Using Asymptotic Techniques in Call Acceptance Management for ATM networks

C. Courcoubetis G. Fouskas

Department of Computer Science, University of Crete
and Institute of Computer Science, FORTH.

Abstract

We discuss the issue of providing effective Call Acceptance Management (CAM) for ATM networks in connection with the work presented in [1, 7, 8]. Two different approaches are described which can be used for CAM. The first approach is based on a fast estimation method for the spare capacity in the network. The second approach is based on an adaptive method for the calculation of the effective bandwidths of traffic sources. This method obtains a good approximation of the effective bandwidths in the ATM environment by estimating the index of dispersion for each traffic source. Both approaches use information which is extracted from the actual traffic flows and are justified by a formal mathematical development based on large deviation asymptotics.

1 Introduction

An ATM network provides support for a wide range of services which have differing bandwidth requirements, traffic pattern statistics and Quality of Service (QoS) constraints. Cells belonging to a number of different calls are statistically multiplexed in order to make efficient use of network resources. At the same time, the various QoS criteria must be maintained for each of the calls being served. The network is required to fulfill, with the help of the Call Acceptance Management (CAM) function, the two conflicting goals of maximizing resource usage and providing acceptable QoS. The CAM function should be able to decide, in real time, if a request for a connection can be satisfied without having QoS degradation in the network.

One approach for CAM, which is presented in [1], is to estimate, for each link, the additional fraction of cells that would be lost if the new call were routed through this link. This method does not require models describing the traffic. By doing direct measurements at the switches it estimates the current cell loss probability and the existing spare capacity of the switch. To reduce the variance of the estimates (so that the information is timely and accurate), each switch estimates the cell loss that would occur with smaller buffers (virtual buffers). Then an extrapolation using some scaling property justified by large deviations allows the approximation of the actual cell loss probability. The method is summarized in section 3.

An other popular approach for doing CAM is based on the notion of an *acceptance region*. The acceptance region for a given link is the set of traffic loads for which the performance criteria (e.g. cell-loss rate or delay) can be guaranteed. Thus the definition of an acceptance region divides the set of possible traffic loads into two partitions: inside the acceptance region we find loads which produce acceptable cell-loss rates ("good" points); outside the acceptance region we find loads which produce unacceptable cell-loss rates ("bad" points). If an easily-defined acceptance region exists, then the CAM function needs only to evaluate whether or not adding the new call will cause the links to operate within or outside the acceptance region. Since acceptance regions are for most practical purposes impossible to compute analytically (based on inexact data and imperfect models), much research has been devoted to deriving successful approximations of such regions (see [4, 6, 3]).

One possible approximation for the acceptance region is based on the concept of an *effective bandwidth* [4, 2, 5, 9]. The effective bandwidth of a given call is the "amount" of bandwidth that must be allocated at each link and by the call in order to obtain an acceptable level of QoS by the network. The acceptable level of QoS must be the same for all calls. The *traffic load* on a link is therefore the sum of the effective bandwidth of the calls passing through the link. If we define N_i to be the number of calls of type i currently being served by a link and α_i to be the effective bandwidth of type i calls, the current load on a link is $\sum_i N_i \alpha_i$. If this sum is less than c, where c is the service rate of the link, then this load is declared to be acceptable. If the sum is greater than c, this load is declared to be unacceptable. The problem with using the effective bandwidth idea, as introduced in [4, 2, 5, 9], is that the effective bandwith calculation is based on a particular model of the traffic source. Hence, since in reality such models are fairly inacurate, any pre-calculation of the effective bandwidth and hence of the acceptance region is of questionable value. The approach described in [8] is a remedy to the above problem since it allows the adaptive calculation of a good approximation of the effective bandwidth for a large number of types of traffic sources. It requires the estimation of the index of dispersion of the input process produced by a source. This is a well known problem in time-series analysis, and there exist powerfull algorithms for estimating it. This approach is described in section 4. Finally, in section 5 we discuss several directions for further research.

2 Effective Call Acceptance Management

Quality of Service in ATM networks can be measured by a variety of parameters, including average cell delay, delay jitter, cell loss rate, and many others. Not all of these parameters are directly controllable by, or the responsibility of, the call acceptance manager. In this paper, we use cell-loss rate as the sole performance criterion for QoS measurements, since, for most practical considerations, small delay and jitter are usually guaranteed by the architecture of the network. The problem of considering average delay as the QoS criterio has been addressed in [5].

When a call with a given traffic profile makes a request for a connection, the call acceptance algorithm is responsible for deciding if and how this new call can be routed through the network so that the resulting cell-loss rate at each link along the route remains below the acceptable cell-loss rate threshold. Any CAM strategy must therefore be able to estimate, for each link, if it is currently providing acceptable QoS, and if adding the new call will degrade the QoS.

The most promising way in which this could be achieved is to explicitly measure, at each link, the cell-loss rate that is being incurred by the current mix of traffic. Thus we are basing our estimate on the way in which the traffic is *actually* behaving, not on an inaccurate parametric model. An obvious problem arises in estimating cell loss at the individual links, since the mix of traffic passing through the network is constantly changing. There is no guarantee that a link will serve a static mix of traffic for a length of time which is adequate for obtaining an accurate estimate of its cell-loss rate by simply counting the number of cells arriving and the number of cells being dropped. For example, to measure, with reasonable accuracy, a cell-loss rate of 10^{-6}, one would have to observe in the order of 10^{9} cells passing through a buffer. Assuming that a 150 Mb/s link is serving a buffer of ATM cells, and that the load on the link is .80, the estimate would require approximately 10^{3} seconds of network time. This is, quite obviously, unacceptable because the decisions related to QoS must be made within a much smaller time frame. The remedy for this is the virtual buffer technique from [1], were the variance of the estimator is reduced by many orders of magnitude. It allows the estimation of the cell loss probability in the previous example in about 1-2 seconds of real time.

An other option for performing CAM is by not directly estimating the current cell-loss probability; one could estimate the total effective bandwidth of the traffic streams and compare it to the available link capacity. Again the idea is that one should not derive these effective bandwidths from existing traffic models, but directly measure them from the network. There are many reasons that one can not precompute the effective bandwidth of a traffic type. Consider for example digital voice. We know that a reasonable model for this type of traffic is a two-state Markov fluid. The rates of the underlaying Markov chain can not be the same for all voice sources. Hence there are two choices for CAM; either use the "worst case" voice model with the highest effective bandwidth, or use the average effective bandwidth among all voice models (the network operator could obtain such an average based on statistics of the operation of the network). One can easily see that the first choice will result in a underutilization of the network, whereas the second choice will violate the QoS requirements, see [13]. As it turns out, the effective bandwith of a mix of traffic types, whose consistency is random, is larger than the average of the effective bandwidth of the mix (average according to the distribution of the mix). The results in [8] allow the estimation of a good approximation for the effective bandwidth of a traffic stream without knowledge of an apriori model, under very general assumptions. The basic result in this paper shows that asymptotically in the buffer size B of the switch, most general stationary traffic streams have effective bandwidths. The error of the approximation is of the order of $o(1/B)$ ($o(1/B)$ tends to zero faster than $(1/B)$, for B tending to infinity), and in order to evaluate the effective bandwidth one needs to estimate the average bitrate and the index of dispersion of the traffic stream. This is a standard problem in time-series analysis and there are efficient procedures for obtaining the estimators using spectral methods, see [10].

As a last remark, we believe that these methods can be used in a complementary fashion and provide the timely and accurate information needed by the CAM functions.

3 Using Virtual Buffers to Estimate Spare Capacity

In [1] the authors propose a new technique which can be used to reduce the amount of time required to achieve an accurate cell-loss-rate estimate. By observing the overflow process which would occur if the buffers of the switches were reduced in size, one can derive an estimate for the overflow process of the actual system. Also, a scaling property allows the estimation of the cell loss probability if one would increase the traffic by a small percentage. This is implemented as follows.

Let $F(\mathbf{n}, B, c)$ be the cell loss rate on a link served with capacity c, buffer B, and $\mathbf{n} = (n_1, ..., n_I)$ be the vector with components the number of calls of class $i = 1, \ldots, I$

going through the link. As we have mentioned before, typically $F(\mathbf{n}, B, c)$ is required to be small in the order of 10^{-8}. In order to obtain a small variance estimator of $F(\mathbf{n}, B, c)$, we associate with the buffer of size B three *virtual buffers* with sizes B/k_1, B/k_2 and B/k_3. There is a device which estimates the probability of overflow for each of these smaller buffers by performing direct measurements in the actual switch (which operates with buffer size B). Since the virtual buffers are smaller than the actual buffer, these overflow probabilities are much larger than the overflow probability of the actual buffer, and can therefore be estimated with small variance in a reasonably short time. Using the theory of large deviations, for the case of Markov fluid sources, we obtain that

$$F(\mathbf{n}, B/k, c) = A(B/k)^a \exp(-(\frac{B}{k})I(\mathbf{n}, c)).$$

By knowing the values of F (by estimation) for $k = k_1, k_2, k_3$, we solve the system of the three corresponding equations to compute the unknowns A, a, I, and we then use the formula to compute $F(\mathbf{n}, B, c)$. In [1] there is a discussion for how to choose the sizes of the virtual buffers given the tradeoff that if these buffers are small then the original system is overdistorted and there is a large extrapolation error going back to the original B, and if the buffers are large then the variance of the estimators is large.

There is an other important result which enables us to estimate the spare capacity. In [1] it is proved that

$$F(\mathbf{n}(1+\epsilon), B, c) = F(\mathbf{n}, B, \frac{c}{1+\epsilon}).$$

Hence in order to estimate $F(\mathbf{n}(1+\epsilon), B, c)$ which corresponds to the cell-loss rate if the input traffic is increased by a fraction of ϵ, one could estimate instead $F(\mathbf{n}, B, c/(1+\epsilon))$, which corresponds to the cell-loss rate with the current traffic mix and the service rate reduced by the same fraction. One could easily think of devices which keep track of the buffer occupancy when the service rate is $c/1+\epsilon$, for estimating this quantity during the actual operation of the link. This is the MINOS monitor described in [1].

We show some simulation results from [7] for validating the virtual buffer estimation technique. We chose an arbitrary time interval of 2 seconds of network time for obtaining the cell-loss estimate from a 150 Mb/s link. This is a reasonable time frame for a CAM strategy to obtain information necessary for making call acceptance decisions. Each of the traffic sources is modeled by an ON-OFF Markov fluid, with parameters (0.1, 0.01, 10000), (0.1, 0.05, 2500), (1, 0.25, 375) for the traffic type (i), (ii) and (iii) respectively (the parameters $(S_{OFF}, S_{ON}, R_{peak})$ are the average times the system spends in state OFF, ON and the rate of producing cells in state ON, respectively). In the graphs of Figures 1 and 2, we present results for different sets of traffic mixes. On the x-axis is the load ϱ placed on the link by the given traffic mix (i.e., some combination of the two

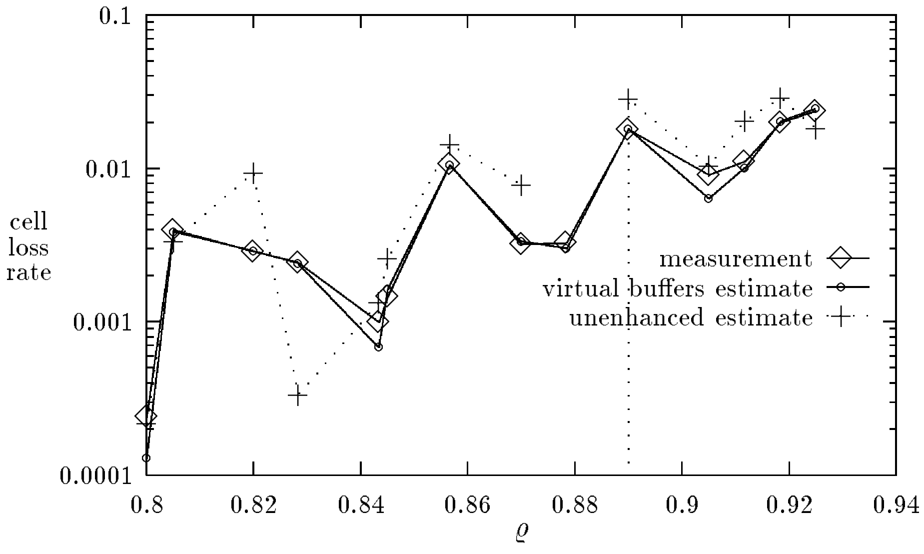

Figure 1: Mix of Traffic Types (i) and (iii)

given traffic types). On the y-axis is the estimated cell-loss rate. For each traffic mix, the graphs contain three separate points. The first point corresponds to an accurate estimate of the cell-loss rate obtained by observing cell losses from the actual buffer over a long period of time. The second point corresponds to the estimate achieved by using virtual buffers in the 2-second time frame. The third point corresponds to the estimate achieved by counting dropped cells in the 2-second time frame. The graphs show that for low cell-loss rates (i.e., below 10^{-2}) the virtual buffer technique much more consistently provides a good estimate than counting dropped cells. For many of these points, the counting technique failed to see a single cell dropped in such a short time period, and it concluded that the cell loss was zero. For some points, the counting technique was lucky enough to make a good guess, but it could not do so with any consistency. The estimate provided by the counting technique at *smaller* cell-loss rates than those presented here is worse and completely unusable. The virtual buffer technique does not suffer from these shortcomings and is able to estimate cell-loss rates in the region of interest. For high cell-loss rates, the counting technique was able to do quite well in many instances, since overflow was frequent and hence a sufficient number of samples were available. However, the network will seldom operate in the region of cell-loss rates which can be accurately estimated by the counting technique.

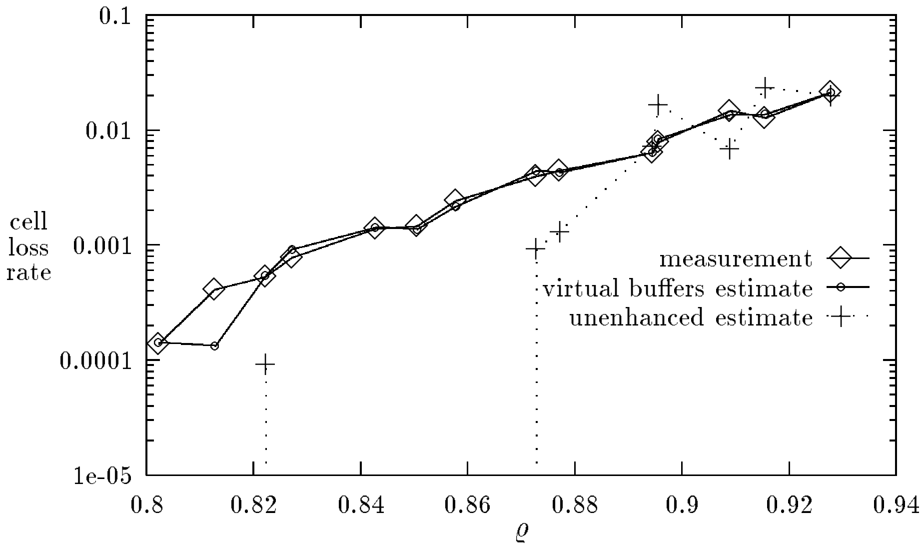

Figure 2: Mix of Traffic Types (ii) and (iii)

4 Adaptive Evaluation of Effective Bandwidths

We briefly review the notion of effective bandwidth already mentioned in the introduction, and summarize the results in [8].

The traffic in an ATM network is packaged in cells and travels between switches over the links of the network. Traffic sources are bursty and therefore over short periods of time the total rate at which cells arrive at a switch can be faster than the total rate c at which they are switched to output links. For this reason cells are buffered and an important problem is to control the total amount of traffic carried by each switch so that the probability of buffer overflow and, hence the resulting cell-loss rate, is very small.

A number of authors have addressed this problem and have shown that for a variety of traffic models, a class i of traffic, $i = 1, \ldots, I$, can be assigned an effective bandwidth α_i, such that if N_i sources of class i are served simultaneously by a link, then the constraint that the probability of buffer overflow is kept smaller than some specified amount can be written

$$c \geq \sum_i \alpha_i N_i.$$

This papers follows on from the work in [5], [2], [9], [11], [12]. In [5] Kelly obtained effective bandwidths for the problem of controlling the cell-loss rate of a $GI/D/1$ queue. (He also considered the problem of controlling average work seen by a customer arriving at a $GI/D/1$ queue). In [2] Gibbens and Hunt presented a method for deriving effective bandwidths if the sources are modeled as two-state Markov fluids and if the buffer serves the fluid at a constant rate. In [9] 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 stationary Gaussian process. In [11] 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 [12] Kesidis and Walrand discuss ways to derive the effective bandwidths for general Markov fluids. The problem with the above results is that they require the model for the traffic sources to be known, which is rarely the case in most practical situations.

In [8] the work of [9] is extended to non-Gaussian stationary sources and allows us to compute effective bandwidth asymptotics that are accurate to within $o(1/B)$, as a function of the buffer size B. The important result is that for general stationary sources and large B, the effective bandwidth of a source is a function of two numbers, its mean bit rate and its index of dispersion. *Thus the index of dispersion is identified as the most appropriate measure of burstiness for ATM traffic sources.* This implies that a good approximation (the error is of the order of $o(1/B)$) of the effective bandwidth of a source can be computed by estimating simple parameters, without knowledge of the model of the source. We summarize now the model and the results in [8].

Time is discrete and at each epoch n a typical source in class i delivers to the buffer a number of cells X_n^i, where $X_1^i, X_2^i, \ldots$ is a stationary process of possibly correlated random variables with mean m_i. Also during each epoch the switch serves c cells. Let

$$\alpha_i = m_i + \frac{\delta \gamma_i}{2B}$$

where $\mathrm{e}^{-\delta}$ is the maximum acceptable cell-loss rate, and

$$\gamma_i = \lim_{n\to\infty} (1/n)\mathrm{E}\left[\left(\sum_{k=1}^{n}(X_k^i - m_i)\right)^2\right] = \lim_{n\to\infty}(1/n)\mathrm{var}\left[\sum_{k=1}^{n} X_k^i\right]$$

is the *index of dispersion* for the source of class i. Observe also that

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

where $\gamma_i(k)$ is the kth order autocovariance. The above converges for sensible second order stationary processes and can be estimated from the data by well known methods (see [10]). The basic theorem in [8] states that the cell loss rate at the switch is less than $\mathrm{e}^{-\delta}$ if and only if

$$c \leq \sum_i N_i \alpha_i + o(1/B).$$

There are some additional technical conditions that must hold in order for the above formula to be correct, which hold indeed for a large class of stationary processes. Note that the above formula allows the calculation of the approximation of the effective bandwidth for general n-state Markov fluids, by neglecting the $1/B^2$ order terms, see [8].

5 Summary and Directions for Further Research

In this paper we describe two approaches which can be used for effective Call Acceptance Management for ATM networks. Both approaches are justified by a formal mathematical development based on large deviaton asymptotics. Thus, these results hold when the cell-loss rate is small and when the buffer size is big. We are currently doing extensive simulations to validate the practical relevance of these methods in the area of ATM traffic parameters and switch characteristics.

The use of the notion of effective bandwidths makes the problem of routing in ATM networks become similar to the traditional problem of routing in circuit-switched networks. One should explore how the "rich" theory of circuit-switched networks (for example, dynamic routing and trunk reservation stategies) can be effectively applied to ATM networks.

Another topic of interest is to extend the existing ATM models in order to describe the overflow processes of high-speed LANs, such as FDDI and DQDB. There are models of the above systems which can be analyzed by using similar asymptotic techniques. We are currently implementing a subset of the above results for predicting acceptable traffic mixes in our FDDI token ring.

References

[1] C. Courcoubetis, G. Kesidis, R. Ridder, J. Walrand and R. Weber, "Admission Control and Routing in ATM Networks Using Inferences from Measured Buffer Occupancy," to appear.

[2] R.J. Gibbens and P.J. Hunt, "Effective Bandwidths for the Multi-type UAS channel," Statistical Laboratory, University of Cambridge, preprint, 1991.

[3] A. Hiramatsu, "ATM Communications Network Control by Neural Networks," *IEEE Trans. on Neural Networks*, Vol. 1, No. 1, March 1990, pp. 122-130.

[4] J. Hui, "Resource Allocation for Broadband Networks," *IEEE Journal on Selected Areas in Comm.*, Vol. 6, No. 9, Dec. 1988, pp. 1598-1608.

[5] F.P. Kelly, "Effective Bandwidths at Multi-class Queues," *Queueing Systems 9*, 1991.

[6] Race Project R1022 "Technology for ATD", Race Document TG_123_0006_FD_CC.

[7] C. Courcoubetis, G. Fouskas, V. Friesen and S. Sartzetakis, "Real-Time Issues in Call Acceptance Management for ATM networks", *Proceedings of the Fifth RACE TMN Conference*, London, 1991.

[8] C. Courcoubetis and R. Weber, "Effective Bandwidths for Stationary Sources", sent for publication.

[9] C. Courcoubetis and J. Walrand, "Note on effective Bandwith of ATM Traffic".

[10] D. Cox and P. Lewis, "The Statistical Analysis of Series of Events", *Chapman and Hall*, 1966.

[11] G. De Veciana, C. Oliver and J. Walrand, "Large Deviations for Birth-Death Markov Fluids", to appear.

[12] G. Kesidis and J. Walrand, "Effective Bandwidths for Multiclass Markov fluids and Other ATM Sources", to appear.

[13] C. Courcoubetis and R. Weber, "The Effective Bandwidth of traffic with Random Consistency", in preparation.