On the Performance of an Effective Bandwidths Formula

Costas Courcoubetis*, George Fouskas*
and Richard Weber†

Abstract

At a buffered switch in an ATM (asynchronous transfer mode) network it is important to know what combinations of different types of traffic can be carried simultaneously without risking more than a very small probability of overflowing the buffer. We argue that for many stationary sources an effective bandwidth is given by a formula involving just the source mean rate, its index of dispersion and the size of the buffer. We report computational experiments that support the practicality of this formula.

Keywords: COMMUNICATIONS, EFFECTIVE BANDWIDTHS, LARGE DEVIATIONS, STATIONARY PROCESSES

1 Effective bandwidths

The traffic in an ATM network is packaged in cells and carried over links between switches in the network. Traffic sources are bursty and so for periods of time cells may arrive at a switch faster than they can be switched to output links. For this reason, switches are buffered and a problem is to know how much total traffic can be carried while keeping the probability of buffer overflow and resulting cell loss very small.

Suppose that a switch handles M classes of traffic, consisting of N_i sources of class i, $i = 1, \ldots, M$. The number of cells that the switch can handle per second is c, the bandwidth of the switch. Quality of service, denoted QoS, is partly determined by the rate and which cells are lost from the buffer. This can be related to the probability that a buffer overflows during a busy period and a number of authors have described models in which QoS equates with satisfaction of the constraint

$$c \geq \sum_i N_i E_i, \tag{1}$$

*Department of Computer Science, University of Crete and Institute of Computer Science, FORTH, Greece
†University Engineering Department, Management Studies Group, Mill Lane, Cambridge CB2 1RX, U.K.

where E_i, is the *effective bandwidth* of a source in class i, $i = 1, \ldots, M$. The effective bandwidth of a bursty source is clearly greater than its average rate. However, because at any given moment some sources will be producing cells above their average rate and other sources will be producing cells below their average rate, there is potential for statistical multiplexing. Thus a source's effective bandwidth need not be as great as its peak rate.

Of course, as other authors have noted, e.g., [9], the motivation for seeking to assign effective bandwidths to bursty ATM sources is that, if this can be done, then problems of admission control and routing in ATM networks resemble those in circuit-switched networks. Subsequent research can focus on how ideas from circuit-switched networks, (such as the well-developed theory of trunk reservation and dynamic routing), can be applied to ATM networks.

This paper builds on the work of Courcoubetis and Walrand [4], Veciana, Olivier and Walrand [5], Kesidis and Walrand [10], Gibbens and Hunt [6] and Kelly [9]. Kelly obtained effective bandwidths for a problem of controlling the average work seen by a customer arriving to a $GI/D/1$ queue. Courcoubetis and Walrand obtained effective bandwidths for a model in which the number of cells that a source delivers to the buffer at discrete time points is a Gaussian stationary process. Gibbens and Hunt, Kesidis and Walrand and Veciana et al. have obtained asymptotics for the probability of buffer overflow when the source rate is modulated by a continuous time Markov process.

In Section 2 we present and comment upon a formula for effective bandwidth for stationary sources. The formula takes the form of the first two terms in an asymptotic expansion that is accurate for large buffer sizes. It expresses the effective bandwidth in terms of just two parameters: the mean source rate and its index of dispersion. The index of dispersion measures the burstiness of the source. In Section 3 we present two models of traffic sources and comment upon on line measurement of effective bandwidths. In section 4 we examine the success of our effective bandwidths in guaranteeing quality of service while utilizing the capacity of the switch as fully as possible.

2 A Simple Formula for Effective Bandwidths

Suppose that at epoch n a typical source in class i delivers to the buffer a number of cells that is distributed as X_n^i, where $\{X_1^i, X_2^i, \ldots\}$ is a stationary process, of correlelated random variables. Courcoubetis and Walrand have given the following formula for the effective bandwidths when this stationary process is Gaussian.

$$E_i = m_i + \frac{\delta \gamma_i}{2B}, \tag{2}$$

where

$$\gamma_i = \lim_{n\to\infty} (1/n) E\left[\left(\sum_{i=1}^{n} X_i\right)^2\right].$$

γ_i is commonly called the *index of dispersion*. It is also π times the spectral density evaluated at 0, i.e.,

$$\gamma_i = \pi f_i(0) = \gamma(0) + 2\sum_{k=1}^{\infty} \gamma(k),$$

where $\gamma(k)$ is the k'th order autocovariance. The above converges for well-behaved, purely nondeterministic second order stationary processes. The importance of γ can be compared with the finding of Whitt that the coefficient of diffusion of the arrival process is important in evaluating the heavy traffic mean queue length for the $G/D/1$ queue. Note that γ can be estimated from the data by spectral estimation techiques. (See, for example Chatfield [2].) It is attractive that effective bandwidths might be estimated observed data, since it is unlikely that any theoretical model is rich enough to adequately model all traffic sources. Estimation of γ is an alternative to the on-line estimation procedure proposed in earlier work [3]. We believe that (2) is an appropriate measures of the effective bandwidths not only for the Gaussian sources considered in [4], but for many other stationary sources. Courcoubetis and Weber have stated this as follows.

Theorem 1 *For B large, an appropriate measure of effective bandwidth is*

$$E_i^* = m_i + \frac{\delta\gamma_i}{2B}. \tag{3}$$

The argument is based on a large deviations calculation. Suppose that over n epochs the N_i sources of type i deliver cells to the buffer at a rate a_i, where a_i exceeds the mean rate m_i, such that $\sum_i N_i a_i = nc + B$. The probability with which this occurs when n is large is approximately,

$$\exp\left\{-n\sum_i N_i I_i(a_i) + o(1)\right\},$$

where $o(1)$ tends to 0 as $n \to \infty$, $I_i(x) = \sup_\theta[\theta x - \phi_i(\theta)]$, and

$$\phi_i(\theta) = \lim_{n\to\infty} (1/n) \log E\{\exp[\theta(X_1^i + \cdots + X_n^i)]\}.$$

The theory of large deviations implies that the probability of overflow has an asymptotic in B determined by the most likely way overflow can occur, and so we maximize the above with respect to the a_i and n and derive an asymptotic expression in terms of powers of $1/B$. This requires certain assumptions on the convergence to $\phi_i(\theta)$ and the observation that when B is large the maximizing value of n is also large. Neglecting terms of order $o(\delta/B)$ the condition that this probability is less than $\exp(-\delta)$ takes the form $c \geq \sum_i N_i E_i^*$.

The following observations lend support to the above heuristic argument. Consider what happens if a source of type i is pre-smoothed, by in a buffer that effects some linear filtering, say $\bar{X}_n^i = a_0 X_n^i + a_1 X_{n-1}^i + \cdots + a_p X_{n-p}^i$, taking $a_0 + \cdots + a_p = 1$, so that $E[\bar{X}_n^i] = E[X_n^i] = m_i$. Then since $\bar{f}_i(0) = |T(0)|^2 f_i(0)$ and the transfer function, $|T(0)| = |\sum_i a_i| = 1$, we have $\bar{\gamma}_i = \gamma_i$.

This suggests that Courcoubetis and Walrand's result for stationary Gaussian source might be viewed as arising from linear filterings a Gaussian process of uncorrelated random variables. Pre-smoothing, by averaging the inflows of several periods, tends to decrease the variance, but it simultaneously increases higher order autocovariances, and the combined effect is that the effective bandwidth is unchanged. This is consistent with what one would expect, because the effects of pre-smoothing are masked by a very large buffer, and it is large buffers with which E_i^* is concerned.

A second observation that supports the use E_i^* is the fact that we obtain exactly the same condition on $(N_1, \ldots, N_M)$ regardless of how we define a time epoch. In other words, if the definition of an epoch is changed from 1 to 2 seconds, so that the numbers of cells produced by a class i source are now $Y_n^i = X_{2n-1}^i + X_{2n}^i$, $n = 1, 2, \ldots$, then the effect is the same as if the process had been smoothed with $a_0 = 0.5$, $a_1 = 0.5$ and then multiplied by two. Since smoothing leaves γ_i unchanged and the multiplication by two doubles it, things are just as they should be, since in the new model, c and m_i will also double.

3 Traffic Source Models

In this section we describe some simple and powerful models, which can be used to characterize two real-time traffic sources: voice and videotelephone.

An Autoregressive Markov model

Maglaris, et al. (see [11]) use the following continuous-state, autoregressive, discrete-time Markov process as a model for videotelephone sources. It is very simple to implement in simulation experiments. Let X_n represent the bit rate of a videotelephone source during the nth frame (assuming that the coder uses the *conditional replenishment* scheme), and

$$X_n = aX_{n-1} + bw_n,$$

where $\{w_n\}$ is a sequence of independent Gaussian random variables, with mean η and variance 1, and a, b are constants, with $|a| < 1$. The steady-state average m, the kth order autocovariance

$\gamma(k)$ and the index of dispersion γ are given by

$$m = \frac{b}{1-a}\eta, \qquad \gamma(k) = \frac{b^2}{1-a^2}a^k, \quad k \geq 0, \qquad \gamma = \left(\frac{b}{1-a}\right)^2.$$

The effective bandwidth from (3) is then

$$E^* = \frac{b}{1-a}\eta + \frac{\delta}{2B}\left(\frac{b}{1-a}\right)^2.$$

Matching the average bit rate and the autocovariances with actual data, Maglaris *et al.* suggest that this model provides a good approximation of the bit rate of a videotelephone source, if $\eta \simeq 0.572$, $a \simeq 0.8781$, $b \simeq 0.1108$. Using these values, we get $X_n \in [0.0, 10.57]$ Mb/s, $m = 3.89$ Mb/s, and $\gamma(0) = 3.01$ Mb^2/s^2.

An On/off Markov fluid model

An alternative and much studied traffic model is the on/off Markov fluid model in which a source is either active and producing a cells per second at constant rate, or idle and producing no cells. The two states alternate according to a continuous time Markov process and the lengths of time spent in the active and idle states are exponentially distributed, with means of $1/\mu$ and $1/\lambda$ respectively. The mean rate of the source is $m = \lambda a/(\lambda+\mu)$ and the effective bandwidth given by (3) can be shown to be

$$E^* = \frac{\lambda a}{\lambda+\mu} + \frac{\delta\lambda\mu a^2}{B(\lambda+\mu)^3}. \tag{4}$$

Of course we expect the effective bandwidth E^* to be less than the peak rate a. This is the case if and only if $B/\delta \geq m/(\lambda+\mu)$. Now $1/(\lambda+\mu)$ can be considered as the time constant of the system; and so for (4) to be meaningful, B/δ should be several times the average number of cells produced during the time constant of the system. This condition is met for realistic values of the parameters.

Gibbens and Hunt [6], have suggested the following formula for the effective bandwidth of an on/off Markov fluid source, when the QoS requirement is $P(\text{buffer overflow}) \leq \exp(-\delta)$.

$$E^\dagger = \left\{a\delta/B - \mu - \lambda + \sqrt{(\mu-\lambda-a\delta/B)^2 + 4\lambda\mu}\right\} \Big/ (2\delta/B). \tag{5}$$

Not surprisingly, (4) is precisely the first two terms of an asymptotic expansion of (5) in powers of δ/B. Note also that Veciana, et al. in [5] have shown that for a two-state Markov modulated fluid the probability of buffer overflow has an asymptotic $\exp(-B\rho)$, where

$$\rho = N\frac{(\lambda+\mu)c - N\lambda a}{c(Na-c)}.$$

Writing $B\rho > \delta$ as a quadratic condition in c/N, and requiring $c > NE^\dagger$ also leads to (5).

Voice source: The model can be used for a voice source with ADPCM modulation and speech activity detection mechanisms. Ibrahim, et al. [8] suggest parameters of $a = 64$Kb/s, $1/\lambda = 650$ms and $1/\mu = 352$ms.

Videotelephone source: Maglaris, et al. [11] use the superposition of 20 on/off Markov fluid sources, or 'video minisources', to model a single videotelephone source. For each video minisource, they use the following parameters: $a = 945$Kb/s, $1/\lambda = 1.273$s, and $1/\mu = 0.321$s. These give values $X_n \in [0, 18.9]$ Mb/s, $m = 3.81$ Mb/s, and $\gamma(0) = 2.87$ Mb^2/sec^2.

Figure 1 shows the effective bandwidths E^* and $E^\dagger$ for these voice and videotelephone sources. Note that the curves are close and not much is lost by using the simpler E^* that can be estimated from the the mean and coefficient of dispersion.

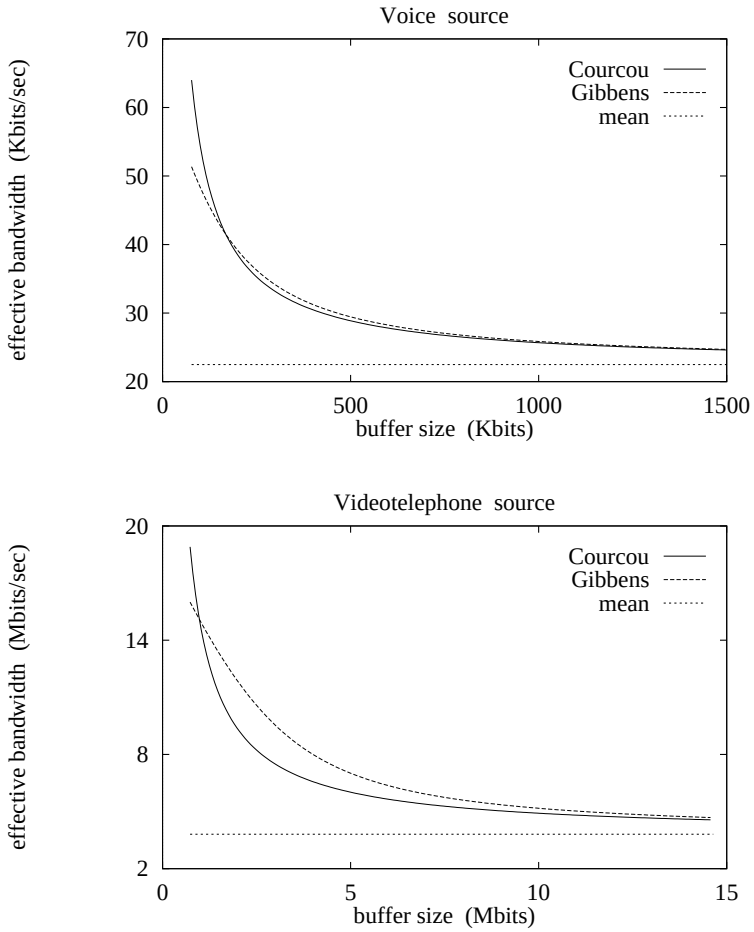

Figure 1: Comparison of E^* and $E^\dagger$

It is interesting to consider the time required to estimate on-line the coefficient of dispersion, and hence the effective bandwidth. Suppose X_j is the number of cells produced during the jth

epoch, and τ is the duration of each epoch. A simple estimation of γ may be obtained from

$$\hat{\gamma} \cdot \tau = c(0) + 2 \sum_{k=1}^{L} c(k).$$

where

$$c(k) = \frac{1}{(n-k)} \sum_{j=1}^{n-k} X_j X_{j+k} - \frac{1}{(n-k)^2} \sum_{j=1}^{n-k} X_j \sum_{j=1}^{n-k} X_{j+k},$$

$k = 0, 1, \ldots, n-1$, are a crude estimates of the covariances. We have found that it works well to take L approximately equal to the time constant of the system, $1/(\lambda + \mu)$.

Figures 2 and 3 show the time required to obtain an accurate estimates of γ for three models. This is about one minute for the voice source model (figure 2) and several minutes for the videotelephone source models (figure 3). (Notice that for the autoregressive Markov model of a video source $\hat{\gamma}$ does not converge to the theoretically calculated value of $b/(1-a)^2$, because the theoretical value ignores the fact that in practice we truncate negative values of X_n to zero.)

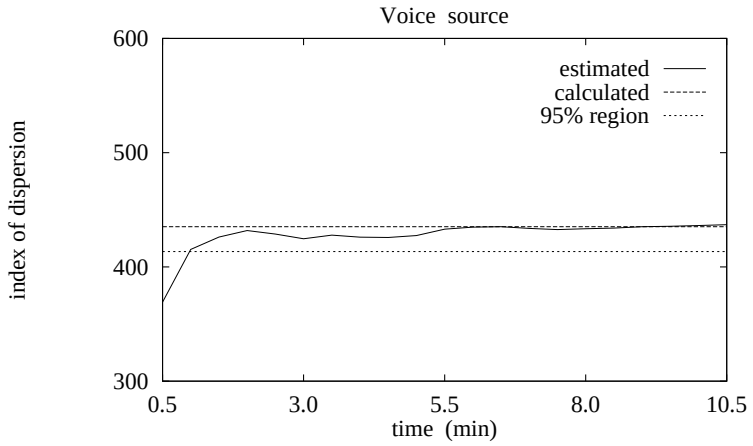

Figure 2: Convergence of estimator of γ, for the voice source model

4 Use of Effective Bandwidths in Switch-loading

In this section we investigate the practicality of using the effective bandwidths E^* to estimate the traffic a switch can carry and examine the how well the actual probability of overflow achieved matches a QoS goal.

Consider a single on/off Markov fluid source model, with $\mu = 1$ and $a = 1$. Following the approach of Anick, et al. [1], the probability that a buffer of size B overflows during a busy

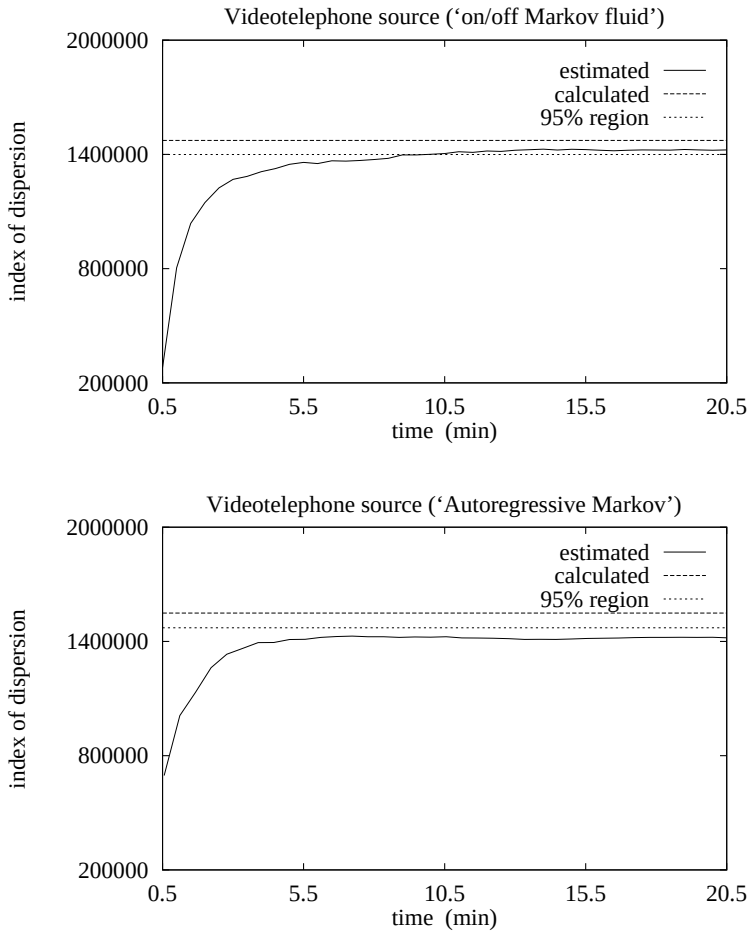

Figure 3: Convergence of estimator of γ, for the videotelephone source models

period when it is served at constant rate c and fed by this single on/off Markov fluid source, is exactly

$$G(B) = \frac{\lambda}{(1+\lambda)c} e^{-(1+\lambda-\lambda/c)B/(1-c)}.$$

Suppose, using (4), c is set equal to the effective bandwith of this source, i.e., $c = \lambda/(1+\lambda) + \delta\lambda/B(1+\lambda)^3$. Then

$$G(B) = \frac{1}{1+\delta\lambda/B(1+\lambda)^2} \exp^{-\delta/(1+\delta/B(1+\lambda)^2)(1-\delta\lambda/B(1+\lambda)^2)}.$$

This may or may not be less than the QoS requirement, $\exp(-\delta)$. But it is easy to check that the QoS is achieved if, for example, $\lambda \geq 1$. Also, $G(B)$ is decreasing in λ for $\lambda \leq 1$. Thus we can conclude that the QoS requirement is acheived if the mean time a source remains off is short compared to the mean time it is on. Figure 4 shows the two different forms of $G(B)$ that we have observed. In the first we take $\lambda = 0.25$, $\delta = 15$, and the QoS is not achieved. In the second we take $\lambda = 2.5$, $\delta = 15$, and the QoS is achieved.

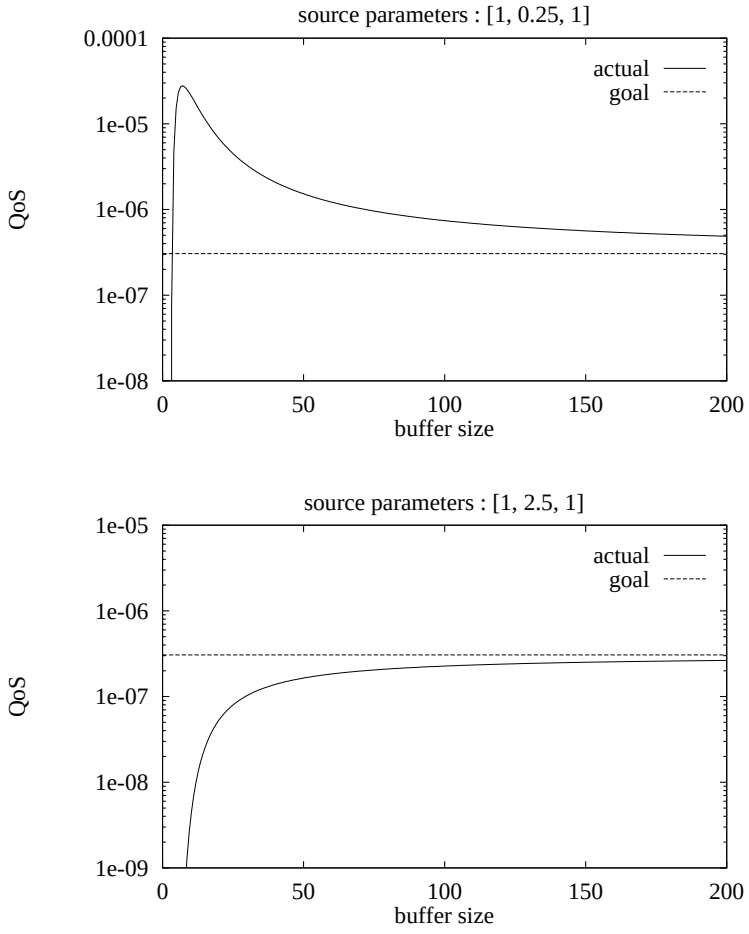

Figure 4: QoS achieved by use of E^*

Figure 5 is for $(a, \lambda, \mu) = (1, 2, 0.8)$, $\delta = 15$, but for 10, 20, 50 and 100 sources sharing the switch. Here, $G(B)$ does not have a closed form, but it can be calculated using the method described by Anick, et al. [1]. This figure is representative of experiments with other parameters and demonstrates that using E^* leads to performance that differs from the specified QoS in a way that is pretty well the same no matter what the number of sources carried by the switch. Note that the use of E^* is conservative, but not too conservative: the probability of overflow is reduced by a factor of $0.5 - 0.8$ for reasonably sized buffers, not by an order of magnitude. This is what we would expect, since the original large deviation approximation is only correct to within a multiplicative constant.

We have considered above traffic composed of identical on/off Markov fluid sources. We conclude the paper with simulation results for mixtures of nonidentical on/off Markov fluid sources and for autoregressive Markov sources. The results for 20 autoregressive Markov source (Figure 6) and a mixture of 10 video and 100 voice sources (Figure 7) are representative of our

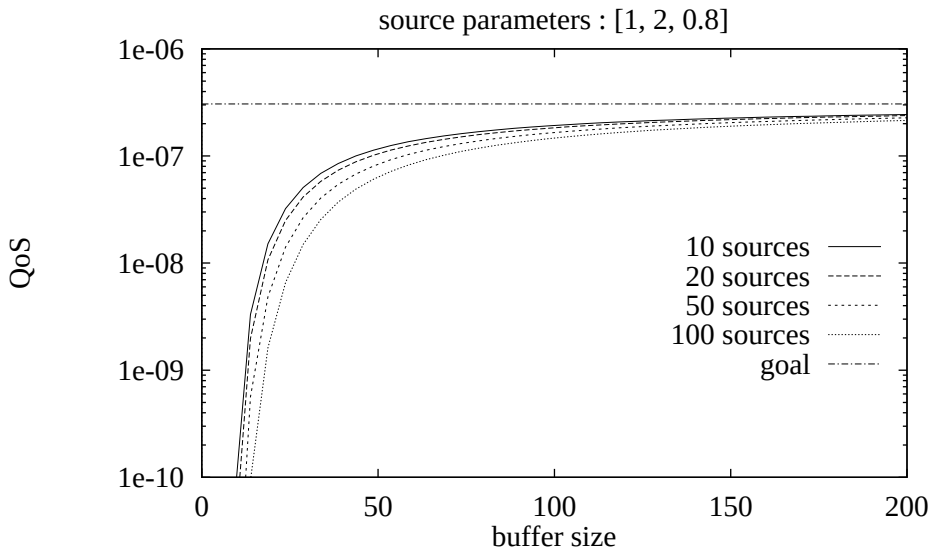

Figure 5: QoS achieved by use of E^* and multiple sources

experimental results with various traffic mixes. In these simulation experiments, the QoS level is set at an unrealistically high overflow probability of $\exp(-9) = 0.0001$, so that simulation results can be obtained more quickly. For buffer sizes of 8Mbits or more, we observe that using effective bandwidths of E^* load the switch conservatively, but not too much so, since the probability of buffer overflow is within an order of magnitude of the QoS goal $\exp(-9)$.

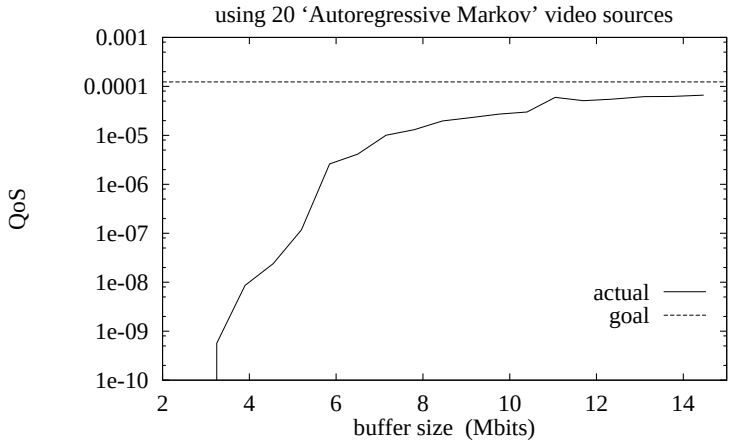

Figure 6: QoS for 20 autoregressive video sources

Figures 8 shows the ratio of the number of voice calls that may be accepted at a 4 Mb/s switch relative to the number that could be accepted if they were constant rate sources of the same mean rate. In other words, they show the proportion of the theoretical maximum

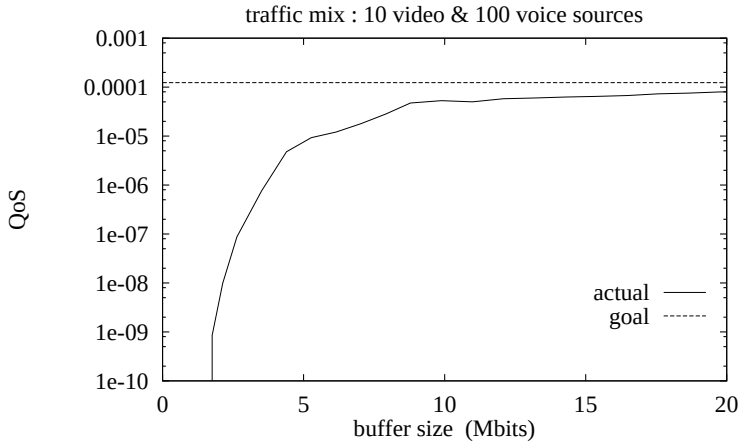

Figure 7: QoS for 10 video and 100 voice sources

utilization of the switch. Figure 9 presents similar information for video minisources at a 25 Mb/s switch. Utilizations are displayed for use of E^* and also $E^\dagger$. The utilizations are very close, though use of E^* results in slightly greater utilization (while providing a conservative use of the switch). The upper curve in each figure is the best utilization that can be attained under this QoS constraint. The lower horizontal line shows the poor utilization that would be achieved if effective bandwidths were taken equal to peak rates. The upper horizontal line shows the utilization using the approach of Kãs and Kleinewillinghõfer-Kopp, [7], [12]. This is based on the idea that the distribution of the bit rate of the superposition of N traffic sources can be approximated by a Gaussian distribution, with mean $\sum_{i=1}^{N} m_i$, and variance $\sum_{i=1}^{N} \sigma_i^2$. Now it can be shown that $P(\text{total bit rate} > c) \leq \epsilon$, if

$$c \geq \sum_{i=1}^{N} m_i + \sqrt{-\log(2\pi\epsilon^2) \sum_{i=1}^{N} \sigma_i^2}. \tag{6}$$

But, clearly $P(\text{buffer is overflowing}) \leq P(\text{total bit rate} > c)$, thus (6) can be used to ensure $P(\text{buffer is overflowing}) \leq \epsilon$. Note, however, that this requirement does not take the form of effective bandwidths since the left hand side of (6) cannot be written in the form of $c \geq \sum_i E_i$.

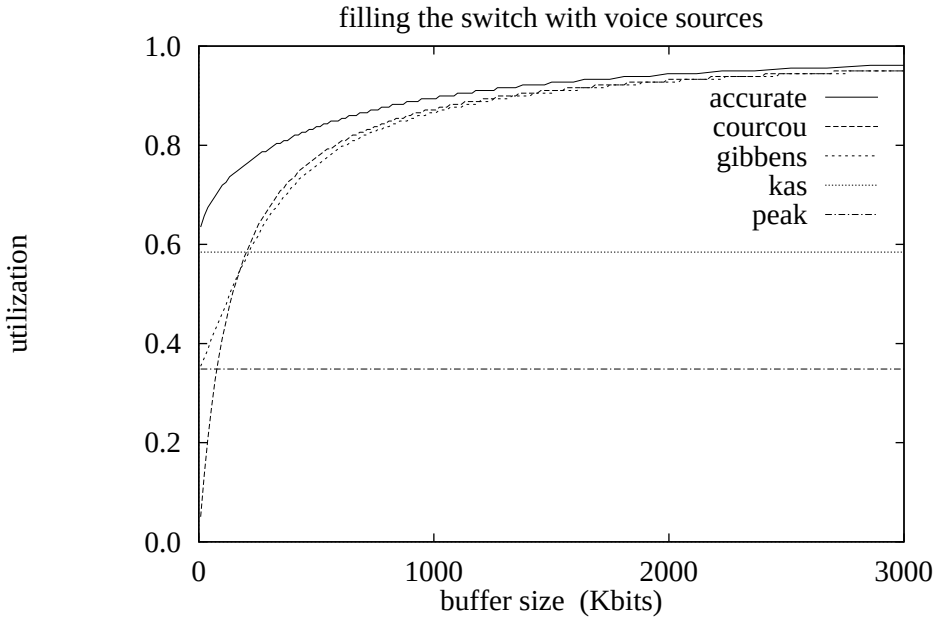

Figure 8: Utilization of switch-loading methods, for the voice source model

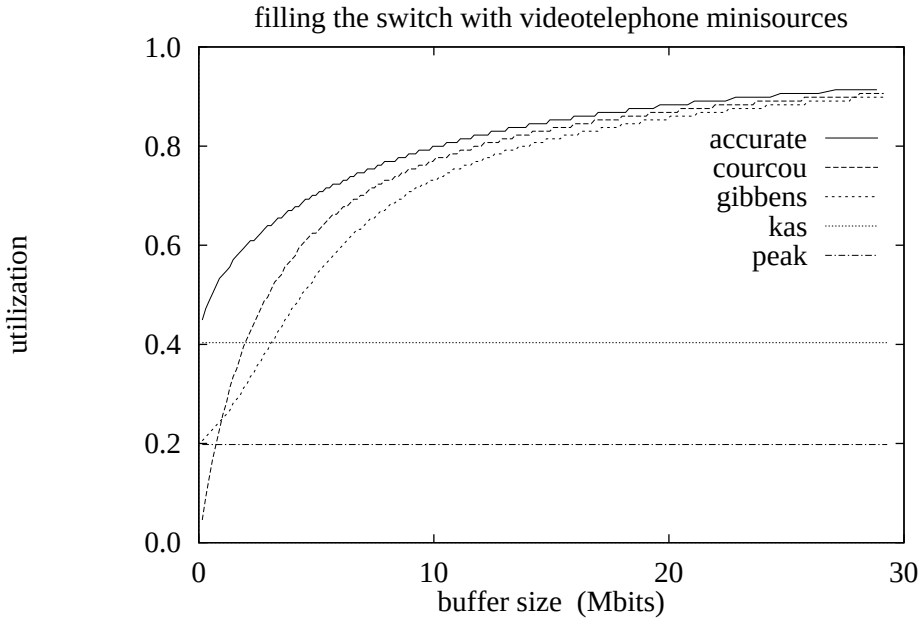

Figure 9: Utilization of switch-loading methods, for the videotelephone minisource model

References

[1] D. Anick, D. Mitra, and M. M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *The Bell System Technical Journal*, 61(8):1871, October 1982.

[2] C. Chatfield. *The Analysis of Time Series: Theory and Practice.* Chapman and Hall, London, 1975.

[3] C. Courcoubetis, G. Kesidis, A. Ridder, J. Walrand, and R.R. Weber. Admission control and routing in ATM networks using inferences from measured buffer occupancy. ORSA/TIMS special interest meeting, Monterey, January, 1991. (to also appear in *IEEE Trans. Communications*).

[4] C. Courcoubetis and J. Walrand. Note on the effective bandwidth of ATM traffic at a buffer. unpublished manuscript.

[5] G. De Veciana, C. Olivier, and J. Walrand. Large deviations for birth death Markov fluids. unpublished manuscript.

[6] R. Gibbens and P. Hunt. Effective bandwidths for the multi-type UAS channel, 1991. Statistical Laboratory, University of Cambridge, preprint.

[7] R. Guérin, H. Ahmadi, and M. Naghshineh. Equivalent capacity and its application to bandwidth allocation in high-speed networks. *IEEE JSAC*, 9(7):968, September 1991.

[8] Ibrahim W. Habib and Tarek N. Saadawi. Multimedia traffic characteristics in broadband networks. *IEEE Communications Magazine*, page 48, July 1992.

[9] F.P. Kelly. Effective bandwidths at multi-class queues. *Queueing Systems*, **9**:5–16, 1991.

[10] G. Kesidis and J. Walrand. Effective bandwidths for multiclass Markov fluids and other ATM sources. Memorandum No. UCB/ERL M92/40.

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

[12] RACE Project R1022. Technology for ATD. RACE document TG-123-0006-FD-CC.