Telecommunication Systems 12 (1999) 167–191

Application of the many sources asymptotic and effective bandwidths to traffic engineering *

Costas Courcoubetis [a,b], Vasilios A. Siris [a] and George D. Stamoulis [a,b]

[a] *Institute of Computer Science (ICS)*
Foundation for Research and Technology - Hellas (FORTH)
P.O. Box 1385, GR 711 10 Heraklion, Crete, Greece
E-mail: {courcou,vsiris,gstamoul}@ics.forth.gr
[b] *Department of Computer Science, University of Crete, Heraklion, Greece*

Accurate yet simple methods for traffic engineering are important for efficient management of resources in broadband networks. The goal of this paper is to apply and evaluate large deviation techniques for traffic engineering. In particular, we employ the recently developed theory of *effective bandwidths*, where the effective bandwidth depends not only on the statistical characteristics of the traffic stream, but also on a link's operating point through two parameters, the *space* and *time* parameters, which can be computed using the *many sources asymptotic*. We show that this effective bandwidth definition can accurately quantify resource usage. Furthermore, we estimate and interpret values of the space and time parameters for various mixes of real traffic demonstrating how these values can be used to clarify the effects on the link performance of the time scales of traffic burstiness, of the link resources (capacity and buffer), and of traffic control mechanisms such as traffic shaping. Our experiments involve a large set of MPEG-1 compressed video and Internet Wide Area Network (WAN) traces, as well as modeled voice traffic.

Keywords: traffic engineering, large deviations, many sources asymptotic, effective bandwidths, time scales, broadband networks

1. Introduction

The rapid progress and successful penetration of broadband communications in the recent years has led to important new problems in traffic modeling and engineering. Among others, call acceptance control and network dimensioning for cases of guaranteed QoS have attracted the attention of researchers. Successful approaches are closely related to the ability of quantifying the usage of resources on the basis of traffic modeling and measurements.

* This work was supported in part by the European Commission under ACTS Project CASHMAN (AC-039). A subset of this paper has appeared in *Proceedings of ACM SIGMETRICS'98/PERFORMANCE'98*, June 1998. The software used for the experiments and other related material are available at URL: http://www.ics.forth.gr/netgroup/msa/

For example, statistical analysis of traffic measurements [15,18,10] has shown a self-similar or fractal behavior; such traffic exhibits long range dependence or slowly decaying autocorrelation. Although the implications of such long range dependence is still an open issue (e.g., see [9,11] and the references therein), recent work [20,11] has shown that these implications can be of secondary importance to the buffer overflow probability when the buffer size is small, which applies to the case where real time communication is supported. This example motivates the need for a methodology to understand the impact of the various *time scales* of the burstiness of real broadband traffic on the performance of the network and on its resource sharing capabilities. In particular, some basic questions for which the network engineer must provide answers are the following: How much does the cell loss probability decrease when the link capacity or buffer size increases? How does traffic shaping[1] affect the multiplexing capability of a link and the amount of resources used by a bursty source? What is the sufficient time granularity of traffic measurements in order to capture the information that is important for performance analysis and network dimensioning? What are the effects of the composition of traffic mix on the multiplexing capability of a link? Traditional queueing theory, which requires elaborate traffic models, cannot answer such questions in the context of large multi-service networks; for such cases asymptotic methods are more appropriate. In this paper we answer such questions by applying and evaluating the many sources asymptotic and the effective bandwidth definition based on this asymptotic for real broadband traffic. This traffic consists of MPEG-1 compressed video, Internet Wide Area Network (WAN) traffic, and traffic resulting from modeled voice.

Problems related to resource sharing have often been analyzed using the notion of *effective bandwidth*, which is a scalar that summarizes resource usage and which depends on the statistical properties and Quality of Service (QoS) requirements of a source. Effective bandwidths are usually derived by means of asymptotic analysis, which is concerned with how the buffer overflow probability decays as some quantity increases. If this quantity is the size of the buffer, we have the *large buffer asymptotic* [8,14]. If the buffer per source and capacity per source are kept constant, and we are interested in how the overflow probability decays as the size of the system (the broadband link and the multiplexed sources) increases, then we have the *many sources asymptotic*; this asymptotic regime has been investigated in [7,2,21].

Effective bandwidth definitions based on the large buffer asymptotic were found, in some cases, to be overly conservative or too optimistic [4]. This occurs because the large buffer asymptotic does not take into account the gain when many independent sources are statistically multiplexed together. Hence, in general the amount of resource usage depends not only on the statistical properties

[1] Related work on how traffic smoothing affects the multiplexing capability of a link employing a guaranteed and renegotiated constant bit rate service model can be found in [22].

and Quality of Service (QoS) requirements of a source, but also on the statistical properties of the other traffic it is multiplexed with and the resources (capacity and buffer) of the multiplexing link. Only recently [13,6] has it been understood how to incorporate such information into the definition of the effective bandwidth. These works have shown that the effective bandwidth of a source depends on the link's operating point through two parameters, the *space* and *time* parameters, which in turn depend on the link resources and the statistical properties of the multiplexed traffic. The space and time parameters can be computed using the many sources asymptotic and, as we will demonstrate with real broadband traffic, have important applications to traffic engineering. Furthermore, since the effective bandwidth gives the amount of resources that must be reserved for the source in order to satisfy its QoS requirements, it helps simplify problems in resource management and network dimensioning.

The rest of this paper is structured as follows. In Section 2 we review basic results from the theory of effective bandwidths, as developed in [13], and many sources asymptotic [7,2,21], and we discuss the application of this framework to traffic engineering, giving emphasis on the interpretation of the space and time parameters. In Section 3 we present a detailed series of experiments which aim to evaluate the accuracy of the above framework for link capacities and buffer sizes that will appear in broadband networks and for real broadband traffic which consists of MPEG-1 compressed video and Internet WAN traces, as well as modeled voice traffic. Finally, in Section 4 we summarize the results of the paper and identify areas for future research.

2. The many sources asymptotic and its implications

In this section we summarize the key results of the many sources asymptotic and the related effective bandwidth definition, and discuss their implications for traffic engineering.

2.1. Effective bandwidths and the many sources asymptotic

Suppose the arrival process at a broadband link is the superposition of independent sources of J types. Let $N_j = Nn_j$ be the number of sources of type j, and let $n = (n_1, \ldots, n_J)$ (the n_js are not necessarily integers). The broadband link has a shared buffer of size $B = Nb$ and link capacity $C = Nc$. Parameter N is the scaling parameter (size of the system), and parameters b, c are the buffer and capacity per source, respectively. We suppose that after taking into account all economic factors (such as demand and competition) the proportions of traffic of each of the J types remains close to that given by the vector n, and we seek to understand the *relative* usage of network resources that should be attributed to each traffic type.

Let $X_j[0,t]$ be the total load produced by a source of type j in the time interval $[0,t)$, which feeds the above link. We assume that $X_j[0,t]$ has stationary increments. Then, the *effective bandwidth* of a source of type j is defined as [13]

$$\alpha_j(s,t) = \frac{1}{st} \log E\left[e^{sX_j[0,t]}\right], \tag{1}$$

where s, t are system parameters which are defined by the context of the source, i.e., the characteristics of the multiplexed traffic, their QoS requirements, and the link resources (capacity and buffer). Specifically (these interpretations follow from equation (2) below), the *time* parameter t (measured in, e.g., milliseconds) corresponds to the most probable duration of the buffer busy period prior to overflow. The *space* parameter s (measured in, e.g., kb^{-1}) is an indication of the degree of multiplexing and depends, among others, on the size of the peak rates of the multiplexed sources relative to the link capacity. In particular, for link capacities much larger than the peak rates of the multiplexed sources, s tends to zero and $\alpha_j(s,t)$ approaches the mean rate of the source, while for link capacities not much larger than the peak rates of the sources, s is large and $\alpha_j(s,t)$ approaches the maximum value of the random variable $X_j[0,t]/t$.

Let $Q(Nc, Nb, Nn) = P(\text{overflow})$ be the probability that in an infinite buffer which multiplexes $Nn = (Nn_1, \ldots, Nn_J)$ sources and is served at rate $C = Nc$, the queue length is above the threshold $B = Nb$. The following result, established in [7], holds for $Q(Nc, Nb, Nn)$:

$$\lim_{N\to\infty} \frac{1}{N} \log Q(Nc, Nb, Nn) = \sup_t \inf_s \left[st \sum_{j=1}^{J} n_j \alpha_j(s,t) - s(b+ct) \right] = -I, \tag{2}$$

where I is called the asymptotic *rate function*. The last equation is referred to as the *many sources asymptotic*, and has been proved for continuous time in [2] and for a special case in [21]. A similar asymptotic holds for the proportion of workload lost through the overflow of a finite buffer of size Nb. Due to equation (2), the overflow probability can be written as $P(\text{overflow}) = e^{-NI+o(N)}$, which leads to the following approximation when N is large:

$$P(\text{overflow}) \approx e^{-NI}. \tag{3}$$

The accuracy of the above approximation and, more importantly, the achievable link utilization for real broadband traffic are investigated in Section 3.2.

Consider the QoS constraint on the overflow probability to be $P(\text{overflow}) \leq e^{-\gamma}$, and assume $\gamma = N\gamma_0$. Let $A(N\gamma_0, Nc, Nb)$ be a subset of $\mathbf{Z}_+^J$ such that $(Nn_1, \ldots, Nn_J) \in A(N\gamma_0, Nc, Nb)$ implies $\log P(\text{overflow}) \leq -N\gamma_0$ (and vice versa), i.e., the QoS constraint on the overflow probability is met. Due to this property, $A(N\gamma_0, Nc, Nb)$ is called the *acceptance region*. The region

$A(N\gamma_0, Nc, Nb)$ is hard to compute. However, for the scaled acceptance region the following holds [13]:

$$\lim_{N\to\infty} \frac{A(N\gamma_0, Nc, Nb)}{N} = A\,,$$

where

$$A = \bigcap_{0<t<\infty} A_t\,, \tag{4}$$

$$A_t = \left\{ (n_1, \ldots, n_J) : \inf_s \left[st \sum_{j=1}^{J} n_j \alpha_j(s,t) - s(b+ct) \right] \le -\gamma_0 \right\}.$$

Hence, the scaled acceptance region $A(N\gamma_0, Nc, Nb)/N$, for large N, can be approximated by A.

If $(n_1, \ldots, n_J)$ is on the boundary of the region A and the boundary is differentiable at that point, then the tangent plane determines the half-space [13]

$$\sum_{j=1}^{J} n_j \alpha_j(s,t) \le c + \frac{1}{t}\left(b - \frac{\gamma_0}{s}\right) = c^*\,, \tag{5}$$

where (s,t) is an extremizing pair in equation (2) and c^* is the “effective capacity” per source at the operating point (s,t). The case for two source types ($J = 2$) is shown in Figure 1.

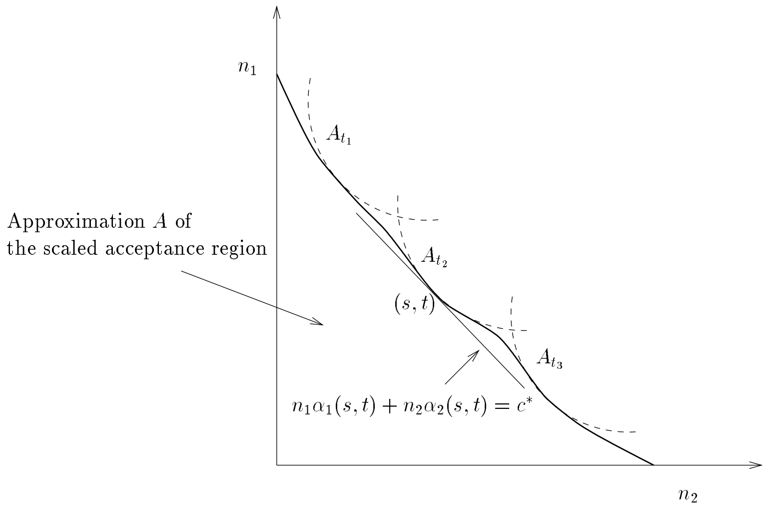


Figure 1. Approximation A of the scaled acceptance region $\frac{A(N\gamma_0, Nc, Nb)}{N}$ for three values of t.

To the extent that $A(N\gamma_0, Nc, Nb)$ can be approximated by NA, it follows from (5) that a point $(N_1, \ldots, N_J) = (Nn_1, \ldots, Nn_j) \in A(N\gamma_0, Nc, Nb)$ can be taken to satisfy

$$\sum_{j=1}^{J} N_j \alpha_j(s,t) \leq C + \frac{1}{t}\left(B - \frac{\gamma}{s}\right) = C^*, \tag{6}$$

where, as in (5), (s,t) is an extremizing pair in equation (2) and C^* is the "effective capacity" of the system at the operating point (s,t).

According to (6), the effective bandwidth $\alpha_j(s,t)$ provides a relative measure of resource usage *for a particular operating point* of the link, expressed through parameters s, t. For example, if a source of type j_1 has twice as much effective bandwidth as a source of type j_2, then, for this particular operating point of the link, one source of the first type can be substituted for two sources of the second type, while still satisfying the QoS constraint. The above measure of resource usage differs from the ordinary measure that is usually reported (i.e., the mean rate), which corresponds to $s = 0$. Note that the QoS guarantees are encoded in the effective bandwidth definition through the value of γ that appears on the right hand side of (6) and influences the form of the acceptance region.

Unlike the effective bandwidth definition (1) which takes into account the effects of statistical multiplexing, the effective bandwidth based on the large buffer asymptotic depends solely on the characteristics of the source and the QoS constraint. Specifically, [8,14] consider the QoS constraint $P(\text{overflow}) \leq e^{-\delta B}$, where B is the total buffer. In this case the effective bandwidth based on the large buffer asymptotic of a source of type j and the corresponding constraint is

$$\alpha_j(s) = \frac{1}{s} \lim_{t\to\infty} \frac{1}{t} \log E\left[e^{sX_j[0,t]}\right], \tag{7}$$
$$\sum_{j=1}^{J} N_j \alpha_j(\delta) \leq C.$$

Observe that (7) is a special case of (1) for $t \to \infty$. Indeed, the effective bandwidth formula (7) gives an accurate measure of resource usage when the link buffer is large, in which case the time parameter t (which is related to the time for buffer overflow) becomes large. However, for finite buffer sizes equation (7) can lead to significant underutilization or even overutilization of link capacity [4]. Section 3 includes experiments that compare the performance of the large buffer asymptotic with that of the many sources asymptotic.

2.2. The Bahadur-Rao improvement

In this section we discuss an improvement of (3), due to [16], that is based on the Bahadur-Rao theorem. Similar ideas were introduced as heuristics in [12,17]. Then we derive an effective bandwidth constraint similar to (6) that takes into account this improvement. An important result is that both the effective bandwidth formula (1) and the computation of parameters s, t remain the same; the latter uses the `supinf` formula (2).

Recently the authors of [16] extended the proof of the many sources asymptotic in [7] to show that as $N \to \infty$

$$P(\text{overflow}) = \frac{1}{\sqrt{2\pi N \sigma^2 s^2}} e^{-NI} \left(1 + O\left(\frac{1}{N}\right)\right), \tag{8}$$

where (s,t) is an extremizing pair of (2) and σ^2 is given by

$$\sigma^2 = \frac{\partial^2}{\partial s^2} \log E\left[e^{sX_j[0,t]}\right] = \frac{M''(s)}{M(s)} - (ct+b)^2,$$

where $M(s) = E\left[e^{sX_j[0,t]}\right]$ is the moment generating function of the traffic process. Based on (8), we have the following approximation:

$$P(\text{overflow}) \approx \frac{1}{\sqrt{2\pi N \sigma^2 s^2}} e^{-NI} = e^{-NI - \frac{1}{2}\log(2\pi N\sigma^2 s^2)}. \tag{9}$$

We will refer to the above equation as the many sources asymptotic approximation with the Bahadur-Rao improvement. The term $\frac{1}{2}\log(2\pi N\sigma^2 s^2)$ can be approximated by $\frac{1}{2}\log(4\pi NI)$ [17]. Hence, equation (9) does not require any additional computations compared to (3).

Next, we derive the effective bandwidth constraint similar to (6) applicable with the Bahadur-Rao improvement (9). If the number of sources of each type $Nn = (N_1, \ldots, N_J)$ is such that the overflow probability given by (9) is equal to the target overflow probability $e^{-\gamma}$, then we have

$$-NI - \frac{1}{2}\log(2\pi N\sigma^2 s^2) = -\gamma.$$

Substituting $\frac{1}{2}\log(2\pi N\sigma^2 s^2)$ with $\frac{1}{2}\log(4\pi NI)$ in this equation gives

$$-NI - \frac{1}{2}\log(4\pi NI) = -\gamma \Leftrightarrow NI = \gamma - \frac{1}{2}\log(4\pi NI).$$

By setting $NI = \gamma + \epsilon$ in the last equation and taking the expansion of the logarithm on the right-hand side, i.e., $\log(4\pi NI) = \log(4\pi(\gamma+\epsilon)) \approx \log 4(\pi\gamma) + \epsilon/\gamma$, we obtain

$$\gamma + \epsilon \approx \gamma - \frac{1}{2}\log(4\pi\gamma) - \frac{1}{2\gamma}\epsilon \Leftrightarrow \left(1 + \frac{1}{2\gamma}\right)\epsilon \approx -\frac{1}{2}\log(4\pi\gamma) \Leftrightarrow \epsilon \approx -\frac{\frac{1}{2}\log(4\pi\gamma)}{1 + \frac{1}{2\gamma}}.$$

Substituting the last equation in $NI = \gamma + \epsilon$ gives

$$NI \approx \gamma - \frac{\frac{1}{2}\log(4\pi\gamma)}{1 + \frac{1}{2\gamma}} = \gamma'.$$

Combining the last equation with (2) gives the following effective bandwidth constraint in the neighborhood of the extremizing pair (s,t) of (2)

$$\sum_{j=1}^{J} N_j \alpha_j(s,t) \leq C + \frac{1}{t}\left(B - \frac{\gamma'}{s}\right) = C^*_{\text{B-R}}\,. \tag{10}$$

It is important to note that the same formula for the effective bandwidth, given by equation (1), is used in both (6) and (10), with the parameters s, t computed using the same formula (2). The Bahadur-Rao improvement only affects (increases) the amount of effective capacity $C^*_{\text{B-R}} > C^*$.

2.3. Implications to traffic engineering

Next we discuss the interpretation of the space and time parameters s, t, and how they can be used for traffic engineering.

For any traffic stream, the effective bandwidth $\alpha_j(s,t)$ in (1) is a template that must be filled with the system operating point parameters s, t in order to provide the correct measure of effective usage by a source for that particular operating point. Although the value of this operating point also depends on this individual source, for a *large* system, due to heavy multiplexing, this dependence can be ignored. Such an engineering approach simplifies considerably the analysis because there is no circle in the definitions of the effective bandwidth and the operating point. Indeed, as we will see in Section 3.4, the values of s, t are, to a large extent, insensitive to small variations of the traffic mix. Furthermore, it has been observed that in networks serving a large number of sources, the traffic mix exhibits strong cyclic behavior. Hence, we can assign particular pairs (s,t) to periods of the day during which the traffic mix remains relatively constant. The values of s, t for a particular period of the day can be computed off-line from traffic measurements taken during that period using the `supinf` formula (2) and the effective bandwidth formula (1); this procedure is discussed in detail in Section 3.1. Alternatively, the parameters s, t can be estimated using their interpretation, which we discuss next (related experimental results are presented in Section 3.3.1).

Recall that the time parameter t corresponds to the most probable duration of the buffer busy period prior to overflow. We now argue that this parameter also identifies the time scales that are important for buffer overflow. Assume that a link is operating at a particular operating point, expressed through parameters s, t. In the effective bandwidth formula (1) the statistical properties of the source appear in $X_j[0,t]$, which is the amount of workload produced by the source in an interval of length t. If two sources have the same distribution of $X_j[0,t]$ for this value of t, then they will both have the same effective bandwidth. A case of practical interest where this result can be applied is traffic smoothing: To have an effect on the amount of resources used by a source, traffic smoothing must be performed on a time scale larger than t, since only then does it affect the distribution of $X_j[0,t]$. We investigate this with real broadband traffic in Section 3.4.3. Based on the above discussion, the time parameter t also indicates the time granularity that traffic measurements must have, since given a value for t it would be sufficient to measure traffic in intervals with length a few times smaller than this value. Traditionally, the time granularity of traffic measurements was chosen in a rather ad-hoc manner.

Next we discuss the interpretation of the parameter s and the product st. Let $\gamma = -\log P(\text{overflow})$. Combining this with (2) we have $\gamma = \sup_t \inf_s [s(b + ct) - st \sum_{j=1}^{J} n_j \alpha_j(s,t)]$. Taking the derivative of the last equation (see also [6]) we obtain

$$s = \frac{\partial \gamma}{\partial B} \quad \text{and} \quad st = \frac{\partial \gamma}{\partial C}\,. \tag{11}$$

Thus, the parameter s is equal to the rate at which the logarithm of the overflow probability decreases with the buffer size for fixed capacity. On the other hand, the product st is equal to the rate at which the logarithm of the overflow probability decreases with the link capacity for fixed buffer size.

3. Multiplexing experiments

In this section we apply and evaluate for real broadband traffic the performance analysis framework discussed in Section 2. The specific issues we address are the following:

- Procedure for numerically solving the `supinf` formula (2). (Section 3.1)
- Comparison of the overflow probability and link utilization using the many sources asymptotic and its Bahadur-Rao improvement to the actual cell loss probability and maximum utilization estimated using simulation. (Section 3.2)
- Comparison of the values of the space and time parameters s, t computed by theory to the values estimated using simulation. (Section 3.3)
- Estimation and interpretation of typical values of parameters s, t for real broadband traffic. (Section 3.3)
- Investigation of how the values of parameters s, t, and subsequently the effective bandwidth, depend on the traffic mix. (Section 3.4)

Our experiments involve real broadband traffic, namely MPEG-1 compressed video and Internet WAN traces, as well as modeled voice traffic. The MPEG-1 sequences, made available[2] by O. Rose [19], were encoded using the UC Berkeley MPEG-1 software encoder with frame pattern IBBPBBPBBPBB. Each sequence consists of 40,000 frames (approximately 30 minutes). For Internet WAN traffic we use the Bellcore Ethernet trace BC-Oct89Ext made available[3] by W. Leland

[2] Available at URL: ftp://ftp-info3.informatik.uni-wuerzburg.de/pub/MPEG/
[3] Available from The Internet Traffic Archive at URL: http://www.acm.org/sigcomm/ITA/

and D. Wilson [15]. The duration of the trace is 122797.83 seconds. For voice traffic we use an on-off Markov fluid model with peak rate 64 kbps and average time spent in the "on" and "off" states 352 msec and 650 msec, respectively [3]. Finally, we consider link capacities $34, 155,$ and 622 Mbps, and buffers with maximum queueing delay up to 50 msec for MPEG-1 traffic, and up to 150 msec for Internet traffic.

3.1. Numerical solution of the `supinf` *formula*

Next we give some details of how the `supinf` formula (2) can be numerically solved in an efficient manner.[4] We assume that the source statistics are available from measurements of aggregate load (e.g., number of cells) in fixed intervals (epochs) with duration τ. From these measurements the value of $X_j[0,t]$ can be computed for values of t that are integer multiples of τ.

The `supinf` formula (2) can be written as

$$-I = \sup_t \inf_s J(s,t)\,, \tag{12}$$

where

$$J(s,t) = \left[st \sum_{j=1}^{J} n_j \alpha_j(s,t) - s(b+ct) \right] .$$

The expectation in (1) can be approximated by the empirical average. Hence if T is the total duration of the trace, then

$$\alpha_j(s,t) = \frac{1}{st} \log \left[\frac{1}{T/t} \sum_{i=1}^{T/t} e^{sX_j[(i-1)t,it]} \right] . \tag{13}$$

Solution of equation (12) involves two optimization procedures: the first consists of finding, for a fixed value t, the minimum $J^*(t) = \min_s J(s,t)$ and $s = \mathrm{argmin}_s J(s,t)$, whereas the second consists of finding the maximum $-I = \max_t J^*(t)$ and $t = \mathrm{argmax}_t J^*(t)$.

The minimization $J^*(t) = \min_s J(s,t)$ can be numerically solved in an efficient manner by taking into account the fact that the logarithmic moment generating function $st\alpha_j(s,t) = \log E[e^{sX_j[0,t]}]$ is convex in s. Due to this, $J(s,t)$ is a unimodal function of s and the minimizer is unique. Hence, to find $J^*(t) = \min_s J(s,t)$ one can start from an initial "uncertainty" interval $[s_a, s_b]$ that contains the minimum (this interval can be found heuristically), and decrease the uncertainty interval using a *golden section* search. The procedure stops when the uncertainty interval has length less than some small value ϵ.

Unlike the previous minimization procedure, there is no general property for $J^*(t)$ that we can take advantage of in order to perform the maximization

[4] Software is available at URL: http://www.ics.forth.gr/netgroup/msa/

$-I = \max_t J^*(t)$ in an efficient manner.[5] For this reason, the maximization is solved by linearly searching the values of t in the interval $[0, \kappa\tau]$ with granularity equal to one epoch τ. The value of κ is determined empirically and depends on the buffer size: the extremizing value of t is larger for larger buffer sizes. Indeed, the experimental results in Section 3.3.1 show that the values of the time parameter t found using this procedure are in good agreement with the interpretation given by the theory, thus validating the correctness of the above procedure.

The run-time required for numerically solving the `supinf` formula (12) depends primarily on the size (number of epochs) of the trace file and the range of values of t that are linearly searched. On the other hand, it does not depend on the number of multiplexed streams. For example, when the trace files contain 40,000 epochs and 50 values of t are searched[6], the solution of the `supinf` formula requires approximately 23 seconds on an ULTRA-1 workstation with one UltraSPARC processor at 170 MHz.

3.2. Overflow probability and link utilization

In this section we compare the overflow probability and link utilization using the many sources asymptotic and its Bahadur-Rao improvement to the actual cell loss probability and maximum utilization estimated using simulation. We also derive a simple heuristic for computing the actual cell loss probability from the overflow probability.

3.2.1. Overflow probability

Figure 2 compares, for a fixed number of streams, the overflow probability estimated using the many sources asymptotic and its Bahadur-Rao improvement with the cell loss probability and frame overflow probability estimated using simulation; the latter is the probability that at least one cell of a frame is lost. Both the cell loss probability and the frame overflow probability are measured using a discrete time simulation model with an epoch equal to one frame time (= 40 msec). In these and subsequent simulations we report the average from a total of 80 independent simulation runs, each with a random selection of the starting frame for every stream. Each simulation run had duration five times the size of the trace. We assume that frame boundaries are aligned and for each stream the trace "wraps around" when the last frame is reached. Both the number of runs and the duration of each run were chosen empirically.

For each method, the decimal logarithm of the overflow probability is plotted against the buffer size (measured in milliseconds), while the link utilization remains constant.

[5] Furthermore, experimentation has shown that $J^*(t)$ can have more than one local minima.
[6] This range of t is typical for the case of MPEG-1 traffic with frame time 40 msec, when $C = 155$ Mbps and the maximum queueing delay in the buffer is less than 15 msec.

In Figure 2, first observe that for small buffer sizes there is a relatively fast decrease of the overflow probability as the buffer size increases. However, this occurs only for buffer sizes up to some value, e.g., $5-8$ msec for a 155 Mbps link; further increase of the buffer above this value has a small effect on the overflow probability. Furthermore, the logarithm of the overflow probability in both of these regimes is almost linear with the buffer size.

Second, observe that although the many sources asymptotic overestimates the Cell Loss Probability (CLP) by approximately 2-3 orders of magnitude, it qualitatively tracks its shape very well. Furthermore, the Bahadur-Rao improvement overestimates the CLP by 1-2 orders of magnitude. On the other hand, the large buffer asymptotic, in addition to overestimating the CLP by many orders of magnitude, also fails to track its shape.

The actual cell loss probability differs from the overflow probability estimated using the many sources asymptotic and its Bahadur-Rao improvement because the latter is a measure of the probability that in an infinite buffer the queue length becomes greater than B, rather than a measure of the CLP. The definition of the buffer overflow probability is closer in spirit to that of the frame overflow probability (the probability that at least one cell of a frame is lost). Indeed, as Figure 2 shows, the overflow probability estimated using the many sources asymptotic with the Bahadur-Rao improvement is very close to the frame overflow probability. This is the case because the simulation epoch is equal to the frame time.

To further explain the above, we derive a simple expression for the cell loss probability in terms of the frame overflow probability L_{f}. If one observes a large number of frames, say M, the average number of frames in which we have at least one lost cell is ML_{f}. Let x be the average number of cells that are lost when a frame overflow occurs. The average number of cells that are lost in M frames is $ML_{\mathrm{f}}x$ from a total of MF, where F is the average number of cells in a frame. We can approximate the cell loss probability with the percentage of lost cells, i.e.,

$$\mathrm{CLP} \approx \frac{ML_{\mathrm{f}}x}{MF} = \frac{x}{F}L_{\mathrm{f}}\,. \tag{14}$$

From the last equation we see that the cell loss probability differs from the frame overflow probability by a correction term $L_{\mathrm{c}} = x/F$. Lets assume that when an overflow occurs only a few cells are lost. This is reasonable to expect for small cell loss probabilities, since the probability of loosing cells in a buffer of size $B+\epsilon$ is exponentially smaller than loosing cells in a buffer of size B. In particular, we will assume that only one cell is lost, hence $L_{\mathrm{c}} \approx 1/F$, and since the average number of cells in one frame is 25 we get $L_{\mathrm{c}} \approx 1/25 = 10^{-1.4}$. This number agrees with the difference between the frame overflow probability and the cell loss probability shown in Figure 2. Indeed, Figure 3 shows the cell loss probability estimated using (14), where $L_{\mathrm{c}} = 10^{-1.4}$. Observe that the cell loss probability using the above heuristic matches the cell loss probability estimated using simulation extremely well.

3.2.2. Link utilization

Let $\varrho = Nm/C$ be the link utilization, where N is the number of streams, m is the mean rate, and C is the link capacity. Figure 4 compares, for a range of buffer sizes and for overflow probability 10^{-6}, the link utilization using the many sources asymptotic and its Bahadur-Rao improvement with the maximum achievable utilization (estimated using simulation). The utilization is computed by finding the largest number of multiplexed streams such that the overflow probability (3), computed using (12) and (13), is less than the target overflow probability 10^{-6}. This is done using a binary search for values of N in the interval $[N_{\min}, N_{\max}]$ with $N_{\min} = C/h$ and $N_{\max} = C/m$, where h is the peak rate of the streams. For the many sources asymptotic with the Bahadur-Rao improvement, (9) is used instead of (3).

Similar to our observations regarding the overflow probability, there are significant gains in increasing the size of the buffer up to a certain value. Increasing the buffer size above this value has a very small effect on link utilization.

Recall that the many sources asymptotic overestimated the CLP by 2-3 orders of magnitude. However, as Table 1 shows, it is more accurate in estimating the maximum utilization. In particular, for $C = 34$ Mbps and $B = 1$ msec the many sources asymptotic achieves a utilization that is approximately 79% of the maximum utilization. The Bahadur-Rao improvement increases this percentage to 88%. Furthermore, this percentage increases for larger link capacities; e.g., for $C = 155$ Mbps and $B = 1$ msec the many sources asymptotic achieves a utilization that is 90% of the maximum utilization (Table 1(b)). Of course, as Figure 5 shows, using the heuristic based on (14) we achieve a utilization that almost coincides with the maximum utilization.

Finally, Figure 6 shows the link utilization in the case of Internet WAN traffic. Observe that while for *Star Wars* traffic the gains of increasing the buffer decrease abruptly, for Internet WAN traffic the gains of increasing the buffer decrease more smoothly as the buffer size increases. This indicates that there are more time scales in Internet traffic which, if not smoothed, become important for buffer overflow, for different buffer sizes.

3.3. Space and time parameters

The space and time parameters s, t characterize a link's operating point and depend on the characteristics of the multiplexed traffic and the link resources. In this section we compare the values of these parameters computed using the `supinf` formula (12) to the corresponding values estimated using simulation. Furthermore, we compute and interpret typical values for these parameters, demonstrating how they can be used for traffic engineering.

Table 1
Link utilization for *Star Wars* traffic and target overflow probability 10^{-6}. The numbers in parentheses indicate the percentage of the maximum utilization (second column).

Buffer	Utilization ϱ		
msec (cells)	Simulation	Many sources	Many sources + B-R
1 (80)	0.57	0.46 (79 %)	0.52 (88 %)
8 (641)	0.70	0.59 (84 %)	0.64 (91 %)
16 (1282)	0.81	0.71 (88 %)	0.77 (96 %)

(a) $C = 34$ Mbps

Buffer	Utilization ϱ		
msec (cells)	Simulation	Many sources	Many sources + B-R
1 (365)	0.82	0.74 (90 %)	0.76 (94 %)
8 (2924)	0.92	0.88 (96 %)	0.89 (97 %)
16 (5849)	0.92	0.89 (97 %)	0.90 (98 %)

(b) $C = 155$ Mbps

3.3.1. Space and time parameters: theory vs. simulation

Recall from Section 2.3 that the space parameter s is equal to the rate at which the logarithm of the overflow probability decreases with the buffer size, equation (11). Motivated by this, we can estimate s using the ratio

$$s = \frac{\Delta\gamma}{\Delta B}, \tag{15}$$

with γ estimated from $-\log(\mathrm{CLP}_{\mathrm{sim}})$, where $\mathrm{CLP}_{\mathrm{sim}}$ is the cell loss probability estimated using simulation. Figure 7(a) shows that the values of parameter s computed using the `supinf` formula (12) are in good agreement with the values estimated using (15).

As discussed in Section 2.1, the time parameter t can be interpreted as the most probable duration of the buffer busy period prior to overflow. Figure 7(b) compares the value of parameter t computed using the `supinf` formula to the average value of the buffer busy period prior to overflow. As was the case for parameter s, the agreement between the two curves is good.

Note that the "steps" in the curves of s, t computed using the `supinf` formula are expected since the many sources asymptotic (and large deviations theory in general) captures only the most likely way overflow can occur. On the other hand, the curves labeled "simulation" in Figures 7(a) and 7(b) represent an average over all events that contribute to overflow.

Additional experimentation with other traffic types (Internet and videoconference traffic) has confirmed the above results.

3.3.2. Typical values and interpretation of the space and time parameters

Next we investigate how parameters s, t and the product st depend on the link capacity and buffer size. The values of s, t are computed using the `supinf` formula for a target overflow probability 10^{-7}.

Figure 8 shows the parameter s as a function of the buffer size, for various link capacities (Figure 8(a)) and video contents (Figure 8(b)). Observe that, initially, s decreases slowly with the buffer size. According to equation (11), s is equal to the rate at which the logarithm of the cell loss probability decreases as the buffer size increases. Hence, for larger buffers, where statistical multiplexing is more efficient, increasing the buffer has a smaller effect on the cell loss probability.

The explanation of the steep decrease of s in Figure 8(a) is similar to the explanation of the "knee" of the curves in Figures 2 and 4. Up to some value, the buffer's effect is to smooth the fast time scales of the multiplexed traffic. Thus, increasing the buffer has a large effect on the overflow probability, and the value of s is high. Once the fast time scales have been smoothed, the slow time scales govern the buffer overflow. Thus, increasing the buffer has a very small effect on the overflow probability, and the value of s is small. Also, observe in Figure 8(a) that the steep decrease of the value of s occurs for smaller buffer sizes (measured in milliseconds) as the link capacity increases. Finally, see Figure 8(b), similar behavior is observed for MPEG-1 traffic with various contents. This indicates that the above behavior of s is due to the MPEG-1 frame structure.

The dependence of parameter t on the buffer size is shown in Figure 9(a). Observe that the steep increases of t occur for the same buffer sizes for which s decreases (Figure 8(a)). Small values of t correspond to the regime where fast time scales are important for buffer overflow, whereas large values of t correspond to the regime where slow time scales are important for buffer overflow.

The product st as a function of the buffer size is shown in Figure 9(b). The initial slow decrease of st as the buffer increases occurs while t remains constant, and is due to the slow decrease of s (see Figure 8(a)). Furthermore, there is a steep increase of st, which occurs for the same buffer sizes for which the changes of s, t occur. The explanation for this steep increase of st is more subtle than the explanation for the behavior of s, t. Recall from Section 2.3 that st is equal to the rate at which the logarithm of the overflow probability decreases with the link capacity, for fixed buffer size, equation (11). Comparing Figures 9(a) and 9(b), we observe that the larger values of st correspond to larger values of t. Larger values of t result in an averaging effect in the amount of workload $X_j[0, t]$ that appears in the effective bandwidth formula (1). Hence, for the overflow phenomenon the traffic appears smoother. But for a link that multiplexes smooth traffic and is operating with a cell loss probability greater than zero, a change of the capacity has a greater effect on the overflow probability compared to a link multiplexing more bursty traffic. This gives the intuition of why st increases sharply for some buffer sizes.

Figure 10 compares the values of s, t for *Star Wars* and voice traffic. Figure 10(a) shows that as the buffer size increases, the value of s for voice traffic decreases smoothly. Furthermore, the rate of decrease is smaller for larger buffer sizes. Comparing the value of s for MPEG-1 and voice traffic, we conclude that, for buffer sizes up to 2 msec and above 10 msec, increasing the buffer has a larger effect for a network carrying voice traffic compared to a network carrying MPEG-1 traffic. This is an example of how the values of the space parameter can be used in buffer dimensioning.

Figure 10(b) shows that the time parameter t for voice traffic increases almost linearly with the buffer size. This can be explained since for a high degree of multiplexing, voice sources (which are modeled as on-off Markov fluids) behave as Gaussian sources. For such sources, it has been shown in [7] that the time parameter t increases linearly with the buffer size.

Figure 11(a) compares parameter s for *Star Wars* and Internet WAN traffic. For MPEG-1 traffic, the values of s form two distinct regimes corresponding to the two distinct time scales that are important for buffer overflow. On the other hand, for Internet traffic the values of s form more than two regimes, indicating that for such traffic there are more time scales which, for different buffer sizes, become important for buffer overflow. Recall that this is also the reason for the smoother dependence of the link utilization on the buffer size for Internet traffic compared to *Star Wars* traffic (Figure 6). Finally, Figure 11(b) shows that s can have different values for different Internet traffic segments from the same source, illustrating that different such segments have different statistical properties.

3.4. Effects of the traffic mix

As discussed in Section 2.1, periods of the day during which the traffic mix remains relatively constant can be characterized by particular pairs (s, t). In this section we investigate the dependence of these parameters, hence of the effective bandwidth, on the traffic mix. The traffic mix we consider consists either of traffic of different type (MPEG-1 video and voice), or of traffic with the same structure but different content (MPEG-1 video with different content), or of smoothed and unsmoothed traffic of the same type and content.

3.4.1. Traffic mix containing Star Wars and voice traffic

We first consider the traffic mix containing *Star Wars* and voice traffic. The horizontal axis in Figures 12(a) and 12(b) depicts the percentage of voice connections, each containing 30 individual voice channels. The vertical axis depicts the effective bandwidth of the *Star Wars* traffic stream. Observe that (1) the effective bandwidth, to a large extent, changes slowly with the traffic mix, (2) the dependence of the effective bandwidth on the traffic mix is smaller for larger capacities and buffer sizes, and (3) there are cases where increasing the percentage of voice connections leads to a sharp decrease of the effective bandwidth.

The first observation supports the argument that particular pairs (s, t) can be assigned to periods of the day during which the traffic mix remains relatively constant. However, the third observation states that there are certain percentages of the traffic mix where the effective bandwidth exhibits sharp changes. If the link's operating point is close to such a percentage, then we can estimate the average amount of resources used by a stream as a weighted sum of the effective bandwidth to the left and to the right of the sharp change. The weights would be determined by the percentage of the time the network was operating on the left and on the right of the change.

The sharp decrease of the effective bandwidth identified above is due to the change of the time scales that are important for buffer overflow. In particular, as indicated in Figure 12(a) above the curve for $C = 155$ Mbps and buffer 4 msec, the time parameter t increases (1, 4, and 7 frames) for the same percentage of voice connection at which the sharp decrease of the value of the effective bandwidth occurs. The increase of t produces an averaging effect (also discussed in Section 3.3.2) in the amount of workload $X_j[0, t]$ that appears in the effective bandwidth formula (1); this averaging results in a smaller effective bandwidth.

3.4.2. Traffic mix containing MPEG-1 traffic with different content

Our previous investigations addressed the case where the traffic mix consists of traffic with different structure. Now we investigate the case where the traffic mix consists of MPEG-1 video traffic with the same encoding parameters but with different content. Figures 13(a) and 13(b) show the effective bandwidth of the *Star Wars* stream as a function of the percentage of news and talk show streams, respectively. These figures show that the content has a very small effect on the effective bandwidth; this implies that the effects of the content on parameters s, t are also very small.

3.4.3. Traffic mix containing smoothed and unsmoothed Star Wars traffic

Our final investigation deals with another important question in traffic engineering: How does traffic smoothing affect the multiplexing capability of a link and the amount of resources used by a traffic stream? We will see that parameter t indicates the minimum time scale at which smoothing must be performed in order for it to affect resource usage.

Figure 14 shows the effective bandwidth of the *Star Wars* stream for different percentages of a traffic mix of unsmoothed and smoothed *Star Wars* traffic; the latter is created by evenly spacing the cells belonging to *two* consecutive frames. Observe that (1) the effects of the traffic mix on the effective bandwidth decrease when the link capacity or buffer size increases, (2) there are cases where increasing the buffer size has a very small effect on the effective bandwidth, e.g., at $C = 622$ Mbps the curves for $B = 8$ msec and $B = 16$ msec practically coincide, and (3) for some buffer sizes smoothing affects neither the effective bandwidth, nor the link's operating point, e.g., in Figure 14(a) the curve for $C = 155$ Mbps and $B = 8$ msec, and the curves for $C = 622$ Mbps and $B = 4, 8$, and 16 msec are *flat*. Next we discuss the third observation in more detail.

Figure 15 shows the effective bandwidth for both the smoothed and unsmoothed *Star Wars* stream. When the percentage of smoothed traffic is small, the time parameter t ($= 40$ msec) is smaller than the time interval over which smoothing was performed (80 msec). For this reason, the amount of workload $X_j[0,t]$ that appears in the effective bandwidth formula (1) is smaller for the smoothed stream than it is for the unsmoothed stream. Hence, the effective bandwidth of the smoothed stream is smaller than the effective bandwidth of the unsmoothed stream. For some percentage of smoothed traffic ($\approx 60\%$), the time parameter t ($= 80$ msec) is no longer smaller than the time interval over which smoothing is performed (80 msec). Because of this, the amount of workload $X_j[0,t]$ is the same for both the smoothed and the unsmoothed streams. Hence, the effective bandwidth of both streams is the same.

4. Conclusions

In this paper we employ the recently developed theory of effective bandwidths based on the many sources asymptotic, whereby the effective bandwidth depends not only on the statistical characteristics of the traffic stream but also on a link's operating point. The latter is summarized in two parameters: the *space* and *time* parameters.

We have investigated the accuracy of the above framework, and how it can provide important insight to the complex phenomena that occur at a broadband link with a high degree of multiplexing. In particular, we have estimated and interpreted values of the space and time parameters for various mixes of real traffic demonstrating how these can be used to clarify the effects on the link performance of the time scales of traffic burstiness, of the link resources (capacity and buffer), and of traffic control mechanisms such as traffic smoothing.

Our approach is based on the off-line analysis of traffic measurements, the granularity of which can be determined by the time parameter of the link. For the traffic mixes considered, the space and time parameters are, to a large extent, insensitive to small variations of the traffic mix. Furthermore, the dependence decreases for larger link capacities and buffer sizes. This indicates that particular pairs of these parameters can characterize periods of the day during which the traffic mix remains relatively constant. This result has important implications to charging, since simple pricing schemes that are linear in time and volume and have important incentive properties can be created from tangents to bounds of the effective bandwidth [6]. Furthermore, the above result opens up some new possibilities for traffic modeling. Rather than developing general models that try to emulate real traffic in any operating environment, a new approach would be to develop models, parameterized by s, t, that emulate real traffic for the particular operating point s, t. Such an approach is taken in [5]. If simple and efficient to implement, such models can be the basis for fast and flexible traffic generators.

The application of our approach to traffic engineering and management of traffic contracts in a real multi-service network that involves a large number of different source types is an important area for further research. Specific issues are whether the number of discontinuities of the operating point parameters s, t increases with the number of source types and how the parameters s, t vary for different periods of the day. Another important research topic is the analysis of multiplexers supporting multiple priorities [13,1]. It is interesting to extend our investigations to this case.

Acknowledgements

The authors are particularly grateful to Frank P. Kelly for his helpful discussions and insights, and thank the anonymous reviewers for their constructive comments.

References

[1] A. W. Berger and W. Whitt. Extending the effective bandwidth concept to networks with priority classes. *IEEE Commun. Mag.*, pages 78–83, August 1998.

[2] D. D. Botvich and N. G. Duffield. Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. *Queueing Systems*, 20:293–320, 1995.

[3] P. T. Brady. A model for generating ON-OFF speech patterns in two-way conversations. *Bell Syst. Tech. J.*, 48, September 1969.

[4] G. L. Choudhury, D. M. Lucantoni, and W. Whitt. On the effectiveness of effective bandwidths for admission control in ATM networks. In *Proc. of the 14th International Teletraffic Congress (ITC - 14)*, pages 411–420, North Holland, 1994. Elsevier Science B. V.

[5] C. Courcoubetis, A. Dimakis, and G. D. Stamoulis. Traffic equivalence and substitution in a multiplexer. To appear in *Proc. of IEEE INFOCOM'99*, 1999.

[6] C. Courcoubetis, F. P. Kelly, and R. Weber. Measurement-based usage charges in communications networks. Technical Report 1997-19, Statistical Laboratory, University of Cambridge, 1997. To appear in Operations Research.

[7] C. Courcoubetis and R. Weber. Buffer overflow asymptotics for a switch handling many traffic sources. *J. Appl. Prob.*, 33:886–903, 1996.

[8] A. I. Elwalid and D. Mitra. Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. *IEEE/ACM Trans. on Networking*, 1(3):329–343, October 1993.

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

[10] M. W. Garrett and W. Willinger. Analysis, modeling, and generation of self-similar VBR video traffic. In *Proc. of ACM SIGCOMM'94*, pages 269–280, August 1994.

[11] M. Grossglauser and J-C. Bolot. On the relevance of long-range dependence in network traffic. In *Proc. of ACM SIGCOMM'96*, pages 15–24, August 1996.

[12] I. Hsu and J. Walrand. Admission control for ATM networks. In *IMA Workshop on Stochastic Networks*. Springler-Verlag, 1994.

[13] F. P. Kelly. Notes on effective bandwidths. In F. P. Kelly, S. Zachary, and I. Zeidins, editors, *Stochastic Networks: Theory and Applications*, pages 141–168. Oxford University Press, 1996.

[14] G. Kesidis, J. Walrand, and C.-S. Chang. Effective bandwidths for multiclass Markov fluids and other ATM sources. *IEEE/ACM Trans. on Networking*, 1(3):424–428, August 1993.

[15] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of ethernet traffic. In *Proc. of ACM SIGCOMM'93*, pages 183–193, September 1993.

[16] N. Likhanov and R. R. Mazumdar. Cell loss asymptotics for buffers fed with a large number of independent stationary sources. In *Proc. of IEEE INFOCOM'98*, April 1998.

[17] M. Montgomery and G. de Veciana. On the relevance of time scales in performance oriented traffic characterizations. In *Proc. of IEEE INFOCOM'96*, pages 513–520, April 1996.

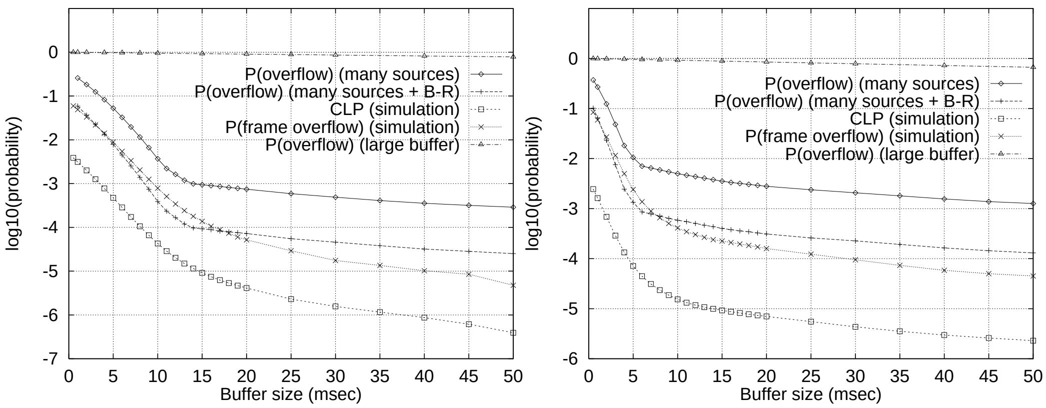
[18] V. Paxson and S. Floyd. Wide-area traffic: The failure of Poisson modeling. *IEEE/ACM Trans. on Networking*, 3(3):226–244, June 1995.

[19] O. Rose. Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems. Technical Report 101, University of Wuerzburg, February 1995.

[20] B. K. Ryu and A. Elwalid. The importance of the long-range dependence of VBR video traffic in ATM traffic engineering: Myths and realities. In *Proc. of ACM SIGCOMM'96*, pages 3–14, August 1996.

[21] A. Simonian and J. Guibert. Large deviations approximations for fluid queues fed by a large number of on/off sources. *IEEE J. Select. Areas Commun.*, 13(7):1017–1027, August 1995.

[22] Z. Zhang, J. Kurose, J. Salehi, and D. Towsley. Smoothing, statistical multiplexing and call admission control for stored video. *IEEE J. Select. Areas Commun.*, 15(6):1148–1166, August 1997.



(a) $C = 34$ Mbps, $\varrho = 0.83$ (b) $C = 155$ Mbps, $\varrho = 0.93$

Figure 2. **Overflow probability: theory vs. simulation for Star Wars traffic.** The many sources asymptotic tracks the shape of the cell loss probability very well. However, it overestimates it by 2-3 orders of magnitude. The Bahadur-Rao improvement overestimates the CLP by 1-2 orders of magnitude. On the other hand, the large buffer asymptotic, in addition to overestimating the CLP by many orders of magnitude, also fails to track its shape.

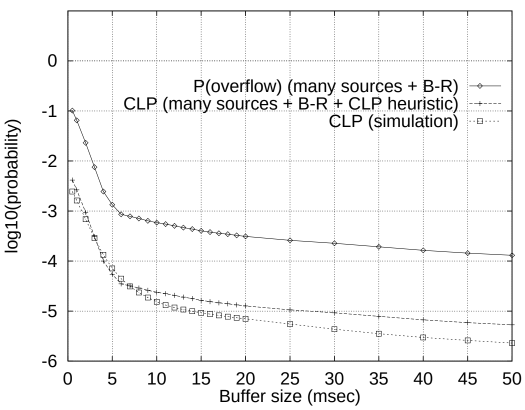
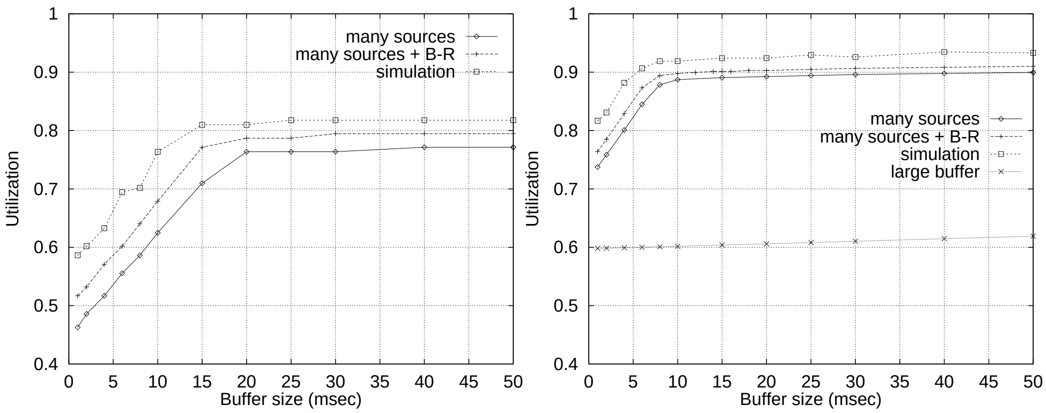


Figure 3. **Overflow probability using the many sources asymptotic with the Bahadur-Rao improvement and CLP heuristic (Star Wars traffic).** The many sources asymptotic with the Bahadur-Rao improvement and CLP heuristic matches the CLP estimated using simulation extremely well. [$C = 155$ Mbps, $\varrho = 0.93$]



(a) $C = 34$ Mbps (b) $C = 155$ Mbps

Figure 4. **Link utilization: theory vs. simulation for Star Wars traffic.** The many sources asymptotic with the Bahadur-Rao improvement performs better in utilizing a link than it does in estimating the cell loss probability (Figure 2). On the other hand, the large buffer asymptotic achieves a very low link utilization. [$P(\text{overflow}) \leq 10^{-6}$]

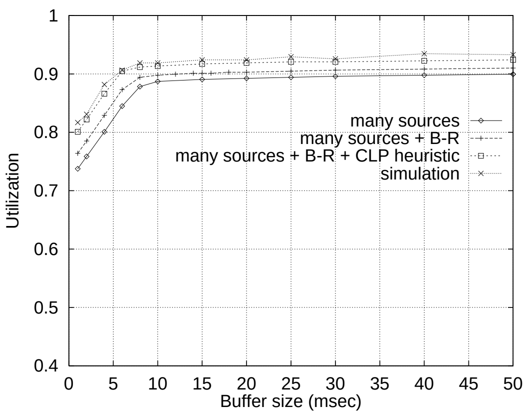


Figure 5. **Link utilization using the many sources asymptotic with the Bahadur-Rao improvement and CLP heuristic (Star Wars traffic).** The many sources asymptotic with the Bahadur-Rao improvement and CLP heuristic achieves practically the same utilization as the maximum utilization (estimated using simulation). [$C = 155$ Mbps]

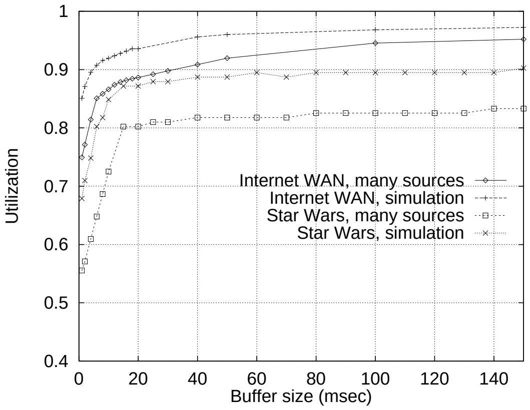
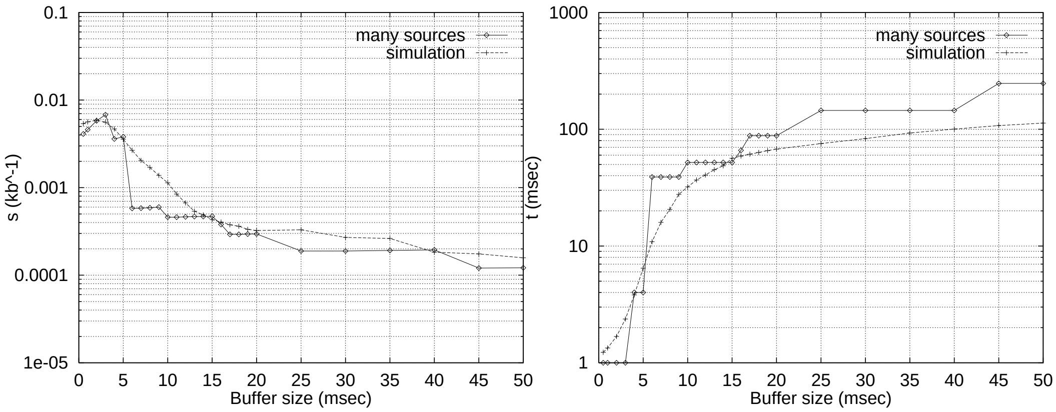
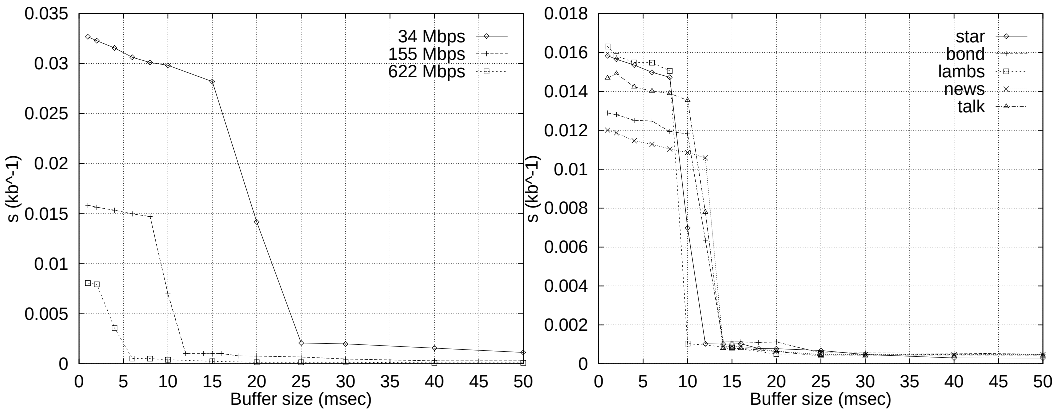


Figure 6. **Link utilization for Internet WAN and Star Wars traffic.** As was the case for MPEG-1 traffic, for Internet WAN traffic the many sources asymptotic achieves a high link utilization. However, while for *Star Wars* traffic the gains of increasing the buffer decrease abruptly, for Internet WAN traffic the gains of increasing the buffer decrease smoother as the buffer size increases. This indicates that there are more time scales in Internet traffic which, for different buffer sizes, become important for buffer overflow. [$C = 34$ Mbps, $P(\text{overflow}) \leq 10^{-4}$]



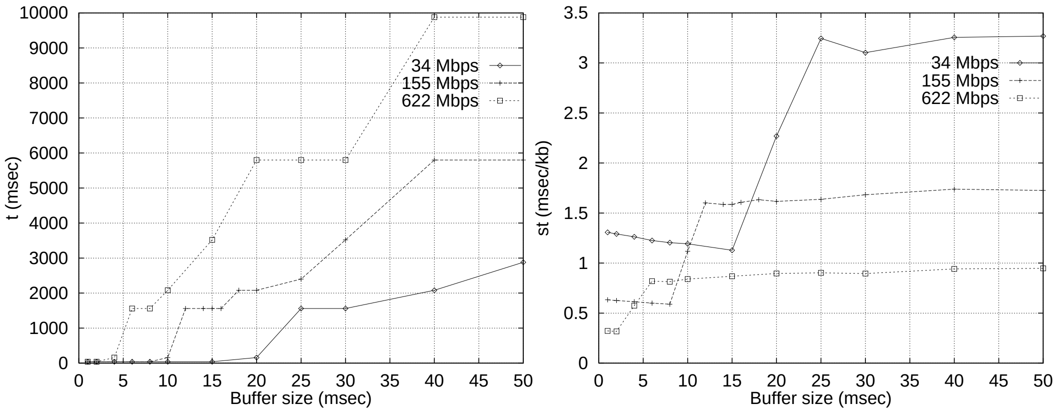
(a) Parameter s (b) Parameter t

Figure 7. **Parameters s, t: theory vs. simulation for Star Wars traffic.** The values of parameters s, t computed by the many sources asymptotic using the `supinf` formula (2) are in good agreement with the values estimated using simulation. [$C = 155$ Mbps, $\varrho = 0.93$]



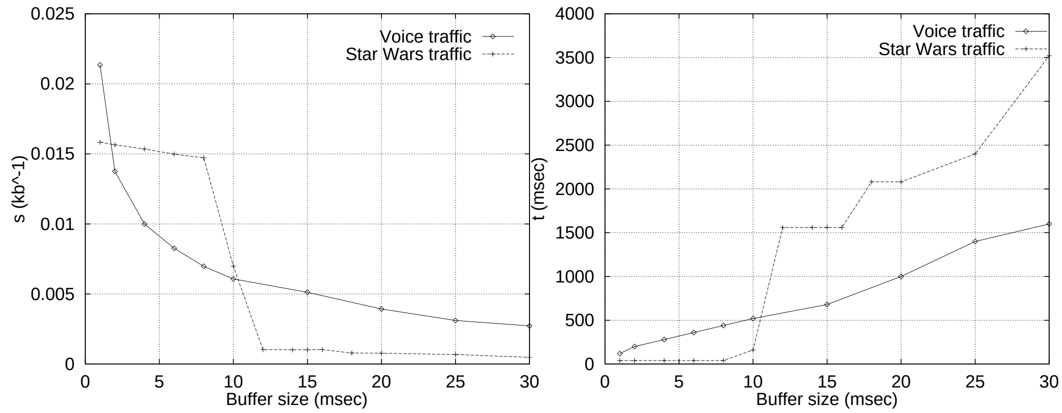
(a) *Star Wars* traffic (b) MPEG-1 traffic with various contents

Figure 8. **Parameter *s* for MPEG-1 traffic.** The space parameter s decreases abruptly at some buffer size; this occurs because the buffer has absorbed the fast time scales and only the remaining slow time scales contribute to buffer overflow. The buffer size (measured in msec) for which this occurs decreases as the link capacity increases (left curve). Similar behavior is observed for MPEG-1 traffic with various contents (right figure). [$P(\text{overflow}) \le 10^{-7}$]



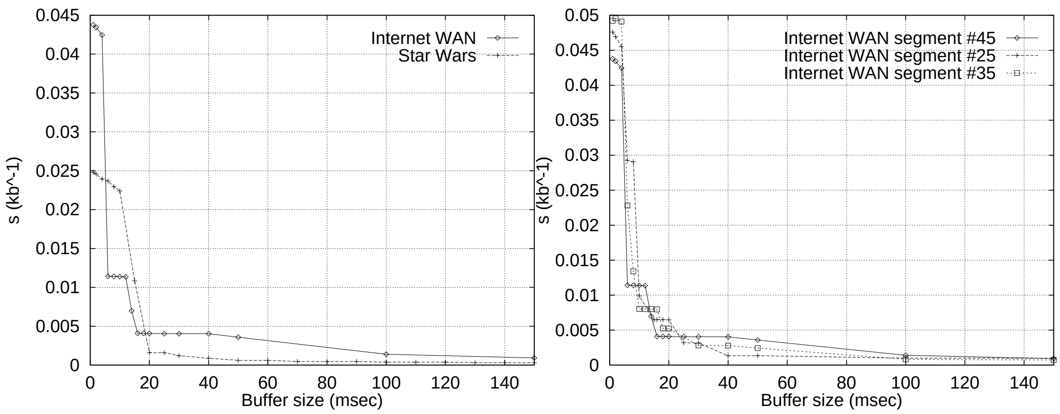
(a) Parameter t (b) Product st

Figure 9. **Parameters *t* and *st* for Star Wars traffic.** The time parameter t increases as the buffer size increases, indicating that slow time scales become important for buffer overflow (left figure). The product st abruptly increases for some buffer sizes (right figure); at this point slow time scales become important for buffer overflow. [$P(\text{overflow}) \le 10^{-7}$]



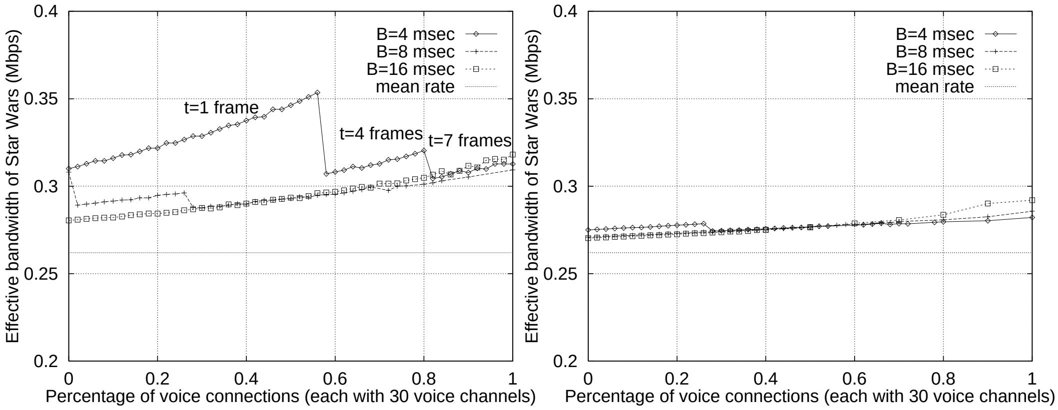
(a) Parameter s (b) Parameter t

Figure 10. **Parameters s, t for Star Wars and voice traffic.** Whereas for MPEG-1 traffic parameter s abruptly decreases for some buffer size (indicating that slow time scales become important for buffer overflow), for voice traffic it gradually decreases, with a rate that also decreases as the buffer size increases (left figure). This indicates a smoother change of the time scales for voice traffic. Parameter t for voice traffic increases linearly with the buffer size, unlike the case of MPEG-1 traffic where it exhibits abrupt jumps. [$P(\text{overflow}) \leq 10^{-7}$]



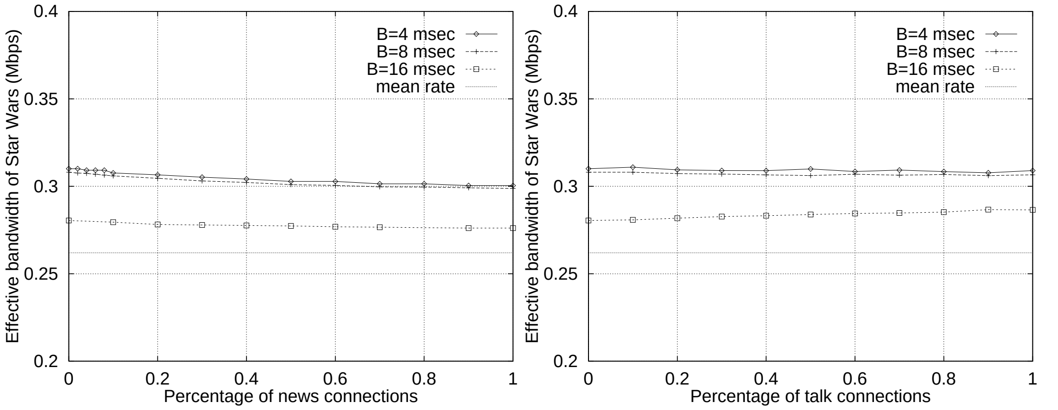
(a) Internet WAN & Star Wars traffic (b) Different Internet WAN segments

Figure 11. **Parameter s for Internet WAN and Star Wars traffic.** Whereas for MPEG-1 traffic the values of parameter s form two distinct regimes, corresponding to the two distinct time scales that are important for buffer overflow, for Internet WAN traffic they form more than two regimes, indicating that for such traffic there are more than two time scales which, for different buffer sizes, become important for buffer overflow (left figure). Also, different segments of Internet traffic have different values for s (right figure). [$C = 34$ Mbps, $P(\text{overflow}) \leq 10^{-4}$]



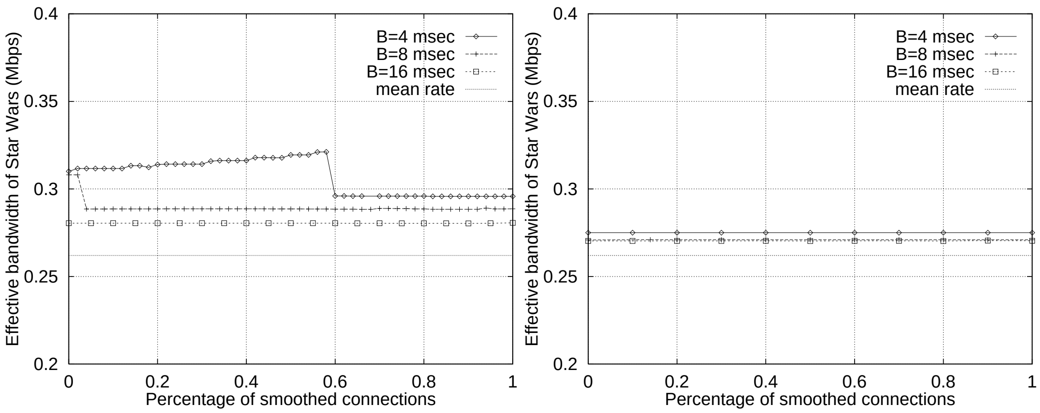
(a) $C = 155$ Mbps (b) $C = 622$ Mbps

Figure 12. **Dependence of the effective bandwidth on the traffic mix containing Star Wars and voice traffic.** The effective bandwidth of the *Star Wars* stream changes slowly for certain ranges of the traffic mix. Furthermore, the sensitivity of the effective bandwidth on the traffic mix decreases as the link capacity or buffer size increases. [$P(\text{overflow}) \leq 10^{-7}$]



(a) *Star Wars* + news (b) *Star Wars* + talk show

Figure 13. **Dependence of the effective bandwidth on the traffic mix containing Star Wars + news/talk show traffic.** The content of MPEG-1 traffic has a small effect on the effective bandwidth; it is the MPEG-1 frame structure that has a larger effect. [$C = 155$ Mbps, $P(\text{overflow}) \leq 10^{-7}$]



(a) $C = 155$ Mbps (b) $C = 622$ Mbps

Figure 14. **Dependence of the effective bandwidth on the traffic mix containing Star Wars and smoothed Star Wars traffic.** Observe that (1) the effects of the traffic mix on the effective bandwidth decrease when the link capacity or buffer size increases, (2) there are cases where increasing the buffer size has a very small effect on the effective bandwidth, e.g., at $C = 622$ Mbps the curves for $B = 8$ msec and $B = 16$ msec practically coincide, and (3) for some buffer sizes, smoothing has no effect on the effective bandwidth, e.g., $C = 155$ Mbps, $B = 8$ msec (left graph) and $C = 622$ Mbps and $B = 4, 8$, and 16 msec (right graph). [$P(\text{overflow}) \leq 10^{-7}$]

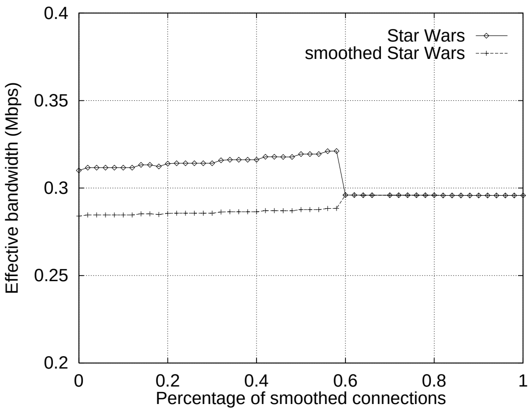


Figure 15. **Effective bandwidth for Star Wars and smoothed Star Wars traffic.** When the percentage of smoothed traffic is small, the time parameter t (= 40 msec) is smaller than the interval over which smoothing was performed (80 msec), and the amount of workload $X_j[0,t]$ that appears in the effective bandwidth formula (1) is smaller for the smoothed stream. Hence the effective bandwidth of the smoothed stream is smaller than the effective bandwidth of the unsmoothed stream. When the percentage of smoothed traffic becomes larger than about 60%, the parameter t becomes equal to the time interval over which smoothing was performed. Hence smoothing has no effect, and the unsmoothed and smoothed streams have the same effective bandwidth. [$C = 155$ Mbps, $B = 4$ msec, $P(\text{overflow}) \leq 10^{-7}$]