Techniques for Adaptive Estimation of Effective Bandwidth in ATM Networks

R. A. Vesilo
School of Mathematics, Physics, Computing and Electronics
Macquarie University, NSW, 2109, Australia

V. Solo
Department of Statistics
Macquarie University, NSW, 2109, Australia

Abstract *This paper examines the effectiveness of adaptive algorithms for estimation of effective bandwidth in ATM networks. Large deviations results for constant capacity single server queues are used to derive a basic effective bandwidth result which is applied to a linear representation of the input process giving the effective bandwidth as a function of the cumulant generating function of the innovations, QoS, mean arrival rate and buffer size. The algorithm is tested by simulation using the least squares lattice algorithm for estimating the linear process coefficients for a range of input processes. The effectiveness of blocking and truncation of the input process on the algorithm is investigated.*

1 Introduction

ATM (asynchronous transfer mode) networks provide multiservice capability whereby traffic sources with different characteristics share network resources through the statistical multiplexing of fixed-size 53 byte cells with buffers used to absorb temporary overloads in traffic flow. To provide quality of service (QoS) guarantees to users, such as for cell loss and delay, switch input traffic flows need to controlled. This is a congestion control problem of which an important component is call admission where network resources, such as buffer space and link capacity, are allocated and the decision to accept a new call is made. The concept of effective bandwidth simplifies resource allocation because it does not include interaction between calls. It associates a bandwidth value with a connection, independent of other connections in a switch, that is greater than the mean rate but less than the peak rate of the connection, that can be used to estimate the link capacity required to support the connection at the required QoS, given the amount of available buffer space. Early work on effective bandwidths was by [6]. More recently by [7] on unbuffered systems and slotted batch models, [5] on the uniform arrival and service (UAS) model, [4] on continuous time Markov modulated fluid sources, [9] on discrete-time and Markov fluid sources and [2] on Gaussian traffic models.

The paper presents adaptive algorithms for estimating effective bandwidth allowing real-time estimates of resources to be made and is an extension of [10] where the underlying theory was introduced using large deviations results for single server queues to obtain an effective bandwidth formula that was applied to a linear representation of the input process. This paper develops the results of [10] to derive a recursive formula for effective bandwidth that involves computing the cumulant generating function of the innovations and the sum of the linear process coefficients. Algorithm effectiveness was tested by simulation using the least squares lattice algorithm to adaptively estimate the linear process coefficients and innovations with four different types of input process: Poisson, slotted batch, autoregressive (AR(1)) and ON-OFF, whose effective bandwidth could be calculated. Two enhancements to the algorithm, blocking and truncation of the input process, were examined to improve stability. The outline of the paper is as follows. Section 2 presents the underlying theory developed by [10]; Section 3 describes the effective bandwidth algorithm; Section 4 presents the algorithm for determining the linear process coefficients; Section 5 describes the traffic models used; and, Section 6 presents the simulation results.

2 Underlying theory

An ATM switch is modelled by a single server queue whose inputs are cells. Time is discretised into intervals of duration τ (not necessarily corresponding to an ATM cell slot) such that the capacity of the queue server is c cells per time interval. Define the cell input process so that δN_k cells arrive in time interval k corresponding to $((k-1)\tau, k\tau]$. The arrival counting process is given by $N_k = \sum_{r=1}^{k} \delta N_r$ = number of cells in $(0, k\tau]$. The queue length is given by the reflection map

$$Q_n \stackrel{\mathcal{D}}{=} \sup_{1\le k\le n} E_k; \qquad E_k = N_k - kc$$

where $\stackrel{\mathcal{D}}{=}$ means equality in distribution. To obtain cell loss probabilities, the probability of buffer overflow at buffer level b is given by $P(Q_n > b) = P(\sup_k E_k > b)$. A large deviations approximation for this is found using the Gartner-Ellis theorem ([4] and [10]), which gives

$$\begin{aligned} P(Q_n > b) &\simeq \sup_{1\le k\le n} e^{-kI_N(c+b/k)} \\ &= e^{-\inf_k kI_N(c+b/k)} \end{aligned}$$

where $I_N(x)$ is the rate function

$$I_N(x) = \sup_{\theta}(x\theta - K_N(\theta))$$

and $K_N(\theta)$ is the pseudo cumulant generating function

$$K_N(x) = \lim_{n\to\infty} n^{-1} \log E(e^{\theta \sum_{k=1}^{n} \delta N_k}).$$

0-7803-4201-1/97/$10.00 (c) 1997 IEEE

Suppose that it is desired to attain a QoS specifed by $P(Q_n > b) \leq e^{-\gamma} = e^{-b\delta}$ for large buffer sizes b, $\delta = \gamma/b$ is the cell loss slope. An expresson for effective bandwidth $\alpha(\delta)$ is obtained by determining the service capacity c which gives

$$\alpha(\delta) \leq c \iff P(Q_n > b) \leq e^{-b\delta} = e^{-\gamma}.$$

In fact $\alpha(\delta)$ is given by $\alpha(\delta) = K_N(\delta)/\delta$ (see [10] for details). If δN_k is stationary with mean m the following theorem is obtained [10].

Theorem. If δN_k is represented as a linear process

$$\delta N_k = \sum_{u=0}^{\infty} h_u \epsilon_{k-u} + m$$

where ϵ_k are independent identically distributed random variables with zero mean and cumulant generating function $K_\epsilon(\theta) = \log E(e^{\epsilon_1 \theta})$ then

(i) $n^{-1} \log E(e^{\theta \sum_{k=1}^{n} \delta N_k/n}) \rightarrow \theta M + K_\epsilon(\theta H)$ where $H = \sum_{u=0}^{\infty} h_u$. Thus the effective bandwidth is

$$\alpha(\delta) = K_N(\delta)\delta = m + K_\epsilon(\delta H)/\delta.$$

(ii) The cell loss slope δ corresponding to service rate c is $\delta = \beta/H$ where β is a scaled cell loss slope for ϵ_k found by solving

$$K_\epsilon(\beta)/\beta = (c - m)/H.$$

This theorem simplifies the calculation of effective bandwidth because it replaces the computation of the pseudo cumulant generating function of δN_k with computation of the cumulant generating function of the innovations.

3 Adaptive estimation of effective bandwidth

The mean m can be estimated adaptively using

$$m_k = m_{k-1} + \mu_m(\delta N_k - m_{k-1})$$

where μ_m is the step size. H is estimated by fitting an AR model of order p to the mean subtracted arrival process i.e. define $\delta\tilde{N}_k = \delta N_k - m$ then

$$\delta\tilde{N}_k = \sum_{r=1}^{p} a_r \delta\tilde{N}_{n-r}.$$

Using an adaptive algorithm of the type given in Section 4, adaptive estimators $\hat{a}_{r,k}, r = 1, \ldots, p$ of $a_r, r = 1, \ldots, p$ can be obtained. The estimator for H is then

$$H_k = \frac{1}{1 - \sum_{r=1}^{p} \hat{a}_{r,k}} \quad k = 1, 2, \ldots$$

To derive an estimator for $K_\epsilon(\beta)$ an estimator, M_k, for the moment generating function $M_\epsilon(\beta) = E(e^{\epsilon\beta})$ is first obtained:

$$M_k = M_{k-1} + \mu_K(e^{\epsilon_k\beta_k} - M_{k-1})$$

where $\beta_k = \delta H_k$, μ_K is the step size and the innovations are be estimated by

$$\epsilon_k = \delta\tilde{N}_k - \sum_{r=1}^{p} \hat{a}_{r,k}\delta\tilde{N}_{k-r}, \quad k = 1, 2, \ldots$$

Since $K_\epsilon = \log M_\epsilon$ and $dK_\epsilon = dM_\epsilon/M$ an estimator K_k for $K_\epsilon(\beta)$ is

$$\begin{aligned}
\delta K_k &= K_k - K_{k-1} = \frac{\delta M_k}{M_{k-1}} = \frac{M_k - M_{k-1}}{M_{k-1}} \\
&= \frac{\mu}{M_{k-1}}\left(e^{\epsilon_k\beta_k} - M_{k-1}\right) \\
&= \mu_K\left(\frac{e^{\epsilon_k\beta_k}}{M_{k-1}} - 1\right) \\
&= \mu_K\left(e^{\epsilon_k\beta_k - K_{k-1}} - 1\right).
\end{aligned}$$

An estimator for the effective bandwidth is then

$$\alpha_k = m_k + K_k H_k.$$

3.1 Enhancements to algorithm

The presence of ϵ_k in the exponent can cause large fluctuations and instability in the estimator α_k. μ_K can be adjusted to control these fluctuations and it must be kept small to reduce the impact of $e^{\epsilon_k\beta_k}$. However this means that the effective bandwidth estimates decay slowly when ϵ_k is small. To provide more flexibility in the estimator, but at the cost of delay in determining the estimator and more memory in its computation, a block estimate can be used. Divide the input sequence ϵ_k into blocks of size B. A block estimate for the moment generating function is obtained by averaging $e^{\beta_k\epsilon_k}$ over a block. The updated estimates are obtained at the end of each block:

$$M_k^B = M_{k-1}^B + \mu_K\left(\frac{1}{B}\sum_{j=1}^{B} e^{\beta_k\epsilon_{k(B-1)+j}} - M_{k-1}^B\right).$$

Following the non-block case, define

$$\delta K_k^B = K_k^B - K_{k-1}^B = \frac{\delta M_k^B}{M_{k-1}^B} = \frac{M_k^B - M_{k-1}^B}{M_{k-1}^B}$$

to give

$$\delta K_k^B = \mu\left(\frac{1}{B}\sum_{j=1}^{B} e^{\epsilon_{k(B-1)+j}\beta_k - K_{k-1}^B} - 1\right).$$

The block effective bandwidth is then

$$\alpha_k^B = m_{kB} + K_k^B H_{kB}.$$

In this estimator, B acts as a smoothing parameter that can independently limit the effect of large values of ϵ_k while μ_K can be used to control the ramp-down of the estimator when ϵ_k is small.

0-7803-4201-1/97/$10.00 (c) 1997 IEEE

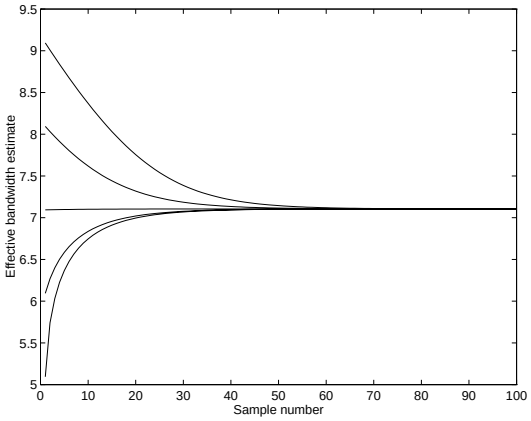

Figure 1: Performance of averaged system

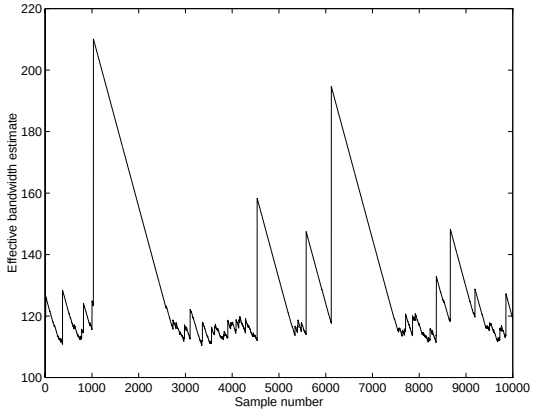

Figure 2: Effective bandwidth of Poisson process ($\mu = 0.02$, $B = 1$)

3.2 Truncation

With blocking there is still the possibility of large values ϵ_k causing instability. One method of limiting this is to apply a threshold T and truncate the input process. In an ATM network truncation of the input process will occur because of peak rate traffic shaping. The effect of truncation is examined in more detail in Section 6.

4 Adaptive AR estimation

Recursive algorithms that can be used to estimate the weights $\hat{a}_{r,k}$ in the linear representation include the least mean squares (LMS), the gradient adaptive lattice (GAL), the recursive least squares (RLS) and the least squares (LSL) algorithms [1]. The one which was used in the simulation experiments in Section 6 was the LSL algorithm and is described in Chapter 10 of [1]. Compared to the other algorithms it gives much faster convergence since it uses conditional predication rather than unconditional prediction to update the estimated coefficients. The algorithm outputs forward and backward reflection coefficients and these are transformed to AR coefficients. The reader is referred to [1] for a detailed description of the algorithm.

5 Traffic models

This section presents four traffic models: Poisson, slotted-batch, AR(1) and ON-OFF, which will be used to test the adaptive effective bandwidth algorithms.

Although traffic processes in ATM are rarely Poisson, the Poisson process can be easily analysed to give results that can be used to test the algorithms. For the Poisson process, $\epsilon_k = \delta N_k - m$ where δN_k are i.i.d Poisson random variables with mean m and $h_0 = 1$ and $h_u = 0$ $(u > 0)$, giving $H = \sum_{u=0}^{\infty} h_u = 1$. Since the cumulant generating function of a Poisson random variable with mean m is $\log E(e^{\theta \delta N_k}) = m(e^{\theta} - 1)$ this gives,

$$\begin{aligned} K_\epsilon(\theta) &= \log E(e^{\theta(\delta N_k - m)}) \\ &= m(e^{\theta} - 1 - \theta) \end{aligned}$$

and the effective bandwidth is

$$\alpha(\delta) = \frac{m}{\beta}\left(e^{\beta} - 1\right).$$

Kelly [8] examines a simple slotted batch model with mean rate m and peak rate h, where the number of cells in a slot are i.i.d with $P(\delta N_k = 0) = 1 - m/h$ and $P(\delta N_k = h) = m/h$. In this case the effective bandwidth is

$$\alpha(\delta) = \delta^{-1} \log\left[1 - \frac{m}{h}(e^{\delta h} - 1)\right].$$

An AR(1) model, which can be used to model videophone traffic, has been examined by [2] who derived an effective bandwidth result using the index of dispersion of a Gaussian linear process. Since H is related to the index of dispersion, the same result as [2] can be obtained as follows. Suppose that the AR(1) process is given by

$$\delta \tilde{N}_k = a\delta \tilde{N}_{k-1} + \epsilon_k$$

then $h_u = a^u$ and $H = 1/(1-a)$. For innovations ϵ_k which are Gaussian with mean 0 and variance σ^2 then $K_\epsilon(\theta) = \sigma^2\theta^2/2$ giving

$$\alpha(\delta) = m + \frac{\sigma^2 \delta}{2(1-a)^2}.$$

Two types of ON-OFF sources are examined: an Interrupted Poisson Process (IPP) and an Interrupted Deterministic Process (IDP). To obtain an exact form for the effective bandwidth, K_ϵ is needed. However, in neither of these cases is an exact form simple to obtain. As an approximation the fluid model results of [4] are used where an ON-OFF source can be considered as a special case of a modelled as Markov modulated fluid source. The effective bandwidth is given by

$$\begin{aligned} \alpha(\delta) &= \frac{1}{2\delta}\left[\lambda_1\delta + q_0 + q_1\right] \\ &\quad -\frac{1}{2\delta}\sqrt{\left\{\lambda_1\delta + q_0 + q_1\right\}^2 - 4q_0\lambda_1\delta}. \end{aligned}$$

where the mean ON time $= 1/q_1$, the mean OFF time $= 1/q_0$, and rate in the ON state is λ_1.

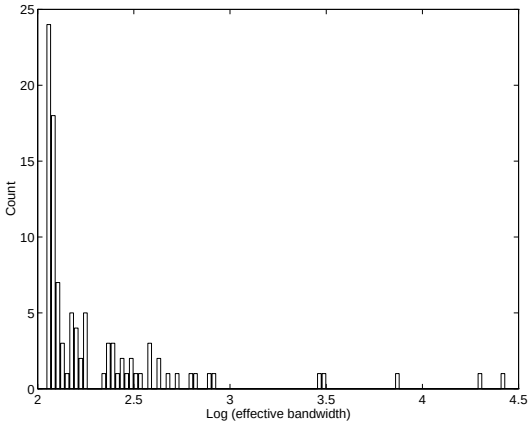

Figure 3: Log histogram of Poisson process effective bandwidth ($\mu = 0.02$, $B = 1$)

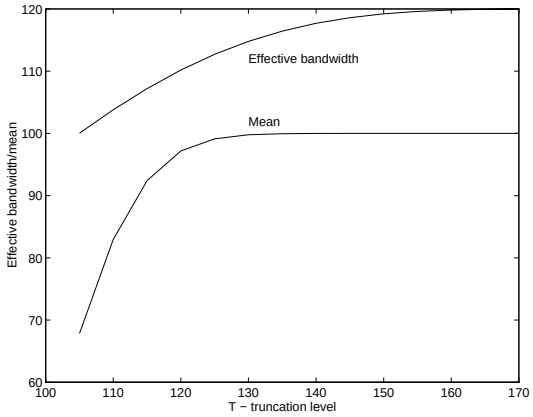

Figure 4: Mean and effective bandwidth of Truncated process

6 Simulation results

Initially, the Poisson process was used to test the viability of the algorithms, then the other traffic models were examined and finally the effect of truncation was examined. The mean rate chosen was $m = 100\,\mathrm{s}^{-1}$. The initial Poisson process simulations used the theoretical values for $H(=1)$ and m in the algorithm. The loss probability chosen was 10^{-4} and the buffer size of $b = 26.0$ was chosen to given an effective bandwidth of $\alpha = 120\,\mathrm{s}^{-1}$. This gave a value of $\delta = 0.3547$ and $\beta = 0.3547$. The stability of the algorithm can be studied via the stability of the averaged system [10]

$$\delta K_k = \mu_K \left(e^{K_\epsilon(\beta) - K_{k-1}} - 1 \right).$$

It can be shown that this is globally stable if $0 < \mu_K < 1$. Behaviour of the averaged system for $\mu_K = 0.02$ for different starting points is shown in Figure 1, veryifying convergence of the algorithm. To test the effective bandwidth algorithm two values of $\mu_K(= 0.02, 0.1)$ and two block values $B(= 1, 10)$ were used. The selection of step sizes and block sizes in adaptive algorithms is a topic of current research which needs to be addressed further before these types of algorithm can be fully utilised, and the values chosen were on the basis of preliminary simulation studies. However, [3] gives a discussion of block sizes for a non-adaptive algorithm.

A trace of a typical simulation is shown in Figure 2 which shows the influence of larger values in ϵ_k causing large jumps in the estimate and the slow steady decay when smaller values are present. To examine this variability, 100 simulations each were conducted for the four possible combinations of μ_K and B. A typical histogram of the estimate after 5000 samples for the case $\mu_K = 0.02$ and $B = 1$ is shown in Figure 3 and shows a general clustering around the expected value of 120 but also the presence significant skewness and outliers. The descriptive statistics for the 100 trials are shown in Table 1. The confidence intervals are based on a normal assumption but because of skewness are only a guide. The results confirm the increase in variability by increasing μ_K and the reduction in variability by increasing B.

6.1 Results with estimated mean and H

Table 2 shows the results of 100 experiments each when the mean is known and H is estimated, the mean is estimated and H is known and when both the mean and H are estimated. In estimating H is was found that although convergence of the LSL was fast (the LMS, GAL, RLS algorithms were also implemented) the estimate of H could vary significantly early on. The results presented include a warm up priod of 200 samples before the mean and/or H values were used to estimate the effective bandwidth. The value of 200 was chosen as a conservative estimate for ease of testing. The results are roughly consistent with the simulations using the theoretical values, taking into consideration the large variability, although the cases where the mean is estimated give better results than where H alone is estimated.

6.2 Non-Poisson models

The slotted batch, AR(1) ($a = 0.8$), IPP and IDP models were examined using adaptive estimation of H but not m. Model parameters in each case were set to give the same theoretical effective bandwidth as the Poisson process. The results support the overall effectiveness of the algorithm but, as in the Poisson case, large values of ϵ_k can case significant variability. Bandwidth limited models (slotted/batch, IDP) have less variability. The AR(1) model shows significant variablity in the estimate even though the innovations are smaller than the Poisson case because of the magnifying effect of $H(= 5)$.

6.3 Truncation

The affect of truncation on a Poisson process was examined through computation of the mean and effective bandwidth of a truncated Poisson process for different thresholds (Figure 4). The mean converges to the untruncated value faster than the effective bandwidth confirming the influence of large values. Simulation results in Table 4 for a threshold of 130 indicate the improvement in stability achieved throug truncation.

0-7803-4201-1/97/$10.00 (c) 1997 IEEE

Case	Value	Min	Max	Median
$\mu_K = 0.02, B = 1$	789.4 (128.5, 1450.21)	111.1	26932	134
$\mu_K = 0.1, B = 1$	5231.3 (201.7, 10260.9)	107.7	241689	301.9
$\mu_K = 0.02, B = 10$	134.8 (122.8, 146.8)	113.9	600.6	118.7
$\mu_K = 0.1, B = 10$	493.3 (57.2, 929.3)	111.8	21689	127. 9

Table 1. Poisson process results

Case	Value	Min	Max	Median
H estimated only	458.1 (268.7, 647.45)	137.6	5518.2	137.6
Mean estimated only	320.7 (210.2, 431.2)	109.5	4448.1	147.1
H and mean estimated	126.3 (120.9, 131.7)	109.5	278.9	118.3

Table 2. Poisson process with H and/or mean estimated ($\mu_K = 0.02$ and $B = 1$)

Case	Value	Min	Max	Median
Slotted batch	120.0 (119.6, 120.0)	119.6	120.3	120.0
AR(1)	2160.7 (157.9, 4163.6)	122.3	100633	247.3
IPP	827.955 (75.1, 1580.8)	126.0	36931	170.7
IDP	122.0 (120.1, 124.0)	114.3	190.6	119.7

Table 3. Non-Poisson process results

Case	Value	Min	Max	Median
$\mu_K = 0.02, B = 1$	120.4 (118.5, 122.3)	110.1	171.4	117.6
$\mu_K = 0.1, B = 1$	231.6 (155.6, 307.6)	106.1	3772.3	140.1
$\mu_K = 0.02, B = 10$	115.8 (115.6, 115.9)	113.5	117.9	115.7
$\mu_K = 0.1, B = 10$	116.7 (116.2, 117.2)	112.4	125.0	116.3

Table 4. Truncated Poisson process results

7 Conclusion

This paper examined adaptive algorithms for effective bandwidth estimation. The main difficulty is variability in the estimate due to the exponential term in the algorithm. It was found that blocking only partially alleviates this problem but truncation at a suitablly low enough level can have a significant effect. The selection of suitable thresholds, step sizes and block sizes is the subject of ongoing work.

References

[1] Alexander, S. T., "Adaptive signal processing thoery and applications", Springer-Verlag, New York, 1986.

[2] Courcoubetis, C., Fouskas, G. and Weber, R., "On the performance of an effective bandwidths formula", Proceedings of ITC-14, 1994, pp. 201 - 212.

[3] Duffield, N. G., Lewis, J. T., O'Connell, N. O., Russell, R. and Toomey, F., "Entropy of ATM traffic streams: A tool for estimating QoS parameters", J. Selected Areas in Communications, Vol. 13, No. 6, August 1995 pp. 981 - 990.

[4] Elwalid, A. I. and Mitra, D., "Effective bandwidth of general Markovian traffic sources and admission control of high speed networks", IEEE/ACM Trans. on Networking, Vol. 1 No. 3, June 1993, pp. 329 - 343.

[5] Gibbens, R. J. and Hunt, P. J., "Effective bandwidths for the multi-type UAS channel", Queueing Systems, 9, 1991, 17 - 28.

[6] Hui, J. Y., "Resource allocation for broadband networks", IEEE Journal on Selected Areas in Communications, Vol. 6, No. 9, December 1988, pp. 1598 - 1608.

[7] Kelly, F. P., "Effective badnwidths at multi-class queues", Queueing Systems, 9, 1991, pp. 5-16.

[8] Kelly, F. P., "On tariffs and admission control for multiservice networks", Operations Research Letters, 15, 1994, pp. 1 - 9.

[9] Kesidis, G., Walrand. J. and Chang, C.-S., "Effective bandwidths for multiclass markov fluids and other ATM sources", IEEE/ACM Trans. on Networking, Vol. 1 No. 4, August 1993, pp. 424 - 428.

[10] Solo, V., "Adaptive estimation of effective bandwidth in ATM networks", Proc. 35th IEEE CDC, 1996.