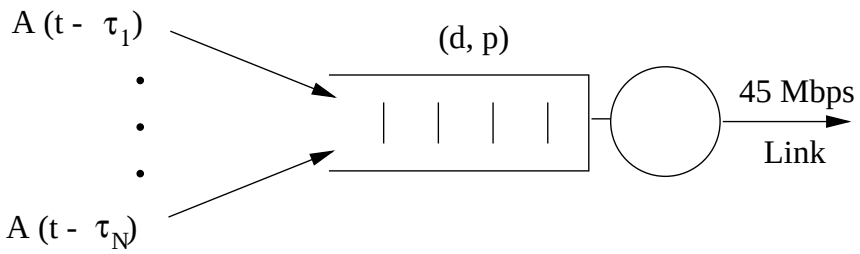


Figure 4: Trace-Driven Simulation Scenario

The trace-driven simulation scenario is shown in Figure 4. For a given 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 (30 minutes). We consider a single QoS pair for all connections that is 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 queue for each simulation is set to be equal to $l \cdot d$ bits. Hence, the outputs of our experiments consist of three-tuples (N, d, p). We then compare the empirical results obtained by multiplexing randomly offset traces with those given by the resource allocation scheme in Section 3.

Clearly, in order to provide a given QoS or (d, p) pair, N can only be so large. The purpose of the resource allocation algorithm of Section 3 is to determine ahead of time the maximum number of connections N that can be multiplexed so that all connections obtain the given QoS. The trace-driven simulations then show the empirical answer to this question by running multiple simulations with various N and measuring the resulting QoS. Thus, comparison of the resource allocation scheme with the trace-driven simulation shows how well the algorithm of Proposition 1 takes the variance characterizations of sources and predicts the resulting QoS.

4.2 Number of Admissible Connections

Figure 5 shows a plot of the average utilization of the multiplexer as a function of the delay bound. Specifically, the vertical axis depicts average utilization which is directly proportional to the number of connections N as given by Equation (13). This N is the maximum number of connections that can be multiplexed such that all connections obtain their required QoS. The QoS is depicted on the horizontal axis with the guaranteed delay bound d. As described above, there is also a corresponding loss or delay-bound violation probability p. Figure 5(a) depicts the case of $p = 10^{-3}$ and Figure 5(b) depicts the case of $p = 10^{-6}$.

In both Figures 5(a) and 5(b), three curves are depicted: the trace-driven simulation, the variance-based resource allocation scheme, and as a base-line, a deterministic, or worst-case resource allocation scheme. The curve labeled "deterministic ($p = 0$)" depicts the maximum number of admissible connections such that

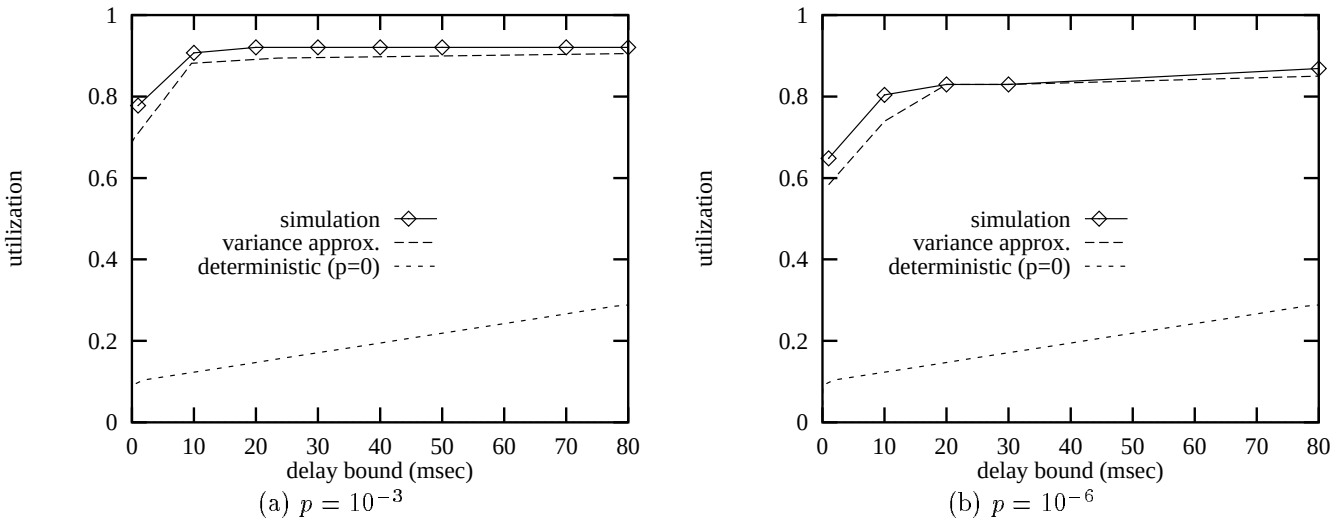


Figure 5: Utilization vs. Delay Bound

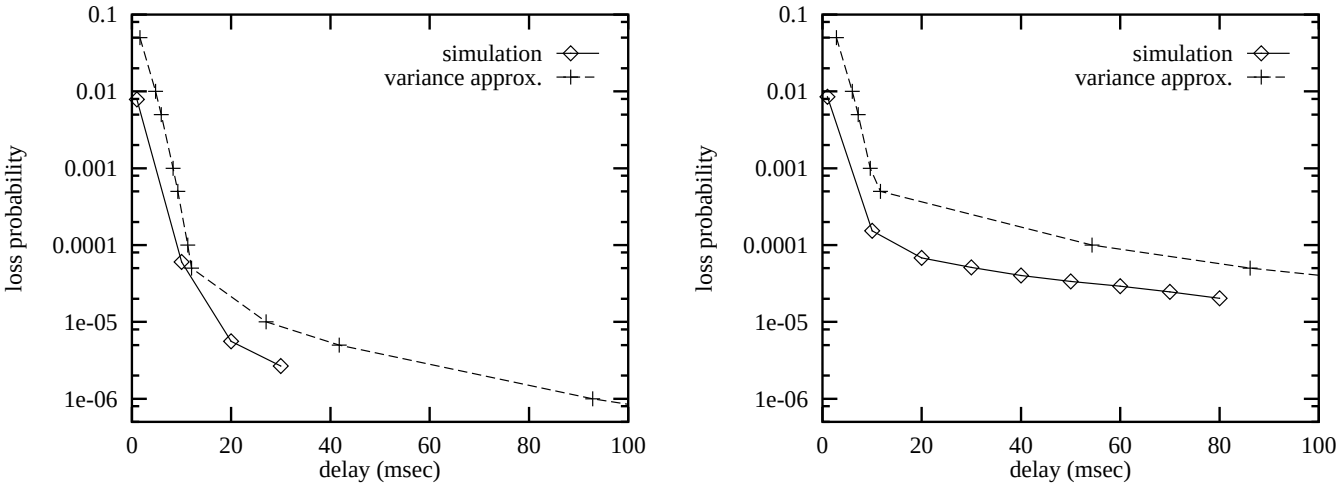
all connections obtain a *deterministic* service such that no packets violate their delay bounds and none are dropped due to buffer overflows: in other words, $p = 0$. This curve is as calculated in [17] and [27] and serves as a base-line for determining how much statistical multiplexing gain is achieved. Thus, we note from Figure 5 that there is a remarkable utilization increase that can be achieved due to the effects of statistical multiplexing. Indeed, the utilizations for the deterministic case are in the range of 10% to 30% while those for the statistical case are in the range of 60% to 90%. Hence, a multi-fold increase in utilization is achievable for network clients that are tolerant to an occasional packet drop.

Next, we focus on the differences between the upper two curves of Figure 5 labeled "simulation" and "variance approximation". For the trace-driven simulation, a point on the curve indicates the maximum number of connections that could be multiplexed so that for a given delay bound d, at most a fraction of 10^{-3} packets violated that delay bound (or a fraction 10^{-6} for Figure 5(b)). Contrastly, the "variance approximation" curve uses the variances-over-intervals traffic characterization, together with the resource allocation scheme of Section 3, and makes an *a priori* determination of how many connections can be multiplexed such that all connections obtain the desired QoS. If Section 3's resource allocation scheme is to be used for capacity allocation, then one would desire that the "variance approximation" curve be as close as possible to, but not greater than, the "simulation" curve. Indeed, if the resource allocation scheme allows for more connections than can actually be supported, then violations of the promised QoS will occur, an undesirable situation for a guaranteed-services network. Hence, Figure 5 shows that our scheme is able to achieve *most* of the achievable statistical multiplexing gain, coming quite close to the results of the trace-driven simulation. Moreover, as desired, the scheme errs slightly on the conservative side rather than over-committing resources.

Finally, we take note of the general shapes of the "simulation" and "variance approximation" curves of Figure 5. Since the delay bound depicted on the horizontal axis corresponds to a buffer size ld, we can see the impact of increasing the buffer size at the network nodes. Both Figures 5(a) and 5(b) show a considerable increase in utilization for minor increases in buffer size for delay bounds in the range of up to 10 msec for Figure 5(a) and 20 msec for Figure 5(b) (10 msec of buffering corresponds to roughly 1000 cells on a 45 Mbps link). However, after the respective delay bounds of 10 and 20 msec, the curves flatten considerably, indicating that increasing the buffer size further is of relatively little use. While the flattening of this curve is most likely due to the longer time-scale characteristics of the traffic streams, we note here that our variance-based resource allocation is able to follow the knee of this curve and to approximate it quite closely.

4.3 Loss Probability and Buffer Size

Figure 6 further describes the results of the experiments by plotting the loss or delay-bound violation probability p vs. delay bound d for a fixed number of connections N and hence a fixed utilization. Figure 6(a) depicts the case of 66 multiplexed connections for an average utilization of 85.6%, and Figure 6(b) depicts the case of 68 multiplexed connections for an average utilization of 88.2%.



(a) 85.6% utilization (b) 88.2% utilization

Figure 6: Probability of Loss vs. Buffer Size

From Figure 6, we take note of the general shape of the loss probability as a function of delay bound, d, or buffer space, ld. As was the case of Figure 5, there is a fairly sharp knee to the curve: for small buffer sizes and delay bounds in the range of 10 to 20 msec, the loss probability falls off roughly exponentially with delay bound and has quite a fast decay rate. However, for larger delay bounds beyond 10 and 20 msec,

the loss probability still decreases exponentially, but with a much slower decay rate. For example, in Figure 5(b) for the curve labeled "simulation", the loss probability falls from roughly 10^{-2} to a little above 10^{-4} as the buffering or delay is increased from 1 msec to 10 msec: nearly two orders of magnitude decrease for a 9 msec increase in buffer size. Alternatively, as the buffering is increased to as much as 80 msec total, the loss probability has barely fallen by another order of magnitude.

We note that such behavior is not well represented by effective bandwidth techniques which model the loss probability by a single exponential $p \approx Ke^{-\eta d}$ (see [4] and [15] for examples). Indeed, several recent works indicate that effective bandwidth approximations may often be considerably inaccurate and severely under-utilize network resources [1], [14]. Alternatively, our resource allocation scheme based on application of the central limit theorem over intervals tracks this loss-behavior quite well. Indeed, as shown in Figure 5, the scheme is able to admit *most* of the allowable connections, and as shown in both Figures 5 and 6, it is able to track the complex behavior of the queue performance across a wide range of loads and QoS parameters.

5 Conclusion

In this paper, we introduced a new scheme for determining priority queue performance based on second moment characterizations of traffic streams, and building on the stochastic bounding techniques of [16] and [18]. We showed how simple computations on the variances of streams' arrivals over intervals can be used to determine QoS parameters.

Our experiments with MPEG traces and trace-driven simulations indicate that our scheme is accurate enough to capture most of the achievable statistical multiplexing gain over a wide range of utilizations, buffer sizes, and loss probabilities, with typical average utilizations in the range of 60% to 90%.

Moreover, our experiments indicate that QoS parameters such as the loss probability have a more complex form than that predicted by effective-bandwidth type approximations. The new variance-based resource allocation scheme is able to track this observed behavior.

Our results have implications for network design and planning as well as capacity allocation in connection admission control algorithms.

References

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

[2] H. Cramér. *Random Variables and Probability Distributions.* Cambridge University Press, 1937.

[3] R. Cruz. A calculus for network delay, part I : Network elements in isolation. *IEEE Transactions on Information Theory*, 37(1):114–121, January 1991.

[4] A. I. 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.

[5] A. Erramilli, O. Narayan, and W. Willinger. Experimental queueing analysis with long-range dependent packet traffic. *IEEE/ACM Transactions on Networking*, 4(2):209–223, April 1996.

[6] K. Fendik and W. Whitt. Measurements and approximations to describe the offered traffic and predict the average workload of a single-server queue. *Proceedings of the IEEE*, 77(1):171–194, January 1989.

[7] D. Le Gall. MPEG: A video compression standard for multimedia applications. *Communications of the ACM*, 34(4):46–58, April 1991.

[8] M. Garrett and W. Willinger. Analysis, modeling and generation of self-similar VBR video traffic. In *Proceedings of ACM SIGCOMM'94*, London, UK, August 1994.

[9] S. Golestani. Duration-limited statistical multiplexing of delay-sensitive traffic in packet networks. In *Proceedings of IEEE INFOCOM'91*, pages 323–332, Bal Harbour, FL, April 1991.

[10] S. Golestani. A framing strategy for congestion management. *IEEE Journal on Selected Areas in Communications*, 9(7):1064–1077, September 1991.

[11] R. Gusella. Characterizing the variability of arrival processes with indices of dispersion. *IEEE Journal on Selected Areas in Communications*, 9(2):203–211, February 1991.

[12] H. Heffes. A class of data traffic processes - covariance function characterization and related queueing results. *Bell System Technical Journal*, 59, 1980.

[13] C. Hwang and S. Li. On input state space reduction and buffer noneffective region. In *Proceedings of IEEE INFOCOM'94*, pages 1018–1028, June 1994.

[14] P. Jelenković and A. Lazar. On the dependence of the queue tail distribution on multiple time scales of ATM multiplexers. In *Proceedings of the 1995 Conference on Information Sciences and Systems*, Baltimore, MD, March 1995.

[15] 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.

[16] 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.

[17] E. Knightly and H. Zhang. Traffic characterization and switch utilization using deterministic bounding interval dependent traffic models. In *Proceedings of IEEE INFOCOM'95*, pages 1137–1145, Boston, MA, April 1995.

[18] 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.

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

[20] W. Leland, M. Taqqu, W. Willinger, and D. Wilson. On the self-similar nature of ethernet traffic. In *Proceedings of ACM SIGCOMM'93*, pages 183–193, San Francisco, CA, September 1993.

[21] S. Li. A general solution technique for discrete queueing analysis of multimedia traffic on ATM. *IEEE Transactions on Communications*, 39:1115–1132, July 1991.

[22] S. Li and C. Hwang. Queue response to input correlation functions: Continuous spectral analysis. *IEEE/ACM Transactions on Networking*, 1(6):678–692, December 1993.

[23] 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.

[24] J. Liebeherr, D. Wrege, and D.Ferrari. Exact admission control for networks with bounded delay services. *IEEE/ACM Transactions on Networking*, 1996. To appear.

[25] O. Rose. Statistical Properties of MPEG Video Traffic and Their Impact on Traffic Modeling in ATM Systems. Technical Report 101, Institute of Computer Science, Universiy of Wurzburg, Germany, February 1995.

[26] B. Ryu and A. Elwalid. The importance of long-range dependence of VBR video traffic in ATM traffic engineering: Myths and realities. In *Proceedings of ACM SIGCOMM'96*, August 1996.

[27] D. Wrege, E. Knightly, H. Zhang, and J. Liebeherr. Deterministic delay bounds for VBR video in packet-switching networks: Fundamental limits and practical tradeoffs. *IEEE/ACM Transactions on Networking*, 4(3). To appear in June 1996.

[28] H. Zhang and D. Ferrari. Rate-controlled static priority queueing. In *Proceedings of IEEE INFOCOM'93*, pages 227–236, San Francisco, CA, March 1993.