Effective Bandwidths for Stationary Sources

COSTAS COURCOUBETIS*
AND
RICHARD WEBER†

first revision 1 September, 1993
final revision 16 May, 1994

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 show that a simple and serviceable measure of effective bandwidths may be computed for stationary traffic sources. For large buffers the effective bandwidth of a source is a function only of its mean rate, index of dispersion and the size of the buffer.

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 the problem is to know how much total 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 and has capacity to handle c cells per second. A number of authors have described models for which a quality of service criterion can be met while the traffic is composed of N_i sources of class i, $i = 1, \dots, M$, if and only if

$$\sum_{i=1}^{M} N_i \beta_i \le c,$$

where β_i, is called the *effective bandwidth* of a traffic source in class i. Intuitively, the bursty character of a source means that its effective bandwidth should be greater than its average rate. However, because at any moment some sources will deliver cells to the buffer at above their average rate and other sources will do so at below their average rate, there is potential for statistical multiplexing. Thus β_i need not be as great as the peak rate of source i.

The notion of effective bandwidths is in the mainstream of present thinking about ATM traffic. Of course 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

*University of Crete, Heraklion, Greece
†University Engineering Department, Management Studies Group, Mill Lane, Cambridge CB2 1RX, U.K.

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 [6], De Veciana, Olivier and Walrand [7], Gibbens and Hunt [10] and Kelly [11]. Kelly obtained effective bandwidths for a problem of controlling the average work seen by a customer arriving to a $D/GI/1$ 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. Asymptotics are obtained in [7] and [10] for the frequency of buffer overflow when the source rate is modulated by a continuous time Markov process. Recently, Kesidis, Walrand and Chang [12] and De Veciana and Walrand [8] have also obtained effective bandwidths for a large class of stationary sources under conditions similar to ours.

In sections 2 and 3 we extend the work of [6] to non-Gaussian stationary sources and compute an effective bandwidth approximation. We argue that the effective bandwidth of a source can be approximated in a manner that makes sense for large buffers and which is a function of just two parameters: the source mean rate and index of dispersion. Thus the index of dispersion is identified as an appropriate measure of burstiness. Our effective bandwidths agree with those of Kelly [11] and they agree with the formula obtained in [7] and [10] for a two-level Markov modulated fluid when the size of the buffer is large.

2 White noise stationary sources

Consider a switch that carries traffic comprised of N_i independent sources of class i, $i = 1, \ldots, M$. Suppose time is discrete and that at each discrete epoch a source in class i delivers to the buffer a number of cells that is independently and identically distributed as X_i, a random variable with mean μ_i and variance σ_i^2. As in [6], we note that the way a buffer can fill during a busy period is for the sources to produce cells at a mean rate that exceeds $\sum_i N_i \mu_i$ for such time until the buffer fills. Define the logarithmic moment generating function,

$$\varphi_i(\theta) = \log E[\exp(\theta X_i)].$$

Let $P(B, \alpha, n)$ denote the probability that during $n = \lceil B/\alpha \rceil$ epochs cells should arrive at an average rate of $(c + \alpha)n$ cells per epoch, and so a buffer of size B that starts empty should be filled during the nth epoch. If the total input to the buffer in epoch i is Y_i, then $P(B, \alpha, n) = P(Y_1 + \cdots + Y_n \approx (c + \alpha)n)$ and this can be estimated from Cramer's theorem (see [2]) as

$$P(B, \alpha, n) = \exp\left(-\frac{B}{\alpha} I(c + \alpha) + o(B)\right),$$

where $o(B) \to 0$ as $B \to \infty$, and $I(\cdot)$ is the 'rate function' defined as

$$I(x) = \sup_\theta \left[\theta x - \sum_{i=1}^{M} N_i \varphi_i(\theta)\right].$$

The probability that the buffer fills during a busy period, say $P(B)$, can be taken as a surrogate for the cell loss rate. For since we are supposing that the buffer is large enough that it does not fill during most busy periods, the mean length of a busy period is nearly independent of B. But if the buffer does fill during a busy period, the amount of cell loss that then ensues is also nearly independent of B. So by a renewal reward calculation in which the start of busy periods are the renewal times, the cell loss rate is just some constant times $P(B)$. Now $P(B)$ is the sum of a number of terms, each of which is the probability of one way it can occur, such as $P(B, \alpha, n)$. If the number of these terms were finite, then we could say that $P(B)$ has the same asymptotic behaviour as the maximum term. This is simply the argument of Laplace, that

$\lim_{B\to\infty}(1/B)\log[\sum_{i=1}^{n}\exp(-B\eta_i)] = \min_i \eta_i$. However, the number of terms is not finite, and it is only a heuristic that $P(B)$ can be approximated by the maximum term. Although it is difficult to make this heuristic rigorous, it is a standard idea in the theory of large deviations that when an unlikely event occurs then it occurs in the most likely of the unlikely ways. The argument will be made with more care in section 3.

Using this heuristic, $\lim_{B\to\infty}(1/B)\log P(B) \leq -\delta$ if and only if $I(c+\alpha)/\alpha \geq \delta$ for all α. This occurs if and only if for each α there exists a θ such that

$$\frac{\alpha(\delta-\theta)+\sum_{i=1}^{M} N_i\varphi_i(\theta)}{\theta} \leq c.$$

The above condition is certainly satisfied if it is satisfied for $\theta = \delta$, that is, if

$$\sum_i N_i\,\varphi_i(\delta)/\delta \leq c.$$

This is the effective bandwidth result given by Kelly [11], based on a bound on the tail of the workload found by a customer arriving to a $D/GI/1$ queue. Since when B is large we will be inclined to take δ small, so that $P(B) \sim \exp(-\delta B)$, it makes sense to expand $\varphi_i(\delta)/\delta$ in powers of δ and ignore terms that are $o(\delta)$. This suggests use of effective bandwidths $\beta_i = \mu_i + \delta\sigma_i^2/2$. With the motivation of this section, we now turn to the general case.

3 Stationary sources

Consider the more realistic case, that from moment to moment there is correlation in the rate at which cells are produced by a source.

3.1 Index of dispersion and effective bandwidth

Suppose that in epoch n a source delivers to the buffer a number of cells that is distributed as X_n, where $\{X_1, X_2, \dots\}$ is a stationary process of correlated random variables, with mean μ. Courcoubetis and Walrand [6] assume the sources are stationary Gaussian and by the argument of section 1, estimate $P(B,\alpha,n)$ by $\hat{P}(B,\alpha,n) = \exp[-(B/\alpha)I(c+\alpha)]$. $\hat{P}(B,\alpha,n) \leq \exp(-\delta)$ for all α if and only if $\beta < c$, where

$$\beta = \mu + \frac{\delta\gamma}{2B}. \tag{1}$$

and

$$\gamma = \lim_{N\to\infty}\frac{1}{N}\mathrm{var}\left(\sum_{n=1}^{N} X_n\right).$$

That β acts as an effective bandwidth can be seen from the fact that if $\{X_n\}_{n=1}^{\infty}$ is actually the superposition of a number of independent sources, consisting of N_i sources of class i, having mean μ_i and asymptotic variance γ_i, $i = 1,\dots,M$, then $\mu = \sum_i N_i\mu_i$ and $\gamma = \sum_i N_i\gamma_i$. In fact, $P(B,\alpha,n) = \exp[-(B/\alpha)I(c+\alpha)+o(B)]$, and considering the $o(B)$ term it is more appropriate to adopt the constraint $\lim_{B\to\infty}(1/B)\log P(B,\alpha,n) \leq -\delta$. This leads to $\beta = \mu + \delta\gamma/2$.

The quantity γ has an interpretation in terms of the autocovariance structure of the process that holds even when the processes are not Gaussian. This interpretation holds under the following assumption.

Assumption A. *Suppose the stationary process $\{X_n\}$ has kth order autocovariance $\gamma(k)$, and spectral density function $f(\omega) = \sum_{k=-\infty}^{\infty} \gamma(k)\exp(i\omega k)$. Suppose the infinite sum of the autocovariances is absolutely summable and $f(\cdot)$ is continuous at 0. Then*

$$\gamma = \pi f(0) = \sum_{k=-\infty}^{\infty} \gamma(k),$$

where γ is called the index of dispersion.

For the assumption to hold we must have $\gamma(k) \to 0$ as $k \to \infty$ and so the process must be purely nondeterministic, without periodicity or long-term dependencies. It is a technical assumption that is plausible under the assumption that the numbers of cells produced during epochs that are widely separated in time are nearly independent. It is easy to show that it is satisfied by Gaussian stationary sources. It is possible to show that it holds for other processes, such as the Markov modulated fluids that we discuss in section 4.

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 techniques. (See, for example Chatfield [3].) It is attractive that effective bandwidths might be estimated from observed data, since it is unlikely that any theoretical model is rich enough to adequately model all traffic classes. Estimation of γ is an alternative to the on-line estimation procedure proposed in earlier work [5].

3.2 Pre-smoothing and time-slicing a source

We shall show in theorem 2 that $\sum_i N_i \beta_i < c$ can be associated with the guarantee of a certain quality of service constraint when $\beta_i = \mu_i + \delta\gamma_i/2$. This formula for effective bandwidths has several attractive properties.

If the source is pre-smoothed, in a buffer that effects some linear filtering, say $\bar{X}_n = a_0 X_n + a_1 X_{n-1} + \cdots + a_p X_{n-p}$, taking $a_0 + \cdots + a_p = 1$, so that $E[\bar{X}_n] = E[X_n] = \mu$. Then to obtain $\bar{f}(0)$ we multiply $f(0)$ by the transfer function, to obtain $\bar{f}(0) = |T(0)|^2 f(0)$. Since $|T(0)| = |\sum_i a_i| = 1$, we have $\bar{\gamma} = \gamma$.

This suggests that Courcoubetis and Walrand's result for a stationary Gaussian source might be viewed as arising from linear filterings of a Gaussian process of uncorrelated random variables. Pre-smoothing, by averaging the inflows of several epochs, 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 within a very large buffer, and it is large buffers with which our effective bandwidths are concerned. It is still an open issue as to how large should be the buffers of ATM switches. Small buffers have the advantage of allowing less delay. However, as buffers are relatively inexpensive, manufacturers may choose to compete by offering switches with large buffers.

De Veciana and Walrand [8] comment that if $|T(0)| = G < 1$, which happens if the source is thinned, then the bandwidth changes to $G\mu + G^2\delta\gamma/2$. Thus bandwidths can be reduced by thinning, but not by smoothing.

A second observation that supports the use of these effective bandwidths is the fact that we obtain exactly the same condition on $(N_1, \ldots, N_M)$ regardless of how we define a time epoch. For example, if the definition of an epoch is changed from 1 to 2 ms, so that the numbers of cells produced by a source in the epoch labelled n is $X_{2n-1} + X_{2n}$, $n = 1, 2, \ldots$, 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 unchanged and the multiplication by two doubles it, things are just as they should be, since in the new model, c and μ will also double.

Although we have seen that pre-smoothing is not helpful in reducing the effective bandwidth of a source, this is because we do an asymptotic analysis for a large buffer, while fixing the amount of pre-smoothing that is carried out upstream. There may still be some interesting questions regarding pre-smoothing that grows at the same time that B increases. A buffer is required to carry out pre-smoothing, and smoothing over a greater number of epochs requires a larger buffer. So there may be a trade off between buffering used for upstream pre-smoothing (e.g., leaky bucket flow control) and downstream buffering.

3.3 General stationary sources

The previous sections have suggested that (1) states an appropriate measures of effective bandwidths for stationary sources and that the result of [6] for Gaussian sources can be applied to general, non-Gaussian, stationary sources. This is made precise in theorem 2. To do this we make use of the Gärtner-Ellis theorem, which holds under the following assumptions.

Assumptions of the Gärtner-Ellis theorem *Let $Z_1, Z_2, \dots$ be a sequence of random vectors in $\mathcal{R}^d$, possibly dependent. Suppose that*

1. *The asymptotic logarithmic moment generating function, defined as*

$$\varphi(\theta) = \lim_{n\to\infty} n^{-1} \log E\left[\exp(n\theta^\top Z_n)\right],$$

exists for all θ, possibly as $\pm\infty$. The set $\{\theta : \varphi(\theta) \le k\}$ is closed for every finite k.

2. *The origin is in the interior of the effective domain, defined as $D_\varphi = \{\theta : \varphi(\theta) < \infty\}$.*

3. *The derivative, $\nabla\varphi(\theta)$, exists in D_φ°, the interior of D_φ, and $|\nabla\varphi(\theta)|$ tends to infinity as θ approaches the boundary of D_φ°.*

Theorem 1 (Gärtner-Ellis) *Let P_n be the probability distribution of Z_n. Then under the assumptions above, P_n satisfies a large deviation principle with good rate function $I(x) = \sup_\theta[\theta^\top x - \varphi(\theta)]$.*

This means that for any subset of the probability space, A,

$$-\inf_{x\in A^\circ} I(x) \le \varliminf_{n\to\infty} \frac{1}{n}\log P_n(A^\circ) \le \varlimsup_{n\to\infty} \frac{1}{n}\log P_n(\bar{A}) \le -\inf_{x\in\bar{A}} I(x).$$

Notice that we require separate statements for open and closed sets, here taken as A° and $\bar{A}$, the interior and closure of A, respectively. That $I(\cdot)$, is 'a good rate function' means that $\{x : I(x) \le \alpha\}$ is a compact set for all $\alpha < \infty$. For the proof see Dembo and Zeitouni [9].

In this section we adopt a different quality of service criterion than in section 2. We pose a constraint in terms of the proportion of time, $\Phi(B)$, that a buffer of size B is full. This is the proportion of time that there is cell loss and it is clearly linked to sensible measures of quality of service. In section 3.1 of [8], De Veciana and Walrand establish a theorem which is similar to (2) of theorem 2. However, they they take as a constraint the proportion of time that a queue with an infinite buffer contains a workload greater than B. The proof in [8] improves upon the heuristic arguments of section 2, which we had used in a draft of this paper. Theorem 2 now follows a reasoning that is similar, though not identical, to that in [8].

Theorem 2 *Suppose X_i is the aggregate input in epoch i to a buffer of size B, where $\{X_1, X_2, \ldots\}$ is a stationary process. Suppose the sequence $\{Z_n\}$, $Z_n = (X_1 + \cdots + X_n)/n$, satisfies the assumptions of the Gärtner-Ellis theorem, with asymptotic logarithmic moment generating function $\varphi(\theta)$. Suppose the buffer is served at the rate of c cells per epoch, with $E[X_1] < c$. Let $\Phi(B)$ be the proportion of time that the buffer is full. Then*

$$\lim_{B\to\infty} (1/B) \log \Phi(B) \le -\delta \iff \varphi(\delta)/\delta < c. \tag{2}$$

If the input is the aggregate of N_i independent processes, each with mean μ_i and index of dispersion γ_i and each satisfies the assumption A, $i = 1, \ldots, M$, then $\varphi(\delta)/\delta < c$ may be be written as $\sum_i N_i \beta_i < c$, where

$$\beta_i = m_i + \frac{\delta \gamma_i}{2} + o(\delta). \tag{3}$$

Proof

Suppose that we observe the contents of the buffer at the ends of epochs $\{t_i\}_{i=-\infty}^{\infty}$, where $t_i = i[B/\xi]$, for some $\xi > 0$. Suppose the buffer is full when observed at the start of epoch 0. Then there must have been a last epoch prior to 0, say $s < 0$, during which the buffer was empty. Let t_{-i-1} be the observation point equal to, or just prior, to the epoch containing s. Let $S_n = X_1 + \cdots + X_n$. Between t_{-i-1} and 0 the buffer must have received at least $i[B/\xi]c + B$ cells. Thus for any $\theta_i > 0$, the probability that the buffer is full, say $\Phi(B)$, satisfies

$$\begin{aligned}
\Phi(B) &\le \sum_{i=1}^{\infty} P(S_{(i+1)[B/\xi]} \ge i[B/\xi]c + B) \\
&\le \sum_{i=1}^{\infty} E[\exp(\theta_i\{S_{(i+1)[B/\xi]} - i[B/\xi]c - B\})] \\
&= \sum_{i=1}^{\infty} \exp\left(-i[B/\xi]\left[\theta_i c - \frac{1}{i[B/\xi]} \log E \exp(\theta_i S_{(i+1)[B/\xi]}) + \frac{B\theta_i}{i[B/\xi]}\right]\right) \\
&= \sum_{i=1}^{\infty} \exp\left(-B\left[\frac{i}{\xi}\theta_i c + \theta_i - \frac{i+1}{\xi}\varphi(\theta_i) + \epsilon(\theta_i, B)\right]\right)
\end{aligned}$$

where $\epsilon(\theta_i, B) \to 0$ as $B \to \infty$. The second inequality is a Chernoff bound and the final equality follows from the assumption that $\varphi(\cdot)$ exists. Let us define

$$f(i,\theta) = \frac{i}{\xi}\theta c + \theta - \frac{i+1}{\xi}\varphi(\theta).$$

Suppose $f(j,\theta_j) = \inf_{i\ge1}\sup_{\theta>0} f(i,\theta)$. It is not hard to see that such a j and θ_j exist. For since $EX_1 < c$ we can choose some $\theta_0 > 0$ such that $\theta_0 c - \varphi(\theta_0) > 0$, and then note that $f(i,\theta_0) \to \infty$ as $i \to \infty$. Hence $j < \infty$ and since $f(j,\theta)$ is concave, θ_j is well-defined. We can also find k and small $\eta > 0$ such that $f(i,\theta_0) - f(j,\theta_j) > i\eta$ for all $i > k$. Taking $\theta_i = \theta_0$ for $i > k$, these facts are enough to show

$$\overline{\lim_{B\to\infty}} \frac{1}{B} \log \Phi(B) \le -\inf_{i\ge1}\sup_{\theta>0}\left[\frac{i}{\xi}\theta c + \theta - \frac{i+1}{\xi}\varphi(\theta)\right].$$

Since ξ is arbitrary we can let $\xi \to \infty$ and obtain

$$\overline{\lim_{B\to\infty}} \frac{1}{B} \log \Phi(B) \le -\inf_{\alpha>0}\sup_{\theta>0}\left[\frac{\theta c}{\alpha} + \theta - \frac{\varphi(\theta)}{\alpha}\right].$$

In the reverse direction, the probability that the buffer is full somewhere within any $i[B/\xi]$ consecutive epochs is at least $P(S_{i[B/\xi]} - i[B/\xi]c > B)$. Hence the proportion of epochs in which the buffer is full is at least $P(S_{i[B/\xi]}/i[B/\xi] - c > B/i[B/\xi])/i[B/\xi]$. Notice that upon taking the logarithm of this the denominator gives a term that is $o(B)$. So to this we can apply the Gärtner-Ellis theorem lower bound, and again use the fact that ξ is arbitrary to get

$$\varliminf_{B\to\infty} \log \frac{1}{B} \log \Phi(B) \geq - \inf_{\alpha>0} \sup_{\theta>0} \left[\frac{\theta c}{\alpha} + \theta - \frac{\varphi(\theta)}{\alpha} \right].$$

Thus $\lim_{B\to\infty}(1/B)\log\Phi(B)$ exists and is $\leq -\delta$ if and only if

$$\sup_{\alpha>0} \inf_{\theta>0} \left[\alpha\left(\frac{\delta}{\theta} - 1\right) + \frac{\varphi(\theta)}{\theta} \right] \leq c. \tag{4}$$

Suppose α^* and θ^* are the optimizing values on the left-hand side of (4). The condition for stationarity with respect to α^* is $\delta/\theta^* - 1 = 0$, and hence $\theta^* = \delta$ and this implies the left-hand side of (4) equals $\varphi(\delta)/\delta$.

If (4) holds with equality then $P(B) = \exp(-\delta B + o(B))$, and so when B is large it makes sense to consider δ small, in order to achieve a probability of overflow that is about $\exp(-\delta B)$. By the assumptions of the Gärtner-Ellis theorem, $\varphi(\cdot)$ is differentiable at 0 so using assumption A and expanding $\varphi(\delta)$, $\varphi(\delta)/\delta < c$ becomes

$$\mu + \frac{\delta\gamma}{2} + o(\delta) \leq c.$$

Note that because sources are assumed to be independent, $\mu = \sum_i N_i\mu_i$ and $\gamma = \sum_i N_i\gamma_i$. ■

4 Markov modulated fluids

4.1 A two-state fluid

Consider a two-state Markov modulated fluid source whose rate is controlled by a Markov process. The state of the Markov process at time t is denoted $X(t)$; it alternates between states 1 and 2 and has holding times in these states that are exponentially distributed with parameters λ and μ. The rate of the fluid source is 0 and a as the Markov process is in states 1 and 2 respectively. The process is in state 2 with probability $\lambda/(\lambda+\mu)$. We can compute the index of dispersion, γ, for this source, either by discretizing the process, or by proceeding directly in continuous time as follows. First we compute the autocovariance as

$$\gamma(t) = a^2 \left\{ \frac{\lambda}{\lambda+\mu} P_{22}(t) - \left(\frac{\lambda}{\lambda+\mu}\right)^2 \right\},$$

where the probability the process is in state 2 at time t given that it starts in this state at time 0, is

$$P_{22}(t) = \frac{\lambda}{\lambda+\mu} + \frac{\mu}{\lambda+\mu} e^{-(\lambda+\mu)t}$$

and so

$$\gamma(t) = \frac{\lambda\mu a^2}{(\lambda+\mu)^2} e^{-(\lambda+\mu)t}.$$

Thus

$$\gamma = 2\int_0^\infty \gamma(t)\,dt = \frac{2\lambda\mu a^2}{(\lambda+\mu)^3}$$

and

$$\beta = \frac{\lambda a}{\lambda + \mu} + \frac{\delta \lambda \mu a^2}{(\lambda + \mu)^3} + o(\delta). \tag{5}$$

This makes sense in various ways. It has the right dimensionality properties, scaling correctly in time and in cells. For small δ, (5) agrees with equation (29) in [7], where for N sources the probability of buffer overflow is determined to have asymptotic behaviour of $\exp(-BE)$, where

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

We can write $E > \delta$ as the quadratic condition, $(c/N)^2\delta + [\lambda + \mu - a\delta](c/N) - \lambda a > 0$. Thus we require $(c/N) > \beta^\dagger$, where $\beta^\dagger$ is the positive root of the quadratic, namely

$$\beta^\dagger = \left\{ -[\lambda + \mu - a\delta] + \sqrt{[\lambda + \mu - a\delta]^2 + 4\lambda a\delta} \right\} \Big/ 2\delta. \tag{6}$$

The above also corresponds to the bandwidth given by Gibbens and Hunt [10]. Expanding β in powers of δ gives (5).

The question arises as to the size of terms neglected in expanding (6). Consider, for example, a model of a voice call source in which, counting time in seconds and bits in 1000's, we take $a = 30$ Kbps, $\lambda = 2$, $\mu = 3$, $\delta = 10$. A buffer of 200 ATM cells, each of 54 bytes, of 8 bits, is about $B = 80$. So for $\delta = 1/8$, we find $\beta = 17.40$ and $\beta^\dagger = 17.47$. For $\lambda = 3$, $\mu = 2$, $\beta = 23.40$ and $\beta^\dagger = 22.29$. For $\lambda = 2.5$, $\mu = 2.5$, $\beta = 20.63$ and $\beta^\dagger = 20.00$. Thus the approximation of $\beta^\dagger$ by β is good. Note that $B = 80$ corresponds to buffering about 5 seconds of peak rate from a single source. On the basis of these calculations a 100 Mbps switch might carry 5700 voice calls of this model class.

4.2 A $M/M/\infty$-modulated fluid

De Veciana et. al. also consider the case of a fluid whose rate is a times the size of an $M/M/\infty$ queue. In fact, this result follows from that which they obtained for the two-state Markov modulated fluid. Simply let the number of such sources, N, tend to infinity and λ tend to zero such that $N\lambda$ is constant. The limit is the $M/M/\infty$ modulated process. The effective bandwidth is

$$\beta = \frac{a}{\mu} + \frac{a^2}{\mu^2}\delta + o(\delta).$$

4.3 A $n-$state Markov modulated fluid

Consider a fluid whose rates are controlled by a n-state Markov process, $\{X(t), t \geq 0\}$, with rate matrix Q and stationary probability vector π, satisfying $\pi^\top Q = 0$. The rate of the fluid when the process is in state i is a_i.

Theorem 3 *The effective bandwidth is*

$$\beta = a^\top \pi + a^\top \left[\pi\pi^\top - \Pi(Q + 1\pi^\top)^{-1} \right] a\delta + o(\delta),$$

where $\Pi = diag(\pi)$.

Proof We shall show how to compute the index of dispersion. Without loss of generality suppose Q has distinct eigenvalues $0, \lambda_1, \ldots, \lambda_n$. The covariance for time lag t is

$$\begin{aligned}
\gamma(t) &= \sum_{ij} \pi_i P_{ij}(t) a_i a_j - \sum_{ij} \pi_i \pi_j a_i a_j \\
&= a^\top \left[\Pi P(t) - \pi \pi^\top \right] a \\
&= a^\top \left[\Pi C \Lambda(t) C^{-1} - \pi \pi^\top \right] a \\
&= a^\top \left[\Pi C \bar{\Lambda}(t) C^{-1} \right] a,
\end{aligned}$$

where $P_{ij}(t) = P(X(t) = j | X(0) = i)$, and the columns of C and rows of C^{-1} are the right and left hand eigenvectors of Q respectively, $\Lambda(t) = \mathrm{diag}(1, \exp(\lambda_2 t), \ldots, \exp(\lambda_n t))$ and $\bar{\Lambda}(t) = \mathrm{diag}(0, \exp(\lambda_2 t), \ldots, \exp(\lambda_n t))$. Then, taking $A = \mathrm{diag}(1, 0, \ldots, 0)$, and any $s \neq 0$,

$$\begin{aligned}
\gamma &= 2 \int_{t=0}^{\infty} \gamma(t)\, dt \\
&= -2 a^\top \Pi C \mathrm{diag}(0, 1/\lambda_2, \ldots, \lambda_n) C^{-1} a \\
&= -2 a^\top \Pi C \left[(\bar{\Lambda}(0) + sA)^{-1} - (1/s) A \right] C^{-1} a \\
&= -2 a^\top \left[\Pi (Q + s 1 \pi^\top)^{-1} - (1/s) \pi \pi^\top \right] a
\end{aligned}$$

The theorem follows from taking $s = 1$. We have found it is it sometimes convenient to use other values of s. ■

5 Conclusions

We have shown that effective bandwidths may be associated with stationary sources. These bandwidths have the advantage that they are simple functions of the average rate of a source and its index of dispersion. The index of dispersion may be estimated from data, and evaluated for simple Markov modulated fluid models using theorem 3. In [4], we have presented, with Fouskas, experimental evidence that in the cases of Markov-modulated and autoregressive source models the use of these bandwidths can achieve a desired quality of service and a good utilization of the switch. There are many other issues relating to the use of effective bandwidths that require investigation. Bean [1] has studied a number of these, including on-line estimation of effective bandwidths, trunk reservation and prioritising calls with different quality of service requirements.

References

[1] N.G. Bean. *Statistical Multiplexing in Broadband Communications Networks.* PhD thesis, University of Cambridge, 1993.

[2] J.A. Bucklew. *Large Deviation Techniques in Decision, Simulation and Estimation.* John Wiley, New York, 1990.

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

[4] C. Courcoubetis, G. Fouskas, and R.R. Weber. On the performance of an effective bandwidths formula. International Teletraffic Conference, 1994.

[5] 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.*

[6] C. Courcoubetis and J. Walrand. Note on the effective bandwidth of ATM traffic at a buffer. unpublished manuscript, 1991.

[7] G. De Veciana, C. Olivier, and J. Walrand. Large deviations of birth death Markov fluids. *Probability in the Engineering and Informational Sciences*, **7**:237–256, 1993.

[8] G. De Veciana and J. Walrand. Effective bandwidths: call admission, traffic policing & filtering for ATM networks. Technical Report UNB/ERL M93/47, U.C. Berkeley, 1993.

[9] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications.* Jones and Bartlett, Boston, 1993.

[10] R. Gibbens and P. Hunt. Effective bandwidths for the multi-type uas channel. *QUESTA*, 1991.

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

[12] G Kesidis, J. Walrand, and C.S. Chang. Effective bandwidths for multiclass Markov fluids and other ATM sources. *IEEE Trans. Networking*, **1**:424–428, 1993.