Investigations of the performance of a measurement-based Connection Admission Control algorithm.

Brian McGurk, Cormac Walsh[1]

Abstract

In this paper we present an algorithm for making connection admission decisions in ATM networks, using measurements made on existing connections and the declared parameters of the new connections. Our scheme makes use of the *shape-function*, a concept developed by Botvich and Duffield [28], which arises in the application of Large Deviation theory to queueing systems. By estimating the shape-function of the existing connections, we can make predictions about the effect on the network of accepting new connections. Using real traffic collected from a network, we compare the performance of this CAC scheme with that of the Mosquito [33] algorithm which is based on estimation of effective bandwidths.

1 Introduction

Asynchronous Transfer Mode (ATM) allows for statistical multiplexing of traffic from many different applications, having widely differing characteristics. This is advantageous from the network's point of view as if affords the opportunity for dramatically increased utilisation of resources: more applications can be allowed to make use of the network than peak rate allocation would suggest if the traffic is buffered at switches and multiplexing points during overload periods. The choice of the buffer size places a hard bound on the queueing delay and so the problem for the network is in deciding how many connections can safely be accepted while keeping the cell-loss ratios sufficiently low.

Connection Admission Control (CAC) is concerned with trying to assess the impact on the network of accepting a new connection, a problem which has been discussed extensively in the literature [2, 3, 4, 5, 14, 21]. One approach which has received recent attention uses the notion of the *effective bandwidth* of a traffic source [1, 6, 7, 9, 10, 14, 15, 16, 17, 18]. This approach is based on the theory of Large Deviations, a probabilistic theory of rare events, which, when applied to queueing systems, can help to quantify the intuitive notion of bandwidth requirement. The term 'effective bandwidth' refers to a particular function which provides a conservative estimate of the bandwidth requirement of a source. This function depends on the QoS constraints in a simple manner and on the statistical properties of the traffic in a complex manner; if the sources are independent, these functions are additive.

However, the additive nature of the effective bandwidth function means that it fails to reflect economies of scale arising from statistical multiplexing. This is because it is based on large buffer asymptotics. An alternative approach [34], based on the asymptotics associated with a large number of connections, involves estimation of the *shape-function* [28] for the multiplexed traffic. In the CAC algorithm introduced in [33], the decision to accept or reject a proposed connection is based on its declared parameters and on-line estimation of the bandwidth requirement of existing connections. For the on-line estimation of bandwidth requirement, a variety of estimators may be used. In this paper, we compare the effectiveness of two estimators: the shape-function estimator [34] and the Mosquito estimator [33].

The practical estimation of bandwidth requirements is a difficult problem, as it depends in a complex way on the statistical properties of the traffic. The problem is typically approached in two ways. One approach [11, 12, 13] is to assume a parametric model of the traffic and to fit parameters for the connection in question. This parameter fitting can be done based on

[1]Dublin Institute for Advanced Studies, 10 Burlington Road, Dublin 4, Ireland. phone: +353 1 6680748; fax: +353 1 6680561; email: walsh@stp.dias.ie

information declared by the connection when it requests admission, or measurements made on the traffic generated by the connection, or a combination of both. Once the detailed model is completed, the estimate can be calculated. However, there are problems with this approach. Firstly, unless on-line measurement is employed, the application is required to deliver a detailed self-characterisation before it has transmitted any traffic. Furthermore, given such a characterisation, the network still has to fit parameters to a model which adequately describes the traffic source. This may prove a difficult problem; the solution of which contains redundant information if what is actually required is just a knowledge of the bandwidth requirement. Finally, it is impossible to tell what types of traffic it may be necessary to transmit in the future, and it should not be required that each new traffic type be submitted to a complex modeling process in advance of transmission.

An alternative approach [8, 14, 15, 21, 22, 23] is to attempt to measure the bandwidth requirement more directly. This avoids the problem of requiring new traffic types to specify a parameterised model in advance and removes the estimation of redundant information. Perhaps most importantly, this approach requires very little declared information on the part of the application; in the present scheme only declaration of the peak-rate is necessary. However, whenever additional parameters are declared they can be incorporated in the estimation of the bandwidth requirement, thereby offering increased efficiency. Hence, this approach can use a model estimated from declared parameters, if one is available, as an initial estimate of the bandwidth requirement, which is then refined by on-line measurement. However, in the absence of such a model, a more conservative initial estimate can easily be made from the declared peak-rate.

2 Theoretical Framework

Our central concern is the loss of cells due to overflow at a buffer. Consider a multiplex of N ATM streams arriving at a buffer which has finite storage capacity B. Cells are removed from the buffer at fixed rate S, the *line-rate*. Each traffic stream has a finite duration, as might be expected for calls of finite length. Associated with each sample of traffic is a *cell-loss ratio* between zero and one (that is, the ratio of cells lost to those that arrive); we denote the cell-loss ratio for a multiplex of N lines with a buffer-size B and a line-rate S by $CLR(N, B, S)$. Experience with a wide variety of traffic-sources shows that the the logarithm of $CLR(N, bN, sN)$ is asymptotically linear in the number of sources N if the line-rate per source s and buffer size per source b are fixed. A typical example is shown in Figure 1 which plots, for a set of motion JPEG sources on the Fairisle ATM network at Cambridge [25], the logarithm of the observed cell-loss ratio against number of sources N when the buffer size and line-rate are scaled appropriately. This demonstrates the multiplexing gain available in shared resource systems due to the statistical properties of the individual traffic streams. For example, if one doubles the number of (identical) sources to be multiplexed, one need not, generally, double the rate and buffer size in order to maintain the same CLR.

The general features of a plot of the logarithm of cell-loss ratio against number of sources are explained by queuing-theory. We model the arrival streams as stationary stochastic processes $\{A_t^n\}$, the *arrivals processes*; here A_t^n denotes the total number of cells which have arrived up to time t from source n. The following scaling behaviour of queue-tail probabilities has been shown to hold for a very general class of traffic models [28, 29, 30]. When N sources, satisfying certain assumptions detailed below, feed a buffer of size Nb which is being served at rate Ns the proportion of cells lost will satisfy the logarithmic asymptotic

$$\log CLR(N, Nb, Ns) \sim -NI(b), \qquad \text{as } N \to \infty, \tag{1}$$

where the *shape function* $I(b)$ depends on s and the detailed traffic characteristics. Although equation (1) describes the behaviour of the CLR as N goes to infinity, we would expect that $\log \mathrm{CLR}(N, Nb, Ns) \approx -NI(b)$ for N large, but finite. One useful feature of (1) is that it does

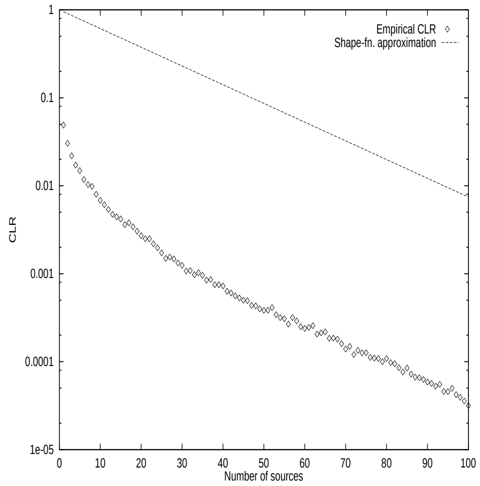

Figure 1: Empirical loss probabilities for samples of JPEG coded video

not require that b, the buffer allocation per source, be large. Thus it can be used to describe cell-level, as well as burst-level, queueing behaviour. In this, it is distinguished from the large body of results about the asymptotic behaviour of tail probabilities for large b, and the consequent effective bandwidth approximation. Another useful feature of (1) is that it holds for a wider class of traffic than the corresponding result for large b asymptotics. It does not require that the traffic be *mixing* and, thus, the estimation based on it, described in this paper, should be valid for long-range dependent traffic [30, 31].

The main condition we require of the multiplex of arrivals processes is that the limit of the finite-time *cumulant generating function* (CGF) per source exists for each time-scale t as N goes to infinity:

$$\lambda_t(\theta) := \lim_{N\to\infty} \frac{1}{Nt} \log E e^{\theta \sum_{n=1}^{N} A_t^n} \tag{2}$$

The finite-time CGFs are related to the *effective bandwidth* of the sources: $\alpha(\theta) = \theta^{-1} \lim_{t\to\infty} \lambda_t(\theta)$. We shall discuss effective bandwidths more fully in section (5).

The assumption above is satisfied by i.i.d. superpositions and also by heterogeneous superpositions where the proportion of each type of source is held constant. In the case of independent heterogeneous superpositions of J types of sources indexed by $j \in \{1,\dots,J\}$, we have that $\lambda_t(\theta) = \sum p_j \lambda_t^{(j)}(\theta)$ where p_j is the proportion of sources of type j and $\lambda_t^{(j)}$ are the finite-time CGFs of a source of type j defined analogously to (2) as

$$\lambda_t^{(j)}(\theta) := \frac{1}{t} \log E e^{\theta A_t^{(j)}}. \tag{3}$$

Thus, when we perform estimation based on (1) we need not assume that the traffic sources are identical.

Under the above assumption, (1) holds with the shape function I given by

$$I(b,s) = \min_{t\geq 0} (t\lambda_t)^*(b+st) \tag{4}$$

where $f^*(x) := \max_{y\in R}\{xy - f(y)\}$ is the *Legendre-Fenchel transform* of the function f; see [28, 29, 30]. The time $\tau = \arg\min_{t\geq 0}(t\lambda_t)^*(b+st)$ at which the minimum above is attained is called the *critical time-scale* and may be interpreted as the most likely time-scale on which the buffer overflows. The theoretical result above offers an explanation of the observed asymptotic behaviour of the logarithm of the cell-loss ratio as a function of the number of sources and relates the slope of its linear asymptote to the CGF of a stochastic process representing the traffic.

The bandwidth requirement of a sample of traffic is defined to be the minimum line-rate at which a *target cell-loss ratio* c is not exceeded in a buffer of storage capacity B:

$$\mathrm{BWR}(N,B,c) := \min\{S : CLR(N,B,S) \leq c\}. \tag{5}$$

Notice that this is an operational definition which does not involve any statistical theory; for a given trace, it can be determined empirically by trial and error.

We are interested in *estimating* the bandwidth requirement of the multiplex. Suppose we have some estimate $\hat{I}$ of the shape-function I. Then we may use this estimate to obtain an estimate of the CLR for any value of S:

$$\widehat{\mathrm{CLR}}(N, B, S) := e^{-N\hat{I}(B/N,S/N)}$$

This leads to a natural estimate of the bandwidth requirement: we adjust S until our estimated CLR just matches the target CLR:

$$\widehat{\mathrm{BWR}} := \min\{S : e^{-N\hat{I}(B/N,S/N)} \leq c\}$$

In the next section we shall explain how to estimate the shape-function.

3 Estimating The Shape-Function

Given a sample realization $\{X_1, X_2, \ldots\}$ of a traffic stream we may estimate the finite-time CGFs of its source as follows. First form all blocks of length t:

$$\tilde{X}_1 := \sum_{i=1}^{t} X_i, \qquad \tilde{X}_2 := \sum_{i=2}^{t+1} X_i, \qquad \ldots$$

Assuming stationarity of the arrival process, we use these overlapping blocks to get an estimate of λ_t by replacing the expectation in (3) with an empirical mean:

$$\hat{\lambda}_t(\theta) := \frac{1}{t} \log \frac{1}{K} \sum_{k=1}^{K} e^{\theta \tilde{X}_k}$$

where K is the number of blocks formed. We assume that the sources are independent and so we may combine the estimates to form an estimate of λ_t:

$$\hat{\lambda}_t(\theta) := \frac{1}{N} \sum_{j=1}^{N} \hat{\lambda}_t^{(j)}$$

Then for each t we merely perform the minimisation and Legendre transform, mirroring (4), in order to form the estimate

$$\hat{I}(b) := \min_{t \geq 0} (t\hat{\lambda}_t)^*(b + st)$$

It is worth distinguishing this procedure from that of section 5 where, for some (large) T, $\hat{\lambda}_T$ will be used to estimate the limiting CGF, $\lambda(\theta) = \lim_{t\to\infty} \lambda_t(\theta)$, and hence estimate the effective bandwidth by $\hat{\alpha}(\theta) := \theta^{-1}\hat{\lambda}_T(\theta)$. There is a tradeoff here in choosing a value for T: too small a value and $\lambda_T(\theta)$ will not be close enough to $\lambda(\theta)$, too large a value and the variance of the estimator will be large (see [32]). Also, it is difficult to automate the choice of block size. However, the shape-function estimator does not suffer from this problem: the value of t where the minimum is empirically observed to occur provides an estimate of $\tau = \arg\min_{t\geq 0}(t\lambda_t)^*(b+st)$ so that this method automatically picks out the time scale relevant to buffer overflow.

4 The CAC Algorithm

In this section we summarise the description, given in [33], of a practical CAC algorithm which makes use of on-line estimation of bandwidth requirement. A measurement based CAC algorithm will, in general, need to combine measurements on the current multiplex with declared parameters from the new call request in order to judge whether there is sufficient bandwidth

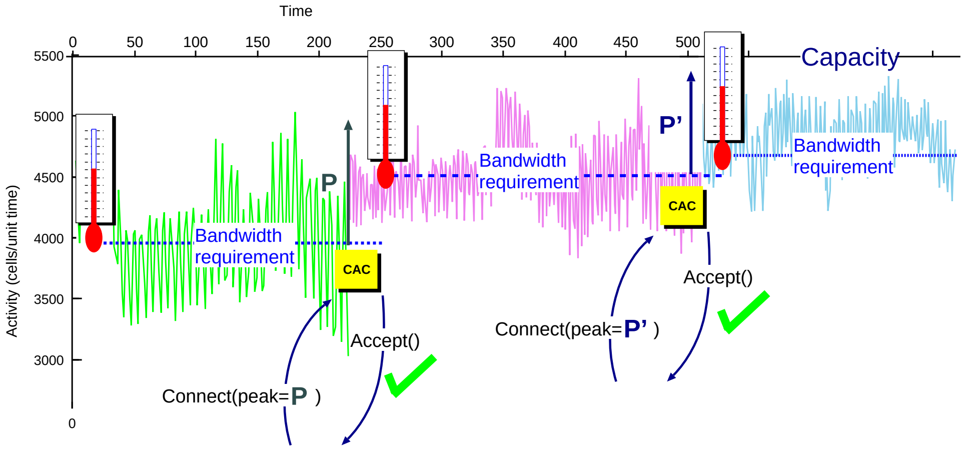

Figure 2: Operation of the CAC algorithm

available to satisfy the QoS requirements. The task of measuring the bandwidth requirement of the current traffic mix is performed by the *estimator*. In this paper, we compare the effectiveness of several estimators, making use for the first time of the shape-function estimator:

$$\widehat{\mathrm{BWR}}(Nb, c) := \min\{sN : e^{-NI(b,s)} \leq c\},$$

where c is the target CLR. For the new call there is no traffic record available and so the CAC algorithm must base its estimate on the the call's declared parameters. These are passed to the *predictor*. In this paper we consider a predictor which requires a new call request to declare only its peak rate, however it is possible to use further information such as the GCRA parameters defined by the ATM forum [20].

Figure 2 shows the behaviour of the CAC algorithm. Given a current multiplex of calls, the system estimates the bandwidth requirement of the multiplex using its chosen estimator (depicted here by a thermometer) and the declared parameters of the connection request. If the estimated bandwidth requirement is less than the link capacity, then the connection can be accepted without violating the QoS of any calls. As soon as the new call commences, the estimator uses measurements of the new multiplex to revise its estimate of the current bandwidth requirement. When the next call attempt arrives, the procedure is repeated, as shown. If a new call attempt arrives before the algorithm has developed an accurate estimate of the new bandwidth requirement the algorithm acts *conservatively*. It uses the most recent stable estimate of the bandwidth requirement, plus the sum of the peak rates of all subsequently admitted calls.

5 The Mosquito Estimator

In this section we review the *Mosquito* Estimator proposed in [33]. Unlike the shape-function estimator, which is based on the large N asymptotics of the system, the Mosquito estimator is based on the large buffer asymptotics. The crucial observation is that for stationary and mixing arrivals processes the loss ratio decays exponentially with buffer size:

$$\mathrm{CLR}(B, S) \sim e^{-B\delta(S)}$$

Since the number of sources being multiplexed is now constant we have dropped the reference to N. The decay rate δ is determined by the line-rate and by the CGF of the multiplex

$$\delta(S) = \max\{\theta : \lambda(\theta) \leq S\theta\}$$

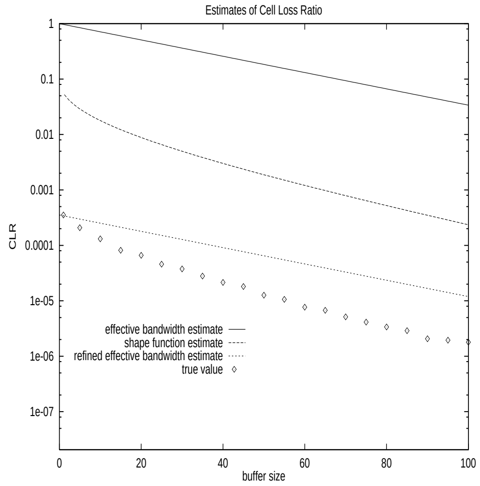

Figure 3: Estimates of the CLR for Markovian traffic using three different approximations plotted against B for fixed N and S.

λ may be estimated by choosing a block length T large enough so that the aggregate arrivals $A_T = \sum_{n=1}^{N} A_T^n$ of the multiplex are approximately independent; in this case

$$\lambda(\theta) := \lim_{t\to\infty} \lambda_t(\theta) \approx \lambda_T(\theta) := \frac{1}{T}\log E e^{\theta A_T}$$

As before, the finite-time CGF λ_T is estimated using the empirical distribution of the blocked arrivals

$$\hat{\lambda}_T(\theta) := \frac{1}{T}\log\frac{1}{K}\sum_{i=1}^{K} e^{\theta \tilde{X}_i}$$

where $\tilde{X}_i$ are the activities in each block. Using the estimate of λ we can estimate the bandwidth requirement; because of the properties of λ this takes the particularly simple form

$$\widehat{\mathrm{BWR}}(B) := \min\{S : -\hat{\delta}(S)B \le \log c\} = \frac{\lambda(\delta^*)}{\delta^*}$$

where $\delta^* = -\log c/B$. The function $\lambda(\theta)/\theta$ is the *effective bandwidth* function of the arriving traffic stream and the estimator based on it is called the *effective bandwidth estimator*.

In practice the effective bandwidth estimator has been found to be too pessimistic. This is because, for finite buffer size, the approximation $\mathrm{CLR}(b) \approx \exp(-B\delta)$ is not exact. An improvement can be made by introducing a pre-factor $e^{-\mu(S)}$ so that the estimate of the bandwidth requirement becomes

$$\widehat{\mathrm{BWR}}(b) := \min\{S : -\hat{\mu}(S) - \hat{\delta}(S)B \le \log c\}.$$

δ is estimated as before and μ is estimated by simulating the system and observing the cell loss in a buffer of size zero.

For a simple Markovian model, figure 3 shows the simple and refined effective bandwidth approximations, and also the shape-function approximation described in the previous section. These are compared with the true value of the CLR produced from simulations.

As well as using the refined effective bandwidth approximation the Mosquito estimator also uses knowledge of the peak rate to make the estimation of λ more robust – see [33] for details.

6 Simulation Results

This section presents the results of simulation experiments using the CAC algorithms of Sections (4) and (5). The aim of the experiments was to evaluate the performance of our approach with respect to several criteria which we discuss below.

Firstly, we are interested in the performance of each algorithm in terms of the resource utilisation it achieves. A *pessimistic* CAC algorithm which allocates resources using the declared peak rate of each source can guarantee that the loss constraints will always be met. Our algorithms, however, attempt to increase the link utilisation by admitting as many calls as possible whilst still maintaining the QoS guarantee. We thus need to compare our approach with the pessimistic system, and with a system which is *optimal* in the sense that the CAC algorithm is assumed to have complete knowledge of the statistical properties of every connection requesting admission. The optimal CAC should achieve maximal link utilisation while ensuring that QoS constraints are met. Secondly, we are interested in the effectiveness of each CAC algorithm in terms of its ability to guarantee the QoS constraints of the traffic.

Simulation Model. In each of our simulations we model a single output buffer and transmission link of an ATM switch. The link speed used is 100 Mb/s, which corresponds to the TAXI transmission rate for the Fairisle ATM network at Cambridge.

The traffic source we use is one which has been widely used in performance studies of ATM systems in the literature, namely a trace of the activity of the Star Wars movie. The Star Wars data set was produced by Garrett and Vetterli at Bellcore, and has been studied in some detail, for example in [24, 26] and [17]. It comprises the information content in bytes per frame for about 2 hours of the film, as transmitted by a DCT based codec similar to JPEG. The byte count per frame is broken down into "slices" of 16 video lines. With 30 slices per frame and 24 frames per second, one slice represents the information transmitted in about 1.4 milliseconds. The Star Wars traffic has been shown to exhibit signs of long-range dependent behaviour, making it potentially difficult for a measurement based CAC to cope with.

The Star Wars traces were constructed by encoding each slice as a single AAL5 PDU which was transmitted (and traced) over Fairisle; cells from each slice were transmitted at a CBR rate equal to the slice rate, reducing the peak rate from the line rate to about 24.2 Mb/s for the worst slice. The mean rate is about 5.3 Mb/s. The buffer is sized according to the requirement that no frame would be delayed more than a frame time; this results in a maximum buffer size of 500 cells. We used a CLR constraint of 10^{-4} for all of the results presented in this paper.

Call Model. We study a scenario in which calls of a particular traffic type arrive according to an exponential inter-arrival time distribution, an assumption which appears to be well founded [27]. In the absence of real-world data we have used call lengths which are exponentially distributed. Calls arrive at a high (Poisson) rate with mean 5 calls/s. Blocked calls are lost, but the high arrival rate means that the system is continually faced with new call attempts. We thus expect the system to remain close to maximum utilisation.

Calls have an exponentially distributed length with a mean length of 60 seconds. Each accepted call transmits an independent *trace* which is derived by randomly selecting a start point in the Star Wars movie. Calls are not correlated in any way.

Results. For each of the CAC algorithms discussed, fig. 4 shows the distribution of the number of connections in progress during the simulation. The left-most histogram was made using peak-rate admission control; this algorithm admits to the system an average of 7.92 connections. From left to right, the other three histograms were made using the simple effective bandwidth algorithm, the shape-function algorithm, and the Mosquito algorithm. The advantage gained by exploiting statistical multiplexing is clearly apparent: the CAC algorithm, using any of the three estimating techniques, admits significantly more calls than the peak rate allocation scheme. As was found in [33], the Mosquito estimator, which makes use of the intercept μ, is less conservative than the simple effective bandwidth estimator. In terms of the number of connections admitted, the performance of the shape-function estimator lies between the other two estimators. This is consistent with fig. 3; we are thus led to believe that it is the different approximations involved in making the estimates which gives rise to the differences in performance.

To obtain bounds on an optimal admission scheme we multiplexed as many segments of the

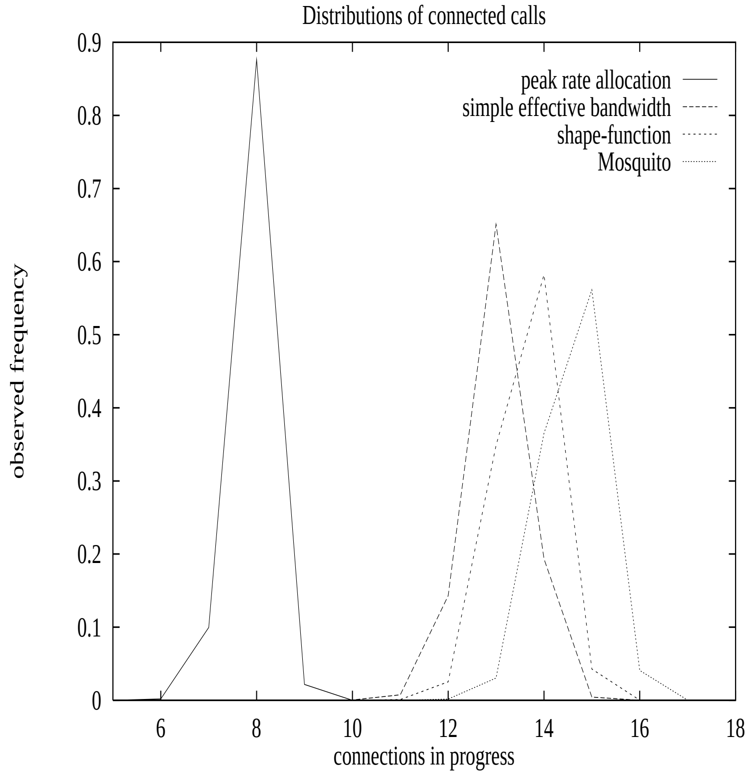

Figure 4: Histogram of the number of calls in progress over the simulation. The mean number of calls in progress under each of the admission schemes were: 7.92 under *peak rate*, 13.0 under *simple effective bandwidth*, 13.6 under *shape-function*, and 14.6 under *Mosquito*.

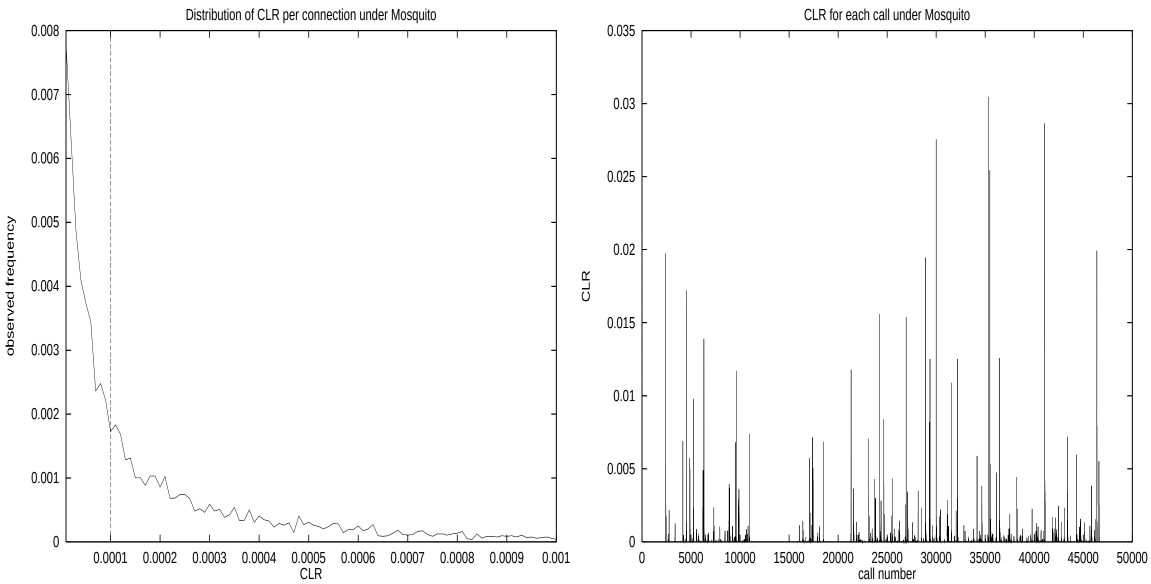

Figure 5: Cell loss under Mosquito. The CLR over the length of the experiment was 3.2×10^{-5} and 4.05% of calls lost more than the target CLR of 10^{-4}. A negligible amount of cell loss occurred when the simple effective bandwidth or shape-function estimators were used.

source traffic as possible such that the CLR constraint was met over the entire experiment. In other words, we found empirically the number of calls which could be multiplexed for a given BWR (the link-rate) and CLR as defined in Equation (5). We found this number to vary between 14 and 16 depending on the sample of traffic used. It is clear that the three algorithms, particularly the Mosquito algorithm, perform very close to optimally.

Turning to the distribution of cells lost under the three admission schemes we find that no loss occurred when using the Simple Effective Bandwidth or the Shape-Function algorithms. This indicates that the Simple Effective Bandwidth algorithm is too conservative: greater link utilisation can be achieved by using the Shape-Function algorithm without in any way compromising QoS. Since the Mosquito algorithm admits more connections than either the Simple Effective Bandwidth or the Shape-Function algorithms we might expect this algorithm to display a higher CLR. We find that this indeed is the case. Mosquito lost 0.0032% of the cells over the length of the simulation which is very close to the target CLR of 0.01%. However this obscures the fact that most connections experience no cell loss while others lose many cells. Overall, 4.05% of the connections lost more than the target CLR under the Mosquito admission policy. The distribution of the CLRs of individual connections and the CLR of each connection are plotted in fig. 5. It is apparent that cell loss occurs in bursts. This is the result of our cell loss allocation scheme: when the buffer overflows an equal number of cells are discarded from each connection that is in progress at that time.

Conclusions and Future Work. We have seen how the CAC algorithm described in [33], using any one of three estimators, can significantly improve on peak rate allocation. The shape-function estimator has been found to perform better than the simple effective bandwidth estimator but less well than the Mosquito estimator in terms of the number of connections admitted. However, unlike Mosquito, the shape-function algorithm did not cause any calls to exceed their target CLR. A possible improvement to the shape-function estimator would be to combine its estimates of the shape-function with observations of cell loss in a simulated buffer in a manner similar to the way in which the Mosquito algorithm uses estimates of μ. The cell loss under such a scheme will require investigation.

References

[1] Joseph Y. Hui. Resource allocation for broadband networks. *IEEE JSAC*, 6(9):1598–1608, December 1988.

[2] N.M. Mitrou and D.E. Pendarakis. Cell level statistical multiplexing in ATM networks: Analysis dimensioning and call acceptance control w.r.t QOS criteria. In *Broadband Technologies*, New Jersey, October 1990. Proceedings Seventh ITC Specialist Seminar.

[3] James Appleton. Modelling a connection acceptance strategy for ATM networks. In *Broadband Technologies*, New Jersey, October 1990. Proceedings Seventh ITC Specialist Seminar.

[4] R.J. Gibbens, F.P. Kelly, and P.B. Key. A decision theoretic approach to call admission control in ATM networks. *IEEE Journal on Selected Areas in Communications, Special issue on advances in the fundamentals of networking – part 1*, 13(6), August 1995.

[5] Peter Key. Connection admission control in ATM networks. *BT Technology Journal*, 13(3), July 1995.

[6] Frank P. Kelly. Notes on effective bandwidths. In Frank P. Kelly, S. Zachary, and I. B. Ziedins, editors, *Stochastic Networks: Theory and Applications*, volume 4 of *Royal Statistical Society Lecture Note Series*, pages 141–168. Oxford University Press, 1996.

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

[8] R. J. Gibbens, F. P. Kelly, Measurement based connection admission control. To appear in 15th International Teletraffic Congress, June 1997.

[9] R. Guerin, H. Ahmadi, and Naghshineh. Equivalent capacity and its application to bandwidth allocation in high speed networks. *IEEE JSAC*, 9:968–981, 1991.

[10] R. J. Gibbons and P. J. Hunt. Effective bandwidths for the multi-type UAS channel. *Queueing Systems*, 9:17–28, 1991.

[11] A.I. Elwalid, D. Mitra, and T.E. Stern. Statistical multiplexing of Markov modulated sources: Theory and computational algorithms. In *Proceedings of the 13th International Teletraffic Congress*, Copenhagen, June 1991.

[12] B. Maglaris, D. Anastassiou, P. Sen, and J.D. Roberts. Performance models of statistical multiplexing in packet video communications. *IEEE Trans. Commun.*, 36:834–844, July 1988.

[13] M. Frater, J.F. Arnold, and P. Tan. A new statistical model for traffic generated by VBR coders for television in the broadband ISDN. *Proceedings IEEE CSVT*, December 1994.

[14] C. Courcoubetis, G. Kesidis, A. Ridder, J. Walrand, and R. Weber. Admission control and routing in ATM networks using inferences from measured buffer occupancy. Technical report, University of California, Berkley, EECS Department, UCB, California, CA94720, 1991.

[15] N. G. Duffield, J. T. Lewis, Neil O'Connell, Raymond Russell, and Fergal Toomey. The entropy of an arrivals process: a tool for estimating QoS parameters of ATM traffic. In *Proceedings of the 11th UK Teletraffic Symposium*, Cambridge, March 1994.

[16] N.G. Duffield, J.T. Lewis, N. O'Connell, R. Russell, and F. Toomey. Entropy of ATM traffic streams. *IEEE Journal on Selected Areas in Communications, Special issue on advances in the fundamentals of networking – part 1*, 13(6), August 1995.

[17] Simon Crosby, Ian Leslie, John Lewis, Neil O'Connell, Raymond Russell, and Fergal Toomey. Bypassing modelling: an investigation of entropy as a traffic descriptor in the Fairisle ATM network. In *Proceedings Twelfth UK Teletraffic Symposium*, Windsor, England, March 1995. IEE, Springer Verlag.

[18] Simon Crosby, Meriel Huggard, Ian Leslie, John Lewis, Fergal Toomey, and Cormac Walsh. Bypassing modelling: Further investigations of entropy as a traffic descriptor in the Fairisle ATM network. In *Proceedings WATM'95 First Workshop on ATM Traffic Management*, Ecole Nationale Superieure des Telecommunications, Paris, December 1995. IFIP W.G. 6.2.

[19] Nils Bjorkman, Alexander Latour, Aziz Miah, Simon Crosby, Ian Leslie, John Lewis, Raymond Russell, and Fergal Toomey. Exploring the queueing behaviour of ATM switches. In *Performance '96*, EPFL, Lausanne, Switzerland, October 1996. IFIP WG 7.3.

[20] ATM Forum. *Traffic Management Specification, Version 4.* ATM Forum/95-0013R8, October 1995.

[21] Sugih Jamin, Peter Danzig, Scott Shenker, and Lixia Zhang. A measurement based admission control algorithm for integrated service packet networks. *IEEE Transactions on Networking*, 5(1):56–70, February 1997.

[22] Sugih Jamin, Peter Danzig, Scott Shenker, and Lixia Zhang. A measurement-based admission control algorithm for integrated services packet networks. In *Proceedings ACM Sigcomm*, Cambridge MA, October 1995.

[23] Sally Floyd. Comments on measurement based admission control for controlled-load services. Submitted to CCR, July 1996.

[24] M. W. Garrett and W. Willinger. Analysis, modeling and generation of self-similar vbr video traffic. In *Proceedings ACM SIGCOMM 94*, pages 269–280, London, UK, August 1994.

[25] Richard Black, Ian Leslie, and Derek McAuley. Experiences of building an ATM switch for the Local Area. In *Proceedings ACM SIGCOMM*, volume 24(4), September 1994.

[26] Song Chong, San-qi Li, and Joydeep Ghosh. Dynamic bandwidth allocation for efficient transport of real-time VBR video over ATM. In *Proceedings IEEE INFOCOM*, pages 81–90, Toronto, Ontario, Canada, June 1994.

[27] V. Paxson and S. Floyd. Wide-area traffic: The failure of poisson modeling. In *Proceedings ACM SIGCOMM 94*, pages 257–268, London, UK, August 1994.

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

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

[30] N.G. Duffield (1996). Economies of scale in queues with sources having power-law large deviation scalings. *J. Appl. Prob.*, **33:** 840–857.

[31] N.G. Duffield (1997). Economies of scale for long-range dependent traffic in short buffers. *Telecommunications Systems.* to appear.

[32] A. J. Ganesh (1996). Bias Correction in Effective Bandwidth Estimation, Computer Systems Group Report, ECS-CSG-23-96, May 1996.

[33] S. Crosby, I. Leslie, B. McGurk, J. T. Lewis, R. Russell, F. Toomey. Statistical properties of a near-optimal measurement-based CAC algorithm. In *Proceedings IEEE ATM'97 Workshop*, pages 103–112, Lisbon, Portugal, May 1997.

[34] C. Walsh, N. G. Duffield. Predicting QoS parameters for ATM traffic using shape-function estimation. In *Proceedings Fourteenth UK Teletraffic Symposium*, Manchester, UK, March 1997.