Providing End-to-End Statistical Performance Guarantees with Bounding Interval Dependent Stochastic Models

Hui Zhang
Lawrence Berkeley Laboratory
MailStop: 50B-229
Berkeley, CA 94720
hzhang@george.lbl.gov

Edward W. Knightly
University of California at Berkeley
and
Sandia National Laboratories
knightly@tenet.berkeley.edu

Abstract

This paper demonstrates a new, efficient, and general approach for providing end-to-end performance guarantees in integrated services networks. This is achieved by modeling a traffic source with a family of bounding interval-dependent (BIND) random variables and by using a rate-controlled service discipline inside the network. The traffic model stochastically bounds the number of bits sent over time intervals of different length. The model captures different source behavior over different time scales by making the bounding distribution an explicit function of the interval length. The service discipline, RCSP, has the priority queueing mechanisms necessary to provide performance guarantees in integrated services networks. In addition, RCSP provides the means for efficiently extending the results from a single switch to a network of arbitrary topology. These techniques are derived analytically and then demonstrated with numerical examples.

1 Introduction

High speed networking has introduced opportunities for new applications such as scientific visualization, network-based medical imaging and video conferencing. These applications have stringent performance requirements in terms of throughput, delay, delay-jitter and loss rate. The best-effort service provided by the current packet-switching networks is not adequate and services that support performance guarantees are needed.

In [5], two types of guaranteed services are proposed: *deterministic service* and *statistical service*. Deterministic service provides a guarantee that performance bounds are met for *all* packets on a connection even in the worst case. Though such a service is important for many applications such as interactive medical imaging applications and high quality video, deterministic service is expensive since resources must be reserved according to the worst case scenario. With deterministic service, the stochastic properties of traffic sources cannot be exploited to achieve a statistical multiplexing gain. Further, when the sources are bursty or require small queueing delays, providing deterministic service results in low average utilization of the network by the guaranteed performance traffic. Alternatively, many applications such as voice and video conferencing can tolerate certain losses of data without significantly affecting the quality. *Statistical service*, in which *probabilistic* or *statistical* performance bounds are guaranteed, can be used to support these applications. The advantage of providing statistical service is that the average utilization of the network by guaranteed performance traffic can be increased by exploiting statistical multiplexing.

There are two important aspects to the problem of providing guaranteed statistical services: developing probabilistic models to characterize traffic sources, and developing techniques to provide end-to-end probabilistic bounds in an internetwork environment.

[1] This research was supported by the National Science Foundation and the Defense Advanced Research Projects Agency (DARPA) under Cooperative Agreement NCR-8919038 with the Corporation for National Research Initiatives, by AT&T Bell Laboratories, Hitachi, Ltd., Hitachi America, Ltd., Pacific Bell, the University of California under a MICRO grant, and the International Computer Science Institute. The views and conclusions contained in this document are those of the authors, and should not be interpreted as representing official policies, either expressed or implied, of the U.S. Government or any of the sponsoring organizations.

In the literature, there have been many models proposed for video or audio traffic sources. Among the more popular ones are the on-off model for voice sources [3] and more sophisticated models based on Markov or other renewal processes for video sources [11]. A good survey for the probabilistic models for voice and video sources is presented in [12]. There are two important limitations to such traffic models. First, in an integrated services network, traffic sources are heterogeneous and will not in general conform to one model. Further, if a traffic source does not conform to the analytical model, no statistical guarantees can be made. Second, the above models cannot capture varying statistical properties of a source over time intervals of different length. It is therefore important to investigate traffic models that capture this *interval-length dependent* property of traffic sources.

In [10], Kurose proposed modeling a source with a family of random variables that stochastically bounds the number of bits sent over various interval lengths. In this paper, we extend this model so that the random variable that bounds the source over an interval of length t is an explicit function of t. This model can thus be used to characterize interval-length dependent behavior such as the observation that over longer intervals, the total number of bits sent by a source can be bounded by a random variable with expectation very near the source's long-term average rate, while on a shorter time scale, the source's bounding random variable must be weighted more towards the peak rate.

Having a traffic model for sources only solves part of the problem. In a networking environment, packets from different connections are multiplexed at each switch. Even if the traffic can be characterized at the entrance to the network, complex interactions among connections will destroy many properties of the traffic inside the network, and the traffic model at the source may not be applicable inside the network. Thus, even if performance guarantees may be made for a single switch, it may not be possible to make end-to-end performance guarantees.

In this paper, we address these two aspects of the problem by using bounding interval-dependent (BIND) stochastic traffic models to characterize traffic sources, and using rate-controlled service disciplines inside the network to reconstruct traffic patterns. By using interval-dependent stochastic traffic models and rate-controlled service disciplines, we can provide per-connection end-to-end statistical guarantees on throughput, delay, and delay-jitter in a network of *arbitrary* topology. The result is quite general. Unlike most existing solutions which work only in feed-forward and some restricted classes of feedback networks, our results hold in arbitrary networks. Also, unlike most existing solutions which assume a *constant* link delay between switches, we need only assume a *bounded* link delay. This is particularly important in internetworks where switches are connected by subnetworks. The delays of packets traversing subnetworks can be *bounded*, but may be *variable*.

The remainder of this paper is organized as follows. In Section 2, the bounding interval dependent (BIND) source model is developed and examples are presented for both discrete and continuous bounding distributions. Section 3 investigates multiplexing the discrete model of Section 2 for both homogeneous and heterogeneous sources. In addition, the switch utilization for various model parameters is investigated along with other trends involved with providing statistical performance guarantees to various sources. The Rate-Controlled Static-Priority (RCSP) scheduler is used to provide different performance guarantees to different connections. Finally, in Section 4, the results of Section 3 are extended to the network to provide end-to-end per-connection statistical performance guarantees.

2 BIND Model

The issue of modeling network traffic sources is gaining importance as networks evolve to provide integrated services and performance guarantees. Source modeling is a prerequisite to providing such services since admission control algorithms inherently require it.

2.1 Stochastic Processes vs. Bounding

The literature contains a wide variety of analysis techniques that model traffic sources by some stochastic process, calculate or approximate the aggregate process, and then solve for quantities in a switch such as the steady-state buffer distribution. Though these techniques provide valuable insights to a certain class of problems, such analysis tools are difficult to extend to integrated services networks since they often encounter problems such as the following:

- Model fitting - for a given stochastic model of traffic (as opposed to a bounding distribution) some sources will not fit the model. In such a case, no statistical guarantees can be made.
- Homogeneous traffic sources - this is not the case in integrated services networks.
- Aggregate results - need per-connection analysis to provide different services to different applications.
- Average results - statistical real-time service needs stronger guarantees than mean results.

- Limiting results - with connections continually being established and torn down, steady state results may not be reached quickly enough.
- Single switch analysis - results are often confined to a single switch and cannot be extended to the network because of the often intractable transformation of a switch on individual connections.
- No priorities - often, results only hold for single priority FCFS queueing. Again, B-ISDN will carry a wide range of traffic types, not only voice or only data.

Because of such difficulties with traditional stochastic models, a great deal of attention has been given to analyzing deterministic traffic models that provide some means of bounding a source's peak and average bandwidth over an averaging interval [4, 6, 7]. Such models are not only practical, but they also result in an analysis that does not suffer from many of the problems mentioned above. Specifically, these analyses can provide end-to-end *per-connection* performance bounds in networks with priority queueing service disciplines. One drawback to such models is that they cannot characterize many of the statistical properties of the source, and without additional assumptions, can only be used to provide deterministic performance bounds, not statistical performance bounds.

Recently, Kurose proposed a general framework for characterizing traffic sources [10]. Under such a framework, source j is characterized by a family of two tuples $\{(\mathbf{R}_{t_1,j}, t_1), (\mathbf{R}_{t_2,j}, t_2), (\mathbf{R}_{t_3,j}, t_3)...\}$, where $\mathbf{R}_{t_i,j}$ is a random variable that is *stochastically larger* than the number of bits generated over any interval of length t_i by source j. A random variable $\mathbf{X}$ is said to be stochastically larger than a random variable $\mathbf{Y}$ (denoted $\mathbf{X} \succeq_{st} \mathbf{Y}$) if and only if $Prob(\mathbf{X} > x) \geq Prob(\mathbf{Y} > x)$ for all x. Instead of modeling the exact arrival process of the source, Kurose's model stochastically *bounds* the number of transmitted bits in intervals of different length. In [16, 17], a stochastic extension to Cruz's deterministic model [4] is proposed that provides end-to-end stochastic bounds. However, this model does not take into account the interval dependent property considered in this paper. Moreover, these analyses investigate networks of work-conserving servers which, as discussed in Section 4, restricts the results to a certain class of networks.

In [6], a deterministic traffic model $(Xmin, Xave, I, Smax)$ is proposed, where $Xmin$ is the minimum packet interarrival time, $Xave$ is the average packet interarrival time over an averaging interval I, and $Smax$ is the maximum packet size. We propose the extension of the Tenet deterministic model to a probabilistic model within Kurose's framework. Further, we extend Kurose's model to make the bounding random variables explicit functions of the interval length in order to better characterize the properties of the source. As motivated above, there are two general requirements for the stochastic BIND model:

$$\mathbf{R}_t + \mathbf{R}_s \quad \succeq_{st} \quad \mathbf{R}_{t+s} \tag{1}$$

$$\frac{E(\mathbf{R}_t)}{t} \quad \leq \quad \frac{E(\mathbf{R}_s)}{s} \quad if \quad t > s \tag{2}$$

The first property is stochastic sub-additivity. The second property requires that the mean bounding rate over smaller time intervals is greater than the mean bounding rate over larger time intervals. In the next two subsections, we present two examples of stochastic bounding models that capture the interval-dependent property of a source. The example in 2.2 is a discrete model while the one in 2.3 is a continuous model. To keep the notation clearer, we will assume for the remainder of the paper that $Smax$ is fixed and equal to 400 bits and that λ_{pk} is the peak rate $(1/Xmin)$ and λ_{av} is the average rate $(1/Xave)$ both in packets or cells per second.

2.2 Discrete Example

In this section, we introduce a discrete-valued family of random variables with a parameterized binomial distribution to bound the number of packets that can be generated by a source in intervals of different length. Note that this is not to say that the underlying random process is binomial, rather that a binomial random variable is used to bound the process. By choosing different parameters for each of the family's random variables, it is possible to bound different processes with complicated distributions.

For the family of binomial bounding random variables, let the j^{th} source, denoted by S_j, be described by $\{(\mathbf{R}_{t,j}, t) \mid t \geq 0\}$, where $\mathbf{R}_{t,j}$ stochastically bounds the total number of packets that can arrive on connection j during any interval of length t. Dropping the source j subscript, an individual source is assumed to be bound by a binomial distribution with parameters M_t and p_t which are given by the following equations:

$$M_t = \lceil \lambda_{pk} t \rceil$$

$$p_t = \begin{cases} c((\lambda_{pk} - \lambda_{av})e^{-\gamma t/I} + \lambda_{av} - \lambda_{pk}e^{-\gamma}) & t \leq I \\ \lambda_{av}/\lambda_{pk} & t > I \end{cases}$$

where $c = 1/\lambda_{pk}(1 - e^{-\gamma})$ and $\gamma \geq 0$ is a client-specified parameter that controls how rapidly the mean bounding rate over an interval approaches the long-term average rate λ_{av} as the interval length gets larger. A larger γ means that the speed with which λ_{av} is approached is faster. This is illustrated in Figure 1 for $I = 133$ msec. The figure shows the mean bounding rate $(E(\mathbf{R}_t)/t = M_t p_t/t)$ vs. interval length t for various values of γ.

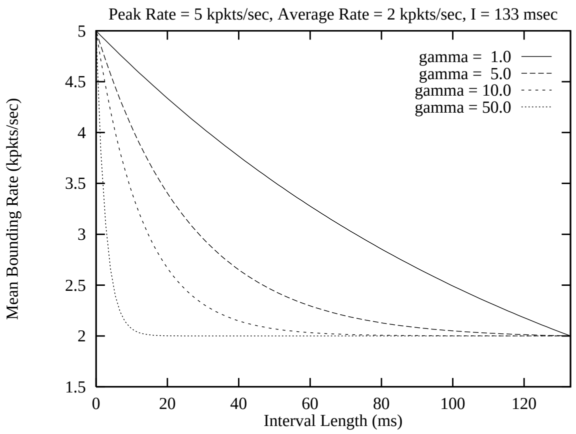


Figure 1: Effect of γ on mean bounding rate

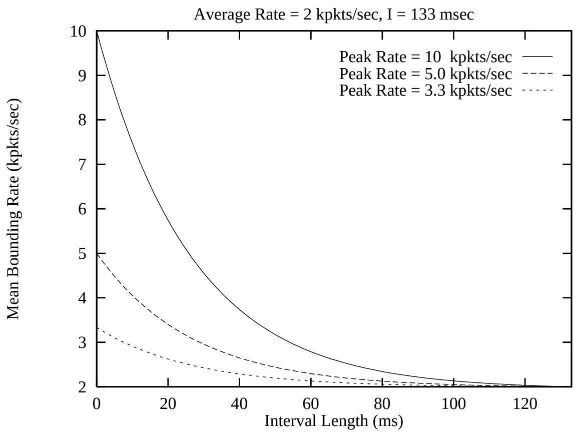


Figure 2: Effect of burstiness on mean bounding rate

This parameterization extends the Tenet deterministic model to a stochastic traffic model within Kurose's framework. Further, the stochastic representation above captures the interval-dependent behavior of a source such that the following properties hold:

1. The property in equation (2) is satisfied since the mean bounding rate over a longer interval is no more than the mean bounding rate over a shorter interval. As well, one can verify that equation (1) is satisfied.

2. The value of the source's mean bounding rate over any interval is greater than λ_{av}, and less than λ_{pk}, i.e., $\lambda_{av} \leq E(\mathbf{R}_t)/t \leq \lambda_{pk}$.

3. If two sources S_j and S_k have the same λ_{av}, I, and γ, but $\lambda_{pk,j} < \lambda_{pk,k}$, then

$$\frac{E(\mathbf{R}_{t,j})}{t} \leq \frac{E(\mathbf{R}_{t,k})}{t} \qquad \forall\ t > 0.$$

Property (3) states that if two connections have the same long-term bounding average rate λ_{av} (with interval length no less than I), the mean bounding rate over any interval is greater for the connection with the higher peak rate λ_{pk}. This property is illustrated in Figure 2. In the figure, the vertical axis is the mean bounding rate, and the horizontal axis is the length of the interval over which the average rate is computed. Curves for three sources are plotted. The three sources have the same I of 133 ms and λ_{av} of 2 kpkts/s, but have different λ_{pk}'s. As shown, a curve with a larger λ_{pk} is always above a curve with a smaller λ_{pk}. The same property holds in the deterministic model: with a fixed λ_{av}, a larger λ_{pk} means burstier traffic.

2.3 Continuous Example

While the previous section presented a discrete valued interval-length dependent traffic model, this section presents a *continuous* valued BIND model. As in the discrete case, this model takes into account the existence of different traffic characteristics over various interval lengths. Additionally, this model illustrates the fact that a wide range of bounding distributions can meet the requirements of equations (1) and (2). We demonstrate such a model with a family of random variables (again dropping the source j superscript) $\{(\mathbf{R}_t, t) \mid t \geq 0\}$, with $\mathbf{R}_t$ described by a family of fluid models where the rate of fluid flow is controlled by a two-state continuous time Markov chain with rate matrix

$$G_t = \begin{pmatrix} -\alpha_t & \alpha_t \\ \beta_t & -\beta_t \end{pmatrix}.$$

Note that α_t and β_t are functions of t indicating a dependence of each member of the family of Markov processes on its interval length. Letting τ be the time of the realization of the process, each Markov chain has a fixed rate matrix based on the fixed interval length t.

This property allows the model to capture the interval-dependent behavior of the underlying random process. Denoting the state of a chain at time τ by Σ_τ, $\Sigma_\tau \in \{1, 2\}$, fluid is generated at rate $r_{1,t}$ when $\Sigma_\tau = 1$, and at rate $r_{2,t}$ when $\Sigma_\tau = 2$ (again, $r_{1,t}$ and $r_{2,t}$ are constant for each Markov chain). The random variable $\mathbf{R}_t$ that bounds the source's rate over intervals of length t is then defined as the distribution of the total fluid content at time τ. Thus, for $t \geq 0$,

$$\mathbf{R}_t = \int_0^t \{r_{1,t} \cdot 1(\Sigma_\tau = 1) + r_{2,t} \cdot 1(\Sigma_\tau = 2)\}\, d\tau \qquad (3)$$

where $1(\cdot)$ is an indicator function.

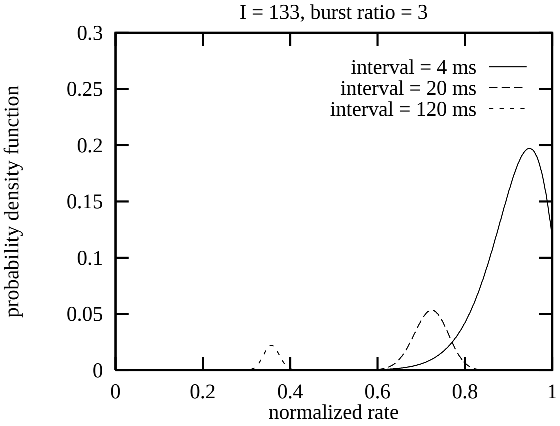


Figure 3: PDF's of rate bounding random variables

The two-state fluid model, a Markov-modulated fluid source (MMFS), is familiar to queueing theory. In [1] for example, the authors calculate the stationary buffer distribution for a number of independent fluid sources served by a link with capacity c. However, the Markov BIND model differs from such traditional models in the following ways:

- $\mathbf{R}_t$ is non-stationary since it is a non-decreasing measure of the fluid content at time t.
- This model uses a *family* of processes as a *bounding* random variable for different time intervals. Further, the random variables are a function of the *interval length.* This is quite different from a MMFS or even a superposition of MMFS's.

The distribution of $\mathbf{R}_t$ may be calculated by taking (without loss of generality) $r_{1,t} = 1$ and $r_{2,t} = 0$ and noting that the distribution of $\mathbf{R}_t$ in equation (3) is the distribution of the occupation time of a two-state Markov chain in state 1 until time t. The details of this derivation and the closed form distribution may be found in [14].

An example of a subset of a family of distributions is shown in Figure 3. The source has a peak rate of 2 Mbps, an average rate of 667 kbps, and $I = 133$ ms. In the figure, the horizontal axis is the rate normalized to the peak rate of 2 Mbps so that a normalized rate of 1 is the peak rate and a normalized rate of 0.333 is the average rate since in this case, the burst ratio $(\lambda_{pk}/\lambda_{av})$ is 3. The vertical axis is the probability density function (PDF) of the bounding random variable $\mathbf{R}_t/t$, where the $\mathbf{R}_t/t$ is the bounding rate over an interval of length t in bits or packets per second. The figure shows the PDF of the normalized rate for intervals of length 4, 20 and 120 ms. As shown, in a longer interval such as 120 ms, the source's rate is bounded by a random variable with expectation very near the source's long-term average rate (0.333 in the figure). Alternatively, on a shorter time scale such as 4 ms, the source's bounding random variable must be weighted more towards the peak rate (1.0 in the figure). The model parameters are $r_{1,t} = 2$ Mbps, $r_{2,t} = 0$, with α_t and β_t chosen so that as t approaches I, the Markov Chain averages 1/3 of its time in state 1 (corresponding to the burst ratio of 3). Different values for $r_{1,t}$, $r_{2,t}$, and the function mapping $\lambda_{pk}, \lambda_{av}, I$, and t onto α_t and β_t can result in a wide variety of density function shapes to tightly bound the traffic source of interest. Thus, these functions will result in bounding distributions of different variance that approach the mean bandwidth with different rates. Also, with $r_2 \neq 0$, the source will always send at least $r_2 \cdot t$ bits in an interval of length t resulting in a minimum rate for the source (and an appropriately weighted delta function at $r_2 \cdot t$ in the source's bounding PDF).

In summary, for both the discrete and continuous random variables, bounding distributions may be calculated that satisfy the properties of equations (1) and (2). These distributions capture the interval length dependent characteristics of traffic sources.

3 Multiplexing Stochastic BIND Sources

In this section, we analyze the multiplexing characteristics of connections specified by the interval-dependent traffic model for the Rate Controlled Static Priority (RCSP) scheduler [19]. The technique of extending a single switch analysis to a networking environment is discussed in Section 4. While Section 4 utilizes the Rate-Controlled aspect of the scheduler to provide end-to-end results, this section utilizes the Static Priority mechanism to analyze the statistical multiplexing behavior of the traffic model. The RCSP scheduler has the advantage of being both flexible so that it can offer a multiple number of delay bounds, and simple so that it can be implemented at very high speeds. We present numerical examples using the discrete traffic model developed in Section 2.2 and consider both a heterogeneous and homogeneous mix of sources.

The scheduler in an RCSP server consists of a number of prioritized real-time packet queues as shown in Figure 4. Packets at priority level 1 have the highest priority. A channel is assigned to a particular priority level at the channel's establishment time and all packets from the channel will be inserted into the real-time packet queue at that priority level. Multiple channels can be assigned to the same priority level. The scheduler services packets using a non-preemptive static priority policy which chooses packets in FCFS order from

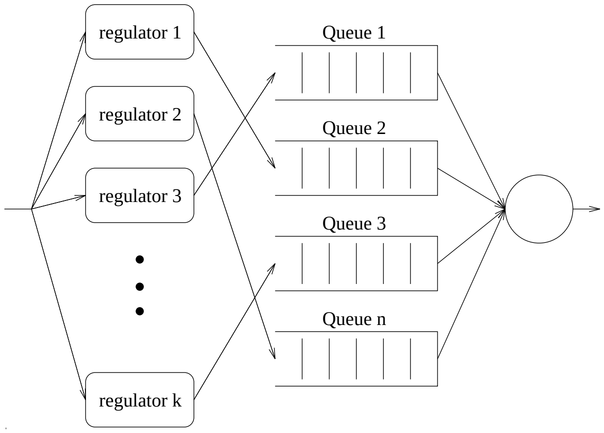


Figure 4: RCSP Scheduler

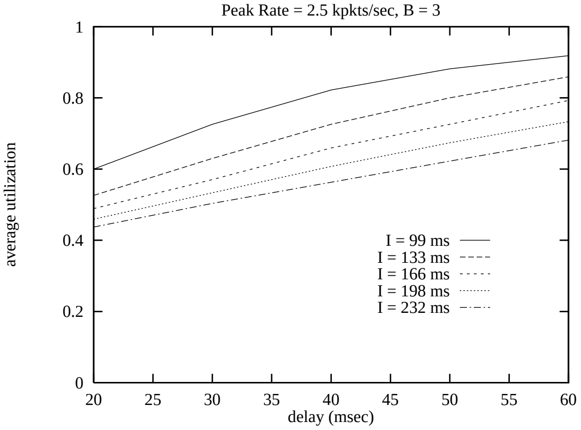


Figure 5: Effect of averaging interval

the highest-priority non-empty queue. Non-real-time packets are serviced from a separate, lower-priority, queue (not shown) only when there are no real-time packets queued. The service policy for non-real-time packets can be arbitrary.

There is a delay bound $\overline{d^m}$ associated with each priority level m. For a connection associated with priority level m, we are interested in calculating the delay-bound violation probability: $Prob\{d^m > \overline{d^m}\}$.

Proposition 1 *Let $\overline{d^1}, \overline{d^2}, \ldots, \overline{d^n}$ ($\overline{d^1} < \overline{d^2} < \cdots < \overline{d^n}$) be the respective delay bounds associated with each of the n priority levels in a Static Priority scheduler. Let C_q be the set of connections at level q and let the j^{th} connection among C_q have traffic specification*

$$\{(\mathbf{R}_{\overline{d^1},j,q}, \overline{d^1}), (\mathbf{R}_{\overline{d^2},j,q}, \overline{d^2}),\ \ldots\ (\mathbf{R}_{\overline{d^n},j,q}, \overline{d^n})\}.$$

With a link speed l and a maximum packet size of $\overline{Smax}$ for all connections,

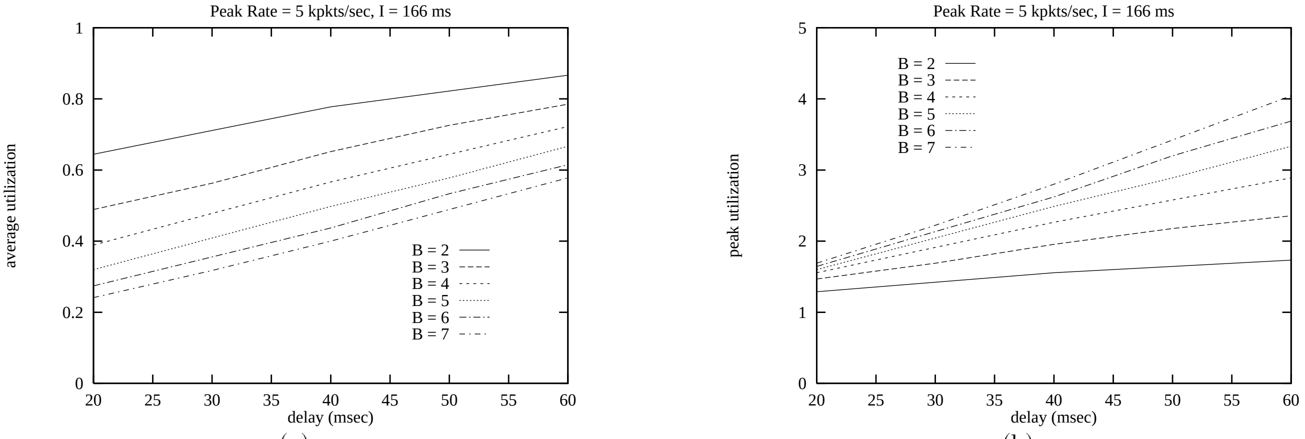
$$Prob\{d^m > \overline{d^m}\} \leq Prob\{\sum_{q=1}^{m} \sum_{j \in C_q} \mathbf{R}_{\overline{d^m},j,q} + \overline{Smax} \geq \overline{d^m} l\}$$

Intuitively, a packet meets its delay bound if all packets with higher priority (the number of which is the double sum of random variables) plus any packet already in service are served before the delay expires. The proposition shows that the tail distribution of the sum of the bounding random variables for all the connections with same or higher priorities can be used to provide an upper bound for the delay-bound violation probability. The result applies to bounding random variables with any distribution.

In the following sections, we present numerical examples using the discrete traffic model developed in Section 2.2. We consider the cases of both homogeneous and heterogeneous sources.

3.1 Homogeneous Sources

As in Section 2.2, we assume that $\mathbf{R}_{\overline{d^q},j,q}$ has a binomial distribution with parameters $M_{\overline{d^q},j,q}$ and $p_{\overline{d^q},j,q}$. For homogeneous sources, $M_{\overline{d^q}} = M_{\overline{d^q},j,q}$ and $p_{\overline{d^q}} = p_{\overline{d^q},j,q}$ for all j and q. With independence among the connections, $\displaystyle\sum_{q=1}^{m} \sum_{j \in C_q} \mathbf{R}_{\overline{d^m},j,q}$ has a binomial distribution with parameters $\sum_{q=1}^{m} J_q M_{\overline{d^q}}$ and $p_{\overline{d^q}}$, where J_q is the number of connections at priority level q, or $\mid C_q \mid$.



(a) (b)

Figure 6: Effect of burst ratio

The following examples show the relationships between the average utilization μ and the delay bound under various conditions. The link transmission rate is 45 Mbps, $\gamma = 6.0$, and the delay overflow probability is 0.1%. The parameters under consideration are I and the burst ratio $B = \lambda_{pk}/\lambda_{av}$. In all cases, the maximum number of connections are accepted such that the required performance guarantees of all connections are satisfied.

In Figure 5, the peak packet rate λ_{pk} and B are fixed and I is varied. As shown, a longer averaging interval I results in lower average utilization of the link. This confirms the intuition for the deterministic case (see [18, 20]). Since a source can send a maximum of $\lambda_{av}I$ consecutive packets at rate λ_{pk}, a larger I means a larger maximum burst length which results in lower average utilization of the link.

Figure 6 shows how B affects statistical multiplexing. Figure 6(a) shows that for a given performance requirement, smoother traffic is easier to multiplex and results in a higher average utilization of the link. Additionally, Figure 6(b) shows that although the average utilization is lower for bursty traffic, the peak utilization, a measure of the statistical multiplexing gain, is higher. This matches the intuition that higher burst ratios provide more opportunity for statistical multiplexing, but result in lower network utilization in order to meet a specific performance guarantee.

3.2 Heterogeneous Sources

This section presents an investigation into the effects of interactions among heterogeneous traffic sources. Since there is not a closed-form solution for the distribution of the sum of binomial random variables with different parameters, the resulting distribution may be calculated by convolving the individual probability distribution functions. This convolution can be efficiently implemented with the Fast Fourier Transform (FFT).

For simplicity, we consider different mixtures of two different types of sources. The algorithm, however, can calculate the case of arbitrary heterogeneous sources with no additional computational cost. A Class 1 source has a peak rate of 2 Mbps, an average rate of 1 Mbps, and $I = 198$ ms. A Class 2 source has a lower bandwidth, but has a greater burst ratio of 6, with a peak rate of 400 kbps and an average rate of 66.7 kbps. The delay overflow probability is 0.1%, and γ is 6.0.

Figure 7 shows the maximum numbers of Class 1 and Class 2 connections that can be accepted under different delay constraints. A point (n_1, n_2) on the curve means that if there are n_1 Class 1 connections traversing the link, at most n_2 Class 2 connections can be accepted over the same link. As in the case of homogeneous sources, more connections can be accepted when the delay bound is larger. The figure is similar in spirit to [8] where a schedulable region of admissible connection combinations is shown. However, for a general service in which traffic is not restricted to classes, these calculations would be made with the FFT rather than by table lookup.

Figure 8 shows the average and peak utilizations of the link under different mixtures of Class 1 and Class 2 connections. Each number n_1 on the horizontal axis in the figures represents a 2-tuple (n_1, n_2) as defined in Figure 7. Since the traffic of a Class 2 connection is burstier than that of a Class 1 connection, a larger number of Class 1 connections means that the mixture has a larger fraction of Class 1 traffic, and is therefore less bursty. Figure 8(a) shows similar characteristics to the case of homogeneous sources shown in Figure 6(a): less bursty traffic is easier to multiplex and re-

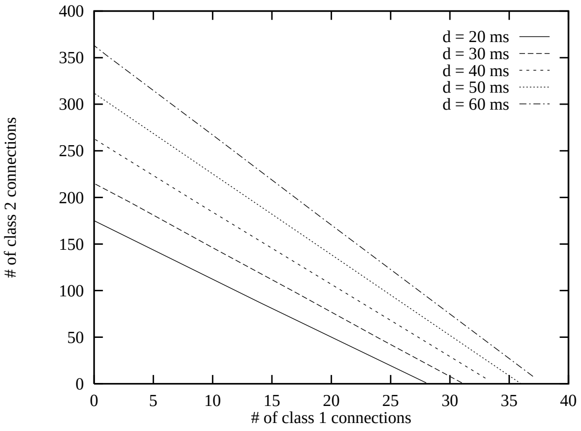


Figure 7: Maximum numbers of accepted connections

sults in a higher average utilization of the link. Figure 8(b) confirms the results obtained for homogeneous sources (in Figure 6(b)): although the average utilization is lower for burstier traffic, the peak utilization, a measure of the statistical multiplexing gain, is higher. Again, the intuition is that higher burst ratios provide more opportunity for statistical multiplexing but result in lower network utilization.

4 Providing End-to-End Statistical Performance Guarantees

In the previous section, we presented the conditions for bounding the delay overflow probability in a Static Priority scheduler and gave numerical examples to illustrate the results. In this section, we extend the analysis from a single scheduler to a network of switches.

In a networking environment, packets from different connections are multiplexed at each switch. Even if the traffic can be characterized at the entrance to the network, complex interactions among connections will distort the traffic pattern and destroy many properties of the traffic inside the network. Thus, the traffic model at the source may not be applicable inside the network.

One solution to this problem is to characterize the traffic pattern's distortion inside the network, and derive the traffic characterization at the entrance to each switch from characterizations of the source traffic and the traffic pattern distortion. This approach, taken in [4, 2, 13, 10], has several limitations.

First, it only applies to networks with *constant* delay links. Constant delay links have the desirable property that the traffic pattern at the receiving end of the link is the same as that at the transmitting end of the link. This property is important for these solutions because central to the analysis is the technique of characterizing the output traffic from a scheduler and using it as the input traffic to the next-hop scheduler. However, in an internetworking environment, links connecting switches may be subnetworks such as ATM or FDDI networks. Though it is possible to bound delay over these subnetworks, the delays for different packets will be *variable*. Thus, these solutions do not apply to an internetworking environment.

Second, most of the solutions characterize traffic in networks with *work-conserving* service disciplines.[1] Characterizing the traffic pattern inside the network is equivalent to solving a set of multi-variable equations [4, 13, 10]. In a feedback network, where traffic from different connections form traffic loops, the resulting set of equations may be unsolvable. Thus, most of these solutions apply only to feed-forward networks or a restricted class of feedback networks.

Finally, in networks with *work-conserving* service disciplines, even if the traffic inside the network can be characterized, the traffic characterization must be more bursty inside the network than at the entrance. For example, in [4], a deterministic fluid model (σ, ρ) is used to characterize traffic sources. A source is said to satisfy (σ, ρ) if during any time interval of length u, the amount of its output traffic is less than $\sigma + \rho u$. In such a model, σ is the maximum burst size, and ρ is the average rate. If the traffic of connection j is characterized by (σ_j, ρ_j) at the entrance to the network, its characterization will be

$$(\sigma_j + \sum_{i'=1}^{i-1} \rho_j \overline{d}_{i',j}, \rho_j) \tag{4}$$

at the entrance to the $i-th$ switch along the path, where $\overline{d}_{i',j}$ is the local delay for the connection at the $i'-th$ switch. Compared to the characterization of the source traffic, the maximum burst size in (4) increases by $\sum_{i'=1}^{i-1} \rho_j \overline{d}_{i',j}$. This increase of burst size grows linearly along the path.

In [10], a family of bounding random variables is used to characterize the source. In a work-conserving service discipline, if the traffic of connection j is characterized by $\{(\mathbf{R}_{t_1,j}, t_1), (\mathbf{R}_{t_2,j}, t_2), ...\}$ at the entrance to the network, its characterization will be

$$\{(\mathbf{R}_{t_1+\sum_{i'=1}^{i-1} b_{i',j}}, t_1), (\mathbf{R}_{t_2+\sum_{i'=1}^{i-1} b_{i',j}}, t_2), ...\}$$

at the i'_{th} switch, where $b_{i'}$ is the maximum busy period at switch i'. The same random variable that bounds the maximum number of packets over an interval at the entrance of

[1] In a work-conserving discipline, the link is never idle when there are packets waiting in the queue [21].

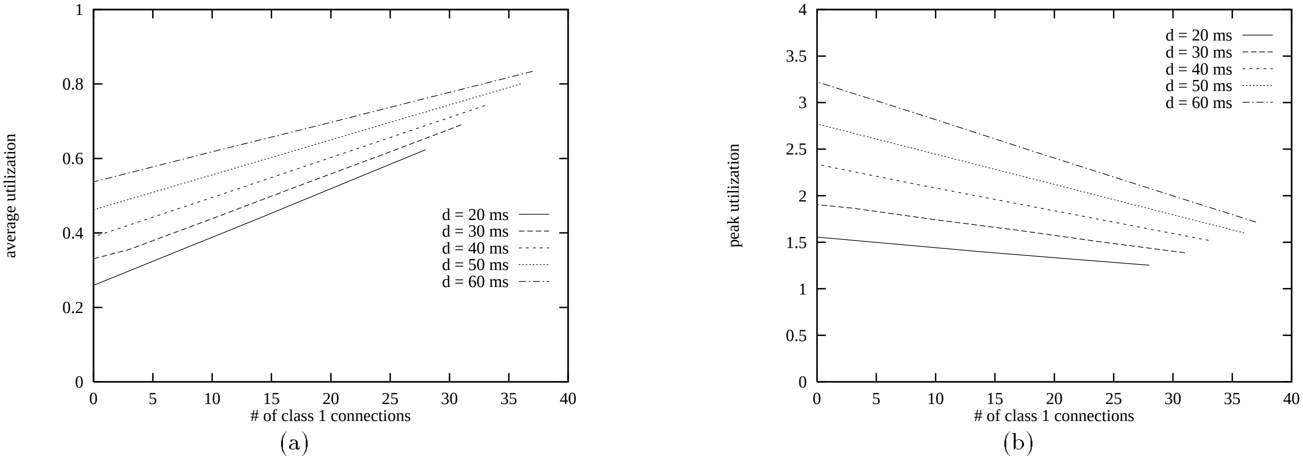


Figure 8: Average and peak utilizations with different mixtures

the network bounds the maximum number of packets over a much *smaller* interval at switch j. I.e., the traffic characterization is burstier at switch j than at the entrance.

Thus, in both the (σ, ρ) and $\{(\mathbf{R}_{t_1}, t_1), (\mathbf{R}_{t_2}, t_2), ...\}$ analysis, the burstiness of a connection's traffic accumulates at each hop along the path from the source to the destination. This will result in low utilization of the network by the real-time traffic.

Another solution to the traffic pattern distortion problem, which we adopt in our approach, is to reconstruct the traffic pattern at each switch with a class of non-work-conserving service disciplines called rate-controlled service disciplines [18]. As in Figure 4, a rate-controlled service discipline consists of two components, a rate-controller and a scheduler. The rate-controller shapes the input traffic from each connection into the desired traffic pattern by assigning an eligibility time to each packet. The scheduler then orders the transmission of eligible real-time packets from all connections. Many types of regulators and schedulers can be used. Different combinations of regulators and schedulers result in different service disciplines. The class is quite general. Most non-work-conserving disciplines proposed for high speed networks such as Delay-EDD [15], Stop-and-Go [7], Hierarchical Round Robin [9], and Rate-Controlled Static Priority [19], either belong to this class, or can be implemented by a rate-controlled service discipline with the appropriate choices of rate-controllers and schedulers [18].

Rate-controlled service disciplines have several important properties:

(1) If a connection's traffic satisfies certain traffic characteristics at the entrance to the network, with use of the appropriate rate-controllers, the same characteristics will be satisfied by the traffic at the entrance to each scheduler along the path. One type of rate-controller, a delay-jitter controlling regulator [19, 18], completely reconstructs the original traffic pattern at each switch. If a connection traverses a path of rate-controlled servers with delay-jitter controlling regulators, the traffic pattern at the entrance to each of the schedulers is *exactly the same* as that at the entrance to the network. This allows us to analyze each scheduler using the *same* traffic characterization.

(2) The end-to-end delay of a packet in a network with rate-controlled servers consists of the following components: waiting time in the schedulers, holding time in the rate-controllers, and the link delays. In [18], it is shown that the end-to-end delay can be bounded by the sum of bounds on link delays and bounds on waiting time in the schedulers; holding packets in rate-controllers will not increase the end-to-end delay bound, although it may increase the end-to-end average delay.

Properties (1) and (2) are significant. Property (1) means that we can analyze the delay characteristic of each scheduler along a path with the *same* traffic characteristics of the original source. The traffic characteristics need not be the ones discussed in this paper. For example, if a connection can be characterized by a MMPP at the entrance to the network, it can be characterized by the *same* MMPP at each of the schedulers. Property (2) means that we can combine the delay analysis of each individual scheduler and obtain the end-to-end delay characteristics of a connection. If we assume that any packet missing the local delay bound at a scheduler is dropped immediately, the end-to-end delay overflow probability Z, can be decomposed into local delay overflow probabilities z_i, where $Z = \prod_{i=1}^{n} z_i$ and n is the total number of switches along the path traversed by the connection. This is done in [6].

In summary, using rate-controlled service disciplines and analyzing delay characteristics using the same traffic characterization at each scheduler will achieve higher network utilization than using work-conserving disciplines and characterizing the traffic inside the network. This is due to the fact that in the latter case, a connection's traffic characterization must become more bursty with each hop. Additionally, using rate-controlled service disciplines allows us to obtain end-to-end performance bounds in much more general networking environments than previous solutions allow.

5 Conclusions and Future Work

In this paper, we have proposed and demonstrated the efficiency of a new mechanism for providing end-to-end probabilistic performance guarantees in an integrated services network. First, we present a new BIND traffic model, which stochastically bounds the number of bits sent over time intervals of different length and requires that these distributions are explicit functions of the interval length. Second, we analyze the multiplexing of such sources in a single switch served by an RCSP scheduler considering both homogeneous and heterogeneous sources. Finally, using delay-jitter control, we efficiently extend these results to the network to provide end-to-end per-connection statistical performance guarantees.

The focus of our current and future work is to characterize a wide variety of real-time traffic sources using both the deterministic and stochastic BIND model. With such characterizations, we will investigate various facets of multiplexing these sources. The goal is to achieve a high network utilization while providing mathematically provable statistical or deterministic performance guarantees. Other areas of the work will include compacting the model to its most significant parameters and comparing the model and its performance metrics to other models in the literature.

6 Acknowledgements

The authors are grateful to Domenico Ferrari and to all members of the Tenet Group for their ideas, comments, and criticisms.

References

[1] D. Anick, D. Mitra, and M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *Bell System Technical Journal*, 61:1971–1894, 1982.

[2] A. Banerjea and S. Keshav. Queueing delays in rate controlled networks. In *Proceedings of IEEE INFOCOM'93*, pages 547–556, San Francisco, CA, April 1993.

[3] P. Brady. A techniques for investigating on-off patterns in speech. *Bell System Technical Journal*, 44:1–22, January 1965.

[4] R. Cruz. A calculus for network delay, parts I and II. *IEEE Transaction of Information Theory*, 37(1):114–141, 1991.

[5] D. Ferrari. Client requirements for real-time communication services. *IEEE Communications Magazine*, 28(11):65–72, November 1990.

[6] 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.

[7] S. Golestani. A stop-and-go queueing framework for congestion management. In *Proceedings of ACM SIGCOMM'90*, pages 8–18, Philadelphia Pennsylvania, September 1990.

[8] J. Hyman, A. Lazar, and G. Pacifici. Real-time scheduling with quality of service constraints. *IEEE Journal on Selected Areas of Communications*, pages 1052–1063, September 1991.

[9] C. Kalmanek, H. Kanakia, and S. Keshav. Rate controlled servers for very high-speed networks. In *IEEE Global Telecommunications Conference*, pages 300.3.1 – 300.3.9, San Diego, California, December 1990.

[10] J. Kurose. On computing per-session performance bounds in high-speed multi-hop computer networks. In *ACM SigMetrics'92*, pages 128–139, Newport, Rhode Island, June 1992.

[11] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson, and J.D. Robbins. Performance models of statistical multiplexing in packet video communications. *IEEE Transaction on Communication*, 36(7):834–844, July 1988.

[12] I. Nikolaidis and I. Akyildiz. Source characterization and statistical multiplexing in ATM networks. Technical Report GIT-CC-92/24, College of Computing, Georgia Institute of Technology, Atlanta, GA 30332-0280, 1992.

[13] A. Parekh and R. Gallager. A generalized processor sharing approach to flow control in integrated services networks: The multiple node case. In *Proceedings of the INFOCOM'93*, pages 521–530, San Francisco, CA, March 1993.

[14] P. Pedler. Occupation times for two state Markov chains. *Journal of Applied Probability*, 8:381–390, 1971.

[15] D. Verma, H. Zhang, and D. Ferrari. Guaranteeing delay jitter bounds in packet switching networks. In *Proceedings of Tricomm'91*, pages 35–46, Chapel Hill, North Carolina, April 1991.

[16] O. Yaron and M. Sidi. Calculating performance bounds in communication networks. In *Proceedings of IEEE INFOCOM'93*, pages 539–546, San Francisco, CA, April 1993.

[17] O. Yaron and M. Sidi. Performance and stability of communication networks via robust exponential bounds. *IEEE Transaction on Networking*, 1(3):372–385, June 1993.

[18] H. Zhang. Service disciplines for integrated services packet-switching networks. PhD Dissertation. UCB/CSD-94-788, University of California at Berkeley, November 1993.

[19] H. Zhang and D. Ferrari. Rate-controlled static priority queueing. In *Proceedings of INFOCOM'93*, pages 227–236, San Francisco, California, March 1992.

[20] H. Zhang and D. Ferrari. Improving utilization for deterministic service in multimedia communication. In *1994 International Conference on Multimedia Computing and Systems*, Boston, MA, May 1994.

[21] H. Zhang and K. Srinivasan. Comparison of rate-based service disciplines. In *Proceedings of ACM SIGCOMM'91*, pages 113–122, Zurich, Switzerland, September 1991.