An on-line Estimation Procedure for Cell-Loss Probabilities in ATM links

COSTAS COURCOUBETIS*, GEORGE FOUSKAS*
AND RICHARD WEBER†

Abstract

We present a methodology for estimating on-line the cell-loss probability of an ATM link. It is particularly suitable for estimating small probabilities of the order 10^{-6}–10^{-9}, with variance orders of magnitude smaller than traditional estimators. The method is justified by the theory of large deviations, and the information required is based upon the actual traffic flows rather than upon the analysis of some specific traffic model. The method is effective when there is a large degree of statistical multiplexing; in other words, when the number of input traffic sources is large. The statistical properties we require for the traffic are very general and are met by most real-time traffic source models. Experimental results suggest that the estimators obtained are timely and accurate in the case of large systems, and hence that the method can be used for the on-line estimation and control of the congestion in the actual network.

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) requirements. All information flows are organized into fixed-size packets, called cells, and carried over high capacity links between switches in the network. Cells belonging to a number of different calls are statistically multiplexed in order to make efficient use of network resources. Switches are buffered, as a safeguard against those cases that cells arrive at a switch faster than they can be switched to output links. Occasionally, the buffer will not be large enough and cell loss will occur. Therefore, a measure of QoS offered by the network to a traffic stream is its cell

*Department of Computer Science, University of Crete and Institute of Computer Science, FORTH, Greece.
†Statistical Laboratory, 16 Mill Lane, Cambridge CB2 1SB, U.K.

loss probability (CLP); for most types of traffic this probability should be very small, of the order of 10^{-6} to 10^{-9}.

Accurate and timely CLP estimation is important but difficult

The network is required to attain high utilization and to provide acceptable QoS, with the help of the Call Admission Control (CAC) function. The CAC must decide whether or not to accept a new call, based on the new call's traffic characteristics and quality requirements and on the current network status. It is therefore important that the network have a mechanism for obtaining accurate and timely estimators of the CLPs that occur at each link. A generic problem that the network management system faces is that since the events of interest are rare (e.g., cell-loss rates of the order of $10^{-6} - 10^{-9}$), any brute-force on-line estimation procedure would fail because of the large time required to make an accurate estimate. For example, to estimate, with reasonable confidence interval, a cell-loss rate of 10^{-6}, by simply counting the number of cells arriving and the number of cells being dropped, one would have to observe in the order of 10^9 cells passing through the buffer. Assuming that a 150 Mb/s link is serving the buffer, and that the load on the link is 0.80, the estimate would require approximately 10^3 seconds of network time. Such a time is *impractical* since decisions about accepting a new call in the network must be done quickly, and because the composition of the traffic can change in such a large time interval. Furthermore, except for very simple models, it is impossible for a queueing theory analysis to accurately predict the cell loss rate in realistic systems. Similar problems occur when one wants to do simulation; obtaining accurate estimates of CLPs could require simulation time in the order of magnitude of tens of hours.

Our approach

The key idea in our approach, developed in [2], is similar to the one found in [1], but applied on a different asymptotic for the system, which resulted in a large improvement of the variance reduction in systems of practical interest. The following example illustrates the basic idea. Consider an ATM link that is carrying 1000 sources, with a loss rate of about 10^{-10}. As we have already mentioned, it is difficult to measure such a small cell-loss rate directly. However, if we measure the loss rate p in a virtual link, that is carrying 200 sources and its bandwidth and buffer space are one-fifth of the corresponding values of the actual link, then we should obtain a value of about 10^{-2}. Such a cell-loss rate is much larger than that of the actual link, and therefore can be measured with small variance in a relatively short time. The results in [2] justify using p^5 as an estimate of the actual cell-loss rate. Of course there is the opportunity to make five independent estimates of p, using five groups of 200 sources which could reduce even further the variance.

The actual method is quite more sophisticated and is described in section 3. The basic idea is the same: we assume that the cell-loss probability is given in terms of a function, whose form is known (by the results in [2]), but whose coefficients depend on the system and traffic parameters and are unknown. We accurately measure the cell-loss probability in a number of independent virtual subsystems, and use least squares linear regression techniques to determine the unknown coefficients. Then we use the regression to calculate the cell-loss probability for the actual system. Our experimental results have shown a remarkable accuracy of the method. In almost all cases, the error of our estimator was consistently small, less than one order of magnitude.

How to use our method

The impact of our results is twofold. First, they can be used for speeding up the simulation time for such large systems. Our method requires the same amount of computation per unit of network time as the one required for the simulation of the actual ATM link. A second application of the results is towards the on-line estimation of the CLPs in actual ATM links. Experimental results suggest that the time constant required to obtain an accurate picture of the CLPs by our method is small enough to allow its use for controlling congestion. Furthermore, by a small variation of the method we can infer the CLP that would occur if we would accept an extra amount of traffic in the system (similarly to [1]). This variant can be used for implementing dynamic Call Admission Control procedures in ATM networks.

The rest of the paper is organized as follows. In section 2 we explain the model and the CLP formula obtained by the asymptotic used in [2]. In section 3 we describe the estimation procedure. Section 4 contains the simulation results, and in section 5 we conclude with a summary and some directions for further research.

2 The model and the CLP formula

In this section we present our model and the asymptotic CLP formula. The precision of the formula increases as a certain parameter of the systems grows. In our case this is the parameter N defined below, which represents the "size" of the system. For details, the reader is referred to [2].

An output link of an ATM switch is modeled as a queue with a constant service rate and a buffer for B cells. We shall suppose that time is discretised into epochs, $n = 1, 2, \ldots$, and that during epoch n the cell service rate and arrival rate are both constant, and equal to C and X_n cells per epoch respectively. This models the usual case where there are dedicated resources C and B for the real-time traffic on the above link (using high priority for real-time traffic ensures the first condition). The workload at the start of epoch n is denoted W_n and

$$W_{n+1} = \max\{0, \min[(W_n + X_n - C), B]\}.$$

We shall assume that $\{X_n\}$ is an ergodic process, and that $E[X_n] < C$. This condition implies that $\{W_n\}$ is ergodic.

Suppose that $\{X_n\}$ is the superposition of sources of M different types, with there being $N\rho_i$ sources of type i, $i = 1, \ldots, M$. Thus the traffic is defined by the parameter $\rho = (\rho_1, \ldots, \rho_M)$ and the scaling parameter N. Let b and c denote respectively the amounts of buffer space and bandwidth per source, so that $B = Nb$ and $C = Nc$.

The cell loss rate can be expressed as $L(c, b, N) = E[(W_n + X_n - C - B, 0)^+]$, when W_n and X_n have their stationary values. A related measure is the proportion of time the buffer is full, which we denote $\Phi(c, b, N) = P(W_n = B)$. Alternatively, we might suppose the buffer is infinite and study the proportion of time that the content is above level B, say $Q(c, b, N)$. Note that $L(c, b, N)$, $\Phi(c, b, N)$ and $Q(c, b, N)$ are all functions of ρ, but to lighten the notation this dependence is not shown. One can prove that exactly the same asymptotic holds for $L(c, b, N)$, $\Phi(c, b, N)$ and $Q(c, b, N)$. In effect, these quantities differ by approximately constant multiplicative factors.

The large N asymptotic

The large N asymptotic with which we are concerned takes the following form.

$$\Phi(c, b, N) = \exp(-NI(c, b) + g_1(c, b, N)), \tag{1}$$

where $\lim_{N\to\infty} g_1(c, b, N)/N = 0$. Both $I(c, b)$ and $g_1(c, b, N)$ depend on ρ but as we have done above, this is suppressed. The rate $I(c, b)$ is found as the solution to an optimization problem posed in terms of time dependent logarithmic moment generating functions. Asymptotic (1) can be compared with a well-known asymptotic in B,

$$\Phi(c, b, N) = \exp(-NbH(c) + g_2(c, b, N)), \tag{2}$$

where $\lim_{b\to\infty} g_2(c, b, N)/b = 0$. This is an asymptotic that has been studied in a long series of papers, by [3], [4], [6], [13], [14] and [20]. Either (1) or (2) may be used to estimate Φ, as $\exp(-NI(c,b))$ or $\exp(-NbH(c))$ respectively. In [2] we show that $I(c,b)/b \to H(c)$ as $b \to \infty$, thus, if the buffer space per source is large then either of these estimates might be used.

However, we find that it is usually better to base an estimate of Φ on (1) rather than (2). For one thing, to base an estimate upon the asymptotic for large N takes more account of what really happens in practice. In the output link of an actual ATM switch, the number of sources is of the order 10^3–10^4 and so it is correct to apply a result that assumes N is large. One cannot say the same for the approach that assumes the buffer is large; in many ATM switch designs, the total buffer space per link may be moderate, say $100 - 200$ cells. This means that in actual systems there is a much smaller amount of buffering per source than would be required for use of the estimate based on the large b asymptotic (see [7]).

The utilization of the switch by the bursty traffic also has an impact on the relative merits of the approaches. In an ATM network, some of the real-time traffic, such as voice and video, will be assigned high priority and some, such as file transfer and interactive traffic, can be delayed and subject to flow control. The real-time traffic will only utilize a proportion of the bandwidth. If we are concerned to model the high-priority traffic, $\{X_n\}$ can be considered to be a model for this traffic alone. Simulations show that when the utilization of the switch is low, say $X_n = (2/3)C$, the estimation of the overflow rate using the large N asymptotic remains accurate, whereas the estimate based on the large b asymptotic is very poor, overestimating the overflow rate by as much as 10^5 (see [7]).

The formula

Both the large N and large b asymptotic are expressed in terms of logarithmic moment generating functions. Suppose $\{Y_1^i, Y_2^i, \ldots\}$ denotes cells generated in epochs $1, 2, \ldots$, by a traffic source of type i, and define

$$\varphi_m^i(\theta) = \frac{1}{m} \log E\left[\exp\left(\theta \sum_{n=1}^{m} Y_n^i\right)\right].$$

We shall also suppose that the asymptotic logarithmic moment generating function exists, defined as

$$\varphi^i(\theta) = \lim_{m\to\infty} \varphi_m^i(\theta).$$

We write $\varphi_m(\theta) = \sum_i \rho_i \varphi_m^i(\theta)$ and $\varphi(\theta) = \lim_{m\to\infty} \varphi_m(\theta)$.

The large N asymptotic result is expressed in the following theorem. This result has also recently been proved independently by [5] and by [17] for the case of on-off Markov fluid sources.

Theorem 1 *For the model above and under appropriate assumptions,*

$$\lim_{N\to\infty} \frac{1}{N} \log \Phi(c,b,N) = I(c,b),$$

where

$$I(c,b) = \inf_m \sup_\theta [\theta(b+mc) - m\varphi_m(\theta)]. \tag{3}$$

For a fixed m the supremum over θ in (3) calculates a large deviation coefficient (appropriate to an asymptotic in N) for the probability that the buffer overflows in m periods; the way in which this overflow occurs is by each of the N sources producing $mc + b$ cells, and so contributing an equal share to the total of $mC + B$ cells that causes the buffer to fill. The infimum fixes upon the m for which this probability is greatest.

3 The estimation method

The form of the CLP formula from the previous section suggests that for large values of N (number of traffic sources) the frequency of buffer overflow $\Phi(N)$ can be approximated by $AN^\psi \exp(-NI(c,b))$, where A and ψ are constants, and $I(c,b)$ is given by the solution to the optimization problem described previously. This is because we could neglect the terms varying slower than $\log(N)$ in an expansion of $g_1(c,b,N)$ and absorb them in the constant A.

The next observation is key to our approach. Suppose that we can make an on-line measurement to estimate the cell loss rate that would occur if $N\rho/k$ sources were routed through a link that has total bandwidth C/k and total buffer B/k. This can be done by a special device at the switch level which simulates a 'virtual' link whose capacity is only $1/k$ as large as that of the actual link, operating on a representative fraction of the actual input of size $1/k$. The device provides an estimator of the buffer overflow rate that occurs in the virtual system of size $1/k$. Our asymptotic result suggests that the loss rate in the virtual system should be of the order $\exp(-NI(c,b)/k)$, hence it can be estimated accurately in a time that is orders of magnitude smaller than the time required to make an estimate of the buffer overflow rate in the original system, which is of the order $\exp(-NI(c,b))$. *The interesting question consists on how to get the value of $\Phi(N)$ without estimating it directly (since this would be prohibitively slow) but by using estimators of the $\Phi(N/k)$ for $k > 1$.*

The answer to this is that since $\Phi(N/k) \approx A(N/k)^{\psi} \exp(-(N/k)I(c,b))$, we can compute the three unknown coefficients A, ψ, $I(c,b)$ by first evaluating the above equation for at least three different values of k where the value of $\Phi(N/k)$ is obtained by on-line estimation using the idea introduced previously, and then by solving the resulting set of three equations with three unknowns. This is done as follows.

Pick three values $k_1, k_2, k_3 > 1$, and estimate directly the values $\Phi(N/k_i)$, $k = 1, 2, 3$. Solve the linear equations

$$\log \Phi(N/k_i) = \log A + \log(N/k_i)\psi + N/k_i I(c,b), \; k = 1, 2, 3,$$

for the three unknowns $\log A$, ψ, $I(c,b)$. Finally, compute the desired quantity $\log \Phi(N) = \log A + \log N\psi + N/I(c,b)$.

¿From the above procedure, it follows that the time to do the on-line estimation is determined by the size of the largest virtual subsystem. We use now a brute-force method to show the dramatic reduction of the estimation time required even if k_i is only 4. First we use a formula that allows the evaluation of the number n of steps (cell times) required for the estimator of the CLP to converge, following results from [19] pp. 335. Assume that the CLP is equal to p, and that the cell-loss events Z_i, $i = 1, ...,$ (Z_i is 1 if the ith cell is lost) are i.i.d. with $Pr\{Z_i = 1\} = p$. Then, if we require for the estimator $\hat{p}_n = n^{-1}(Z_1 + \cdots + Z_n)$ that $Pr\{|\hat{p}_n - p|/p \leq 0.05\} \geq 0.95$ we need a value of $n \approx 10^3 p^{-1}$. Now suppose that we have a 155Mbs link, with anticipated CLP of the order of 10^{-6}. The above analysis implies that we need $n \approx 3 \cdot 10^3$s (this is the order N system). On the other hand the subsystem of order $N/4$ will incur a CLP of the order 10^{-3} (by using the large N formula and omitting the $o(N)$ terms) which can be estimated in only one second! Of course, this is very approximate since cell-loss events are not independent. This is only useful to show the dramatic decrease in variance which occurs in the virtual subsystems.

The actual method for determining the unknowns $\log A$, ψ, $I(c,b)$ can be quite more sophisticated. Our actual method considers a large number of independent virtual subsystems and then uses regression analysis to find the values for the unknowns which best fit the acquired data. Besides the obvious virtues of such a method, a strong advantage is the possibility of evaluating the error of the fitting, and hence know when the values obtained become accurate. In the following section we provide a detailed outline of our implementation together with some simulation results which support our claims.

A final comment is needed about the choice of the constants k_i. We do not have any answer yet for the problem concerning the optimal choice of the above constants. An interesting discussion can be found in [1]. A simple heuristic is that small values of N/k_i will result in estimators with small variance but which will contribute too large extrapolation errors since we want to compute the CLP for large values of N. Hence one should consider a combination of small and relatively large values of subsystem sizes.

4 Simulation Study

In this section we describe the traffic source models that were used in our experiments, we present our simulation model, and discuss some of our simulation results. Extensive simulation experiments were conducted to validate the accuracy of our on-line estimation technique, and to study its effectiveness for different real-time traffic sources.

4.1 Traffic Source Models

In this subsection we describe some simple and powerful models, which were used to characterize realistic ATM traffic sources.

The ON/OFF fluid model

In the ON/OFF fluid model a source is either active and producing a cells per second at constant rate, or idle and producing no cells. The two states alternate according to a continuous time Markov process and the lengths of time spent in the active and idle states are exponentially distributed, with means of $1/\mu$ and $1/\lambda$ respectively.

The model can be used for a voice source with ADPCM modulation and a speech activity detection mechanism. Ibrahim, et al. [9] suggest parameters of $a = 64$Kb/s, $1/\lambda = 650$ms and $1/\mu = 352$ms.

Autoregressive models

Autoregressive models have been widely used to characterize video sources, without scene changes or scene cuts and with moderate motion [15], [16], [11]. Let X_n represent the bit rate of a video source during the nth frame. Then, an autoregressive model of order k, denoted by AR(k), for the stochastic process $\{X_n \; ; \; n = 1, 2, \ldots, N\}$ is

$$X_n = a_0 + \sum_{i=1}^{k} a_i X_{n-i} + w_n \qquad n = k+1, \ldots, N$$

where $a_i, i = 0, \ldots, k$ are constants and the random variables $\{w_n\}$ are independent and identically distributed.

The AR(1) model: In our study we used the following AR(1) model of Maglaris, et al. [15]

$$X_n = 0.8781 X_{n-1} + 0.1108 w_n \quad \text{Kb/frame},$$

where $\{w_n\}$ is a sequence of i.i.d. Gaussian random variables, with mean 143 and standard deviation 250. These values were obtained by matching the average bit rate and the autocovariances with actual videotelephone data.

The AR(2) model: Heyman, et al. [11] suggest that the following model provides a good approximation of the bit rate of a video-teleconference source

$$X_n = 1.2608 + 1.20682 X_{n-1} - 0.225705 X_{n-2} + w_n \quad \text{Kb/frame},$$

where $\{w_n\}$ is a sequence of i.i.d. Gaussian random variables, with mean 0 and standard deviation 6.487. Regression analysis were used to find the above values for the coefficients which best fit a long (30 min) sequence of real video-teleconference data.

The DAR(1) model

In the DAR(1) model [12], a finite-state Markov chain is used to characterize the amount of traffic produced at each video frame. It was found that the DAR(1) model provides sufficient accuracy in characterizing generic variable bit rate video traffic sources [11], [8], [10]. Let $\{X_n\}$ be a stationary DAR(1) process with finite state space E and first-order serial correlation ρ, $0 \le \rho < 1$. Then

$$X_n = \begin{cases} X_{n-1} & \text{with probability } \rho \\ Y_n & \text{with probability } 1 - \rho \end{cases}$$

where $\{Y_n\}$ is a sequence of i.i.d. random variables taking values in the finite state space E, and X_0 is a random variable independent of $\{Y_n\}$ but with the same probability density function. In our study the random variables $\{Y_n\}$ follow a gamma distribution, because it was found that, during a scene, the number of cells per frame follow that distribution [11]. The shape and scale parameters of gamma disribution's density function are both taken to be 1. Also, the autocorrelation coefficient ρ is taken to be 0.9.

Table 1 summarizes the traffic characteristics of the sources used in our simulation study.

Traffic Source	Mean rate (Mb/s)	Peak rate (Mb/s)	Peak/Mean	std. deviation of bit rate
voice	0.02248	0.064	2.85	0.03055
video-AR(1)	3.89	11.38	2.93	1.735
video-AR(2)	2.08	6.57	3.16	1.061
video-DAR(1)	1.00	9.78	9.78	1.00

Table 1: Traffic source characteristics

4.2 Overview of Simulation Model

In our simulations, we model a single link of the network which is fed by a single buffer of an output-buffered switch. Cells arrive at the buffer from a number of traffic sources and are served in a FCFS manner by the link. Also, we consider that cells produced by video sources, arrive at a constant rate during the interframe interval (which is taken to be 33 ms - the NTSC standard period). The buffer is served at a deterministic rate, corresponding to the capacity of the link.

Measurement of the actual CLP

In our experiments, we measure the actual cell-loss probability by observing the traffic passing through the buffer for a period of time, and simply counting the number of cells arriving and the number of cells being dropped. But what should be the length of such a period so that we estimate, with reasonable confidence interval, the cell-loss rate?

We face that problem by considering a related measure of CLP, which is the probability to overflow the buffer during a *busy cycle*, where a busy cycle is called the evolution of the buffer's queue from the time it ceases to be empty until it becomes empty again. The successive busy cycles are approximately i.i.d. [19]. Let p denote that probability and n and S_n denote respectively the numbers of busy cycles and successful busy cycles, where we call a busy cycle successful if the buffer overflows. Then, the maximum-likelihood estimator of p is given by $\hat{p} = S_n/n$ with $\mathrm{E}\{\hat{p}\} = p$ and $\mathrm{E}\{S_n\} = np$.

The statistic S_n is binomially distributed but, if the sample size n is large we can use the normal approximation to the binomial distribution [18]. Then the 95% confidence interval for

p is obtained from

$$\frac{S_n}{n} - 1.96\sqrt{\frac{S_n(1-S_n/n)}{n^2}} \;<\; p \;<\; \frac{S_n}{n} + 1.96\sqrt{\frac{S_n(1-S_n/n)}{n^2}}$$

Thus, to obtain an estimate of p within the range $[p - p/a,\ p + p/a]$ with probability 0.95, we have to observe the switch for such a period of time, so that $S_n = (1.96a)^2$ successful busy cycles occur. In our simulations, we choose S_n to be 96, so that the error of a brute-force estimate of the actual cell-loss rate should not be greater than the 0.4 of its value.

Estimation Procedure's Implementation

As we have already mentioned, our estimation procedure is based on the brute-force estimates of cell-loss rate occurred in a number of virtual subsystems, which carry much less number of sources than the actual system. That cell-loss rates can be estimated accurately in a time that is orders of magnitude smaller than the time required to estimate the actual cell-loss rate. Recall that the "size" of each subsystem is represented by the number N of its sources, so the capacity and the buffer size of each virtual switch are equal to Nc and Nb respectively, where c and b denote the amounts of bandwidth and buffer space per source. But, what should be the optimal choice of those virtual subsystems? If we choose small values of N, this will result in estimators with small variance but which will contribute too large extrapolation errors. On the other hand, the choice of large values of N, will result in estimators with large variance. We do not yet have the answer to that problem, but trial and error gives values that work well.

We implement our estimation method by using many groups of subsystems, say K. The size of the subsystems in each group is constant, say $N_k,\ k = 1, \ldots, K$. Also, the subsystems of the same group are independent, and its number is denoted by $n_k,\ k = 1, \ldots, K$. While the network operates, we estimate the cell-loss rate in each virtual subsystem, by counting the cells arriving and the cells being dropped. Let $G^t_{i,k}(N_k)$ denote the cell-loss rate measured in virtual subsystem i, $i = 1, \ldots, n_k$, of size N_k, at time t. If 15 or more successful busy cycles occur in each subsystem of size N_k, then we calculate the mean value $\bar{G}^t_k(N_k)$ of $G^t_{i,k}(N_k)$, i.e., $\bar{G}^t_k(N_k) = \sum_{i=1}^{n_k} G^t_{i,k}(N_k)/n_k$. If less than 15 busy cycles occur in one subsystem of size N_k, then we don't use any of the measurements obtained from the subsystems of the same group. We choose the above threshold of busy cycles to be 15, so that the error of a direct measurement of the actual cell-loss rate should not be greater than the one order of magnitude of its value. Using the above criteria, we reduce the variance of the brute-force estimate of cell-loss rate occurring in a virtual subsystem of a specific size.

The cell loss rate $G(N)$ in a system of size N can be approximated by $AN^{\psi}\exp(-NI(c,b))$, where A and ψ are constants, and $I(c,b)$ is described in 3. This can be rewritten as the following linear equation

$$\log G(N) = \log A + \psi \log N - IN \tag{4}$$

So, at any time t, we can use regression analysis to find the values for the unknowns $\log A$, ψ and I which parameterise the best-fit linear regression of $\log \bar{G}_k^t(N_k)$ against $\log N_k$ and N_k, for those groups of subsystems that satisfy the above-mentioned requirements of $G_{i,k}^t(N_k)$. Then, we can compute the actual cell-loss rate and its variance from 4 using the obtained coefficients.

The goodness-of-fit is measured by the following quantity, which is called the *squared multiple regression coefficient*

$$Q = \frac{\sum \left(\widehat{\log \bar{G}_k} - \overline{\log \bar{G}}\right)^2}{\sum \left(\log \bar{G}_k - \overline{\log \bar{G}}\right)^2}$$

where $\overline{\log \bar{G}} = \mathrm{E}\{\log \bar{G}_k\}$, and $\widehat{\log \bar{G}_k}$ is computed from 4. The squared multiple regression coefficient must lie between 0 and 1, and the closer it is to 1, the better the fit. But, in our experiments, we cannot use that coefficient as an indicator of estimate's accuracy because averaging the cell-loss rate amongst virtual subsystems of the same size makes the use of this measure suspect.

4.3 Results of Simulation Experiments

We conduct four sets of simulation experiments in order to study the effectiveness of the proposed on-line estimation procedure, in terms of accuracy and timeliness, for realistic input traffic. Table 2 summarizes the configurations of our simulation experiments.

In the first set of experiments, we make 100 individual runs using 1024 voice sources as input traffic. The values of bandwidth and buffer space per source are taken to be $c = 25$Kb/s and $b = 0.1$Kb respectively, so the total buffer size of the switch is 100Kb and the load on the link is 0.899. In order to accurately measure the actual cell-loss rate, using the technique described in the previous subsection, each run simulated the behaviour of the switch for 215 sec. of real network time on the average. So, the value of the actual cell-loss rate estimator is measured to be $10^{-5.06}$ with a very small 95% confidence interval. Also, the predictive confidence interval of that estimator is $[10^{-4.56},\ 10^{-5.56}]$, i.e., the 95% of the

Expt Set	Input traffic	c (Mb/s)	b (Kb)	actual CLP	95% CI	predictive CI
A	1024 voice	0.025	0.1	$10^{-5.06}$	$[10^{-5.01},\ 10^{-5.11}]$	$[10^{-4.56},\ 10^{-5.56}]$
B	512 video AR(1)	4.16	1.0	$10^{-5.59}$	$[10^{-5.59},\ 10^{-5.59}]$	$[10^{-5.58},\ 10^{-5.60}]$
C	512 video AR(2)	2.16	0.1	$10^{-5.29}$	$[10^{-5.29},\ 10^{-5.29}]$	$[10^{-5.28},\ 10^{-5.30}]$
D	512 video DAR(1)	1.15	0.1	$10^{-5.12}$	$[10^{-5.09},\ 10^{-5.15}]$	$[10^{-4.87},\ 10^{-5.37}]$

Table 2: Simulation experiment configurations

estimates obtained from the 100 individual experiments are lying inside this region. To implement the estimation method described in the previous section, we consider the following 9 groups of independent virtual subsystems : 64 subsystems of size 16, 32 of size 32, 16 of size 64, 10 of size 96, 8 of size 128, 6 of size 160, 5 of size 192, 4 of size 224 and 4 of size 256.

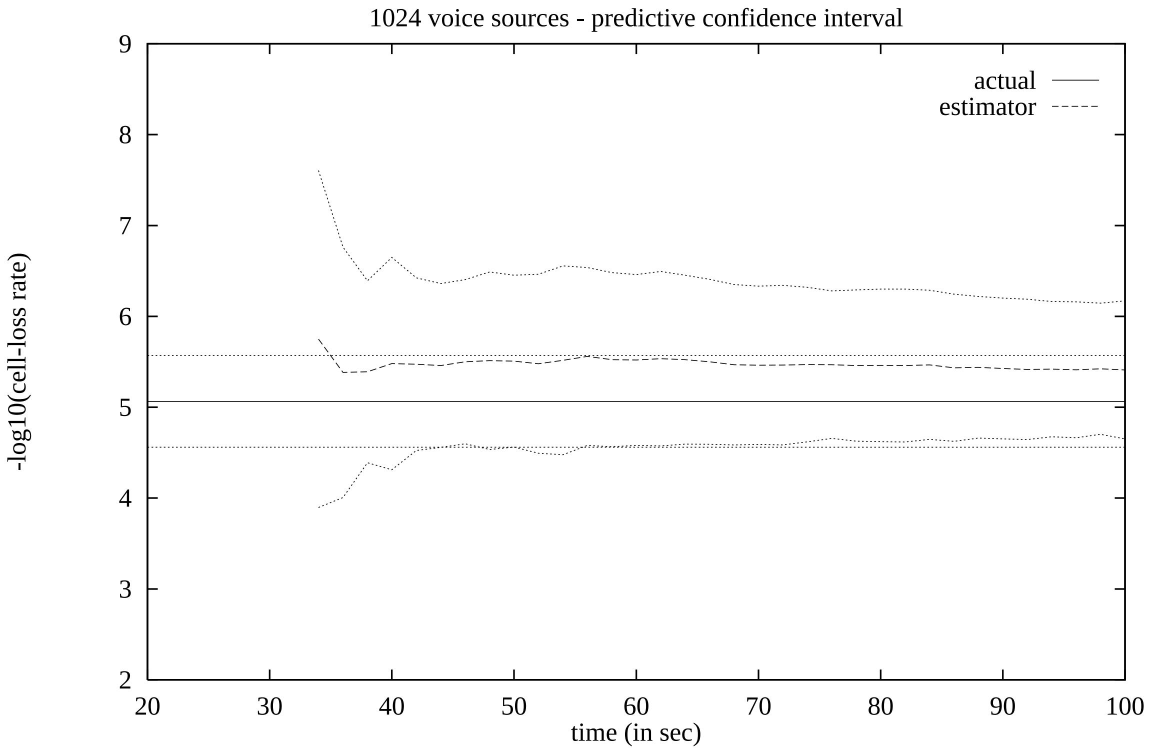

Figure 1: Convergence of cell-loss rate estimator, for 1024 voice sources

Figure 1 shows the actual cell-loss rate and our estimator over time for 1024 voice sources. Notice that the CLP estimator and its predictive confidence interval does not overlap with the actual CLP and its confidence interval. An obvious explanation is that this is showing us the extent of the inaccuracy of the large N asymptotic as an estimate. Another explanation is that the error is produced by ours two-term approximation of the $g_1(\cdot,\cdot,\cdot)$ function of 1. Recall that the $g_1(\cdot,\cdot,\cdot)$ function is $o(N)$ and is approximated by $\log N^{\psi} + \log A$.

From figure 1 we observe that the time required to obtain an estimate of cell-loss rate within an order of magnitude of the actual loss rate is about one minute. It is worth noting that, in 70% of the experiments there were no buffer overflows in the first 100 seconds. This shows how difficult it would have been to try to estimate the CLP directly.

In the second and the third set of experiments, we also make 100 individual runs using 512 AR(1) and 512 AR(2) video sources as input traffic respectively. In both sets, the utilization of the switch is 0.937 and the total buffer size of the switch is taken to be 1Mb in the second set and 100Kb in the third set. In order to obtain an accurate cell-loss rate estimate of the actual system, each run in the second set simulated the behaviour of the switch for 4368 sec. and each run in the third set simulated the switch for 3275 sec. of real network time on the average. Table 2 presents the cell-loss rates and their confidence intervals experienced in those sets of experiments. To implement our estimation method, we consider the following 8 groups of independent virtual subsystems : 32 subsystems of size 16, 16 of size 32, 10 of size 48, 8 of size 64, 6 of size 80, 5 of size 96, 4 of size 112, and 4 of size 128.

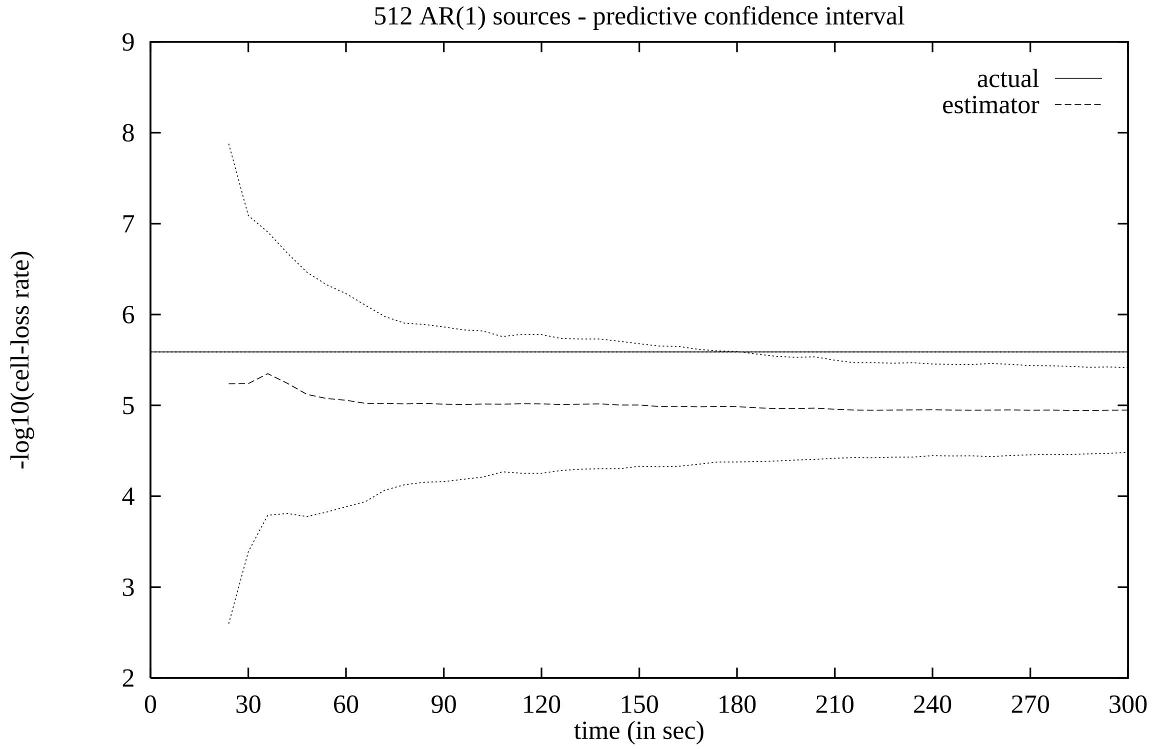

Figure 2: Convergence of cell-loss rate estimator, for 512 video AR(1) sources

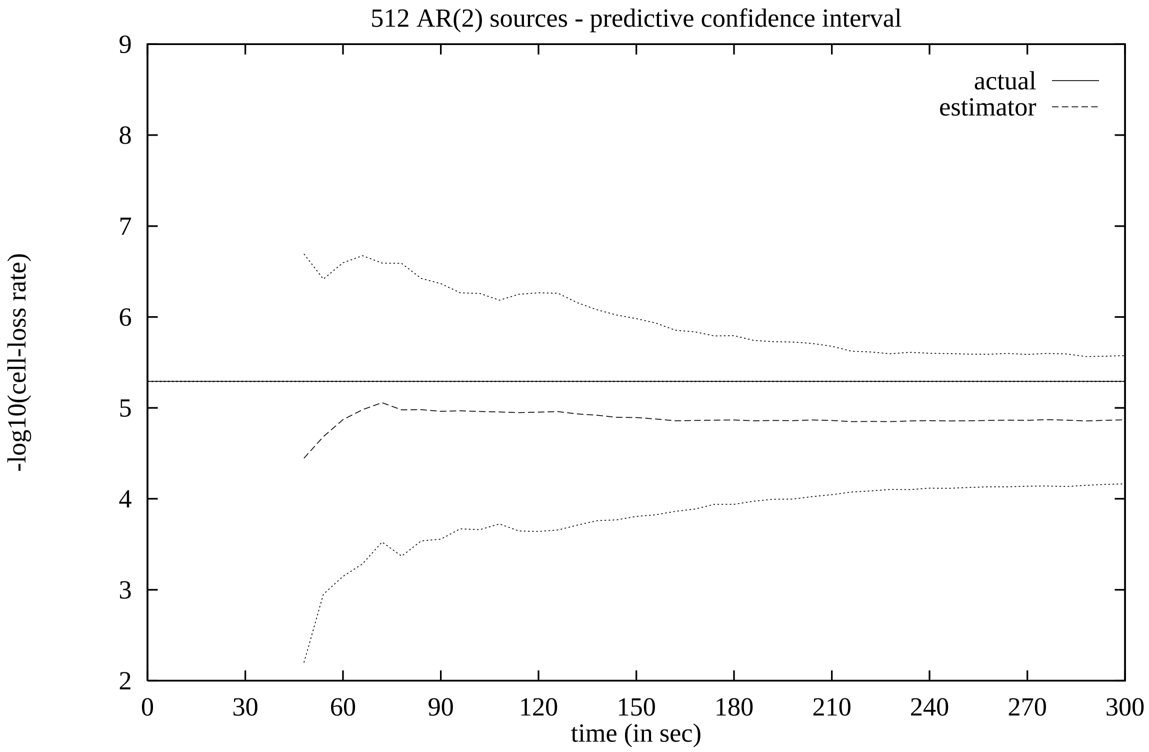

Figure 3: Convergence of cell-loss rate estimator, for 512 video AR(2) sources

Figures 2 and 3 show the time required to obtain a CLP estimate within an order of magnitude of the actual CLP for 512 AR(1) and 512 AR(2) video sources respectively. This is about 150 sec. for AR(1) sources and 240 sec. for AR(2) sources. In each set of runs, there were no buffer overflows, in any of the 100 experiments, in the first 300 sec.

Finally, we run 100 individual experiments using 512 DAR(1) video sources as input traffic. The total buffer size of the switch is taken to be 100Kb and the load on the link is 0.869. Each run simulated the behaviour of the switch for 5472 sec. of real network time on the average, yielding an actual cell-loss rate estimator equal to $10^{-5.12}$ with a very small 95% confidence interval. We use the same groups of virtual subsystems that have been used in experiments of AR(1) model.

Figure 4 shows that the time required to obtain an estimate of cell-loss rate within an order of magnitude of the actual loss rate is about 210 sec. for 512 video DAR(1) sources. Also, note that in only 21% of the experiments there were no buffer overflows in the first 300 sec.

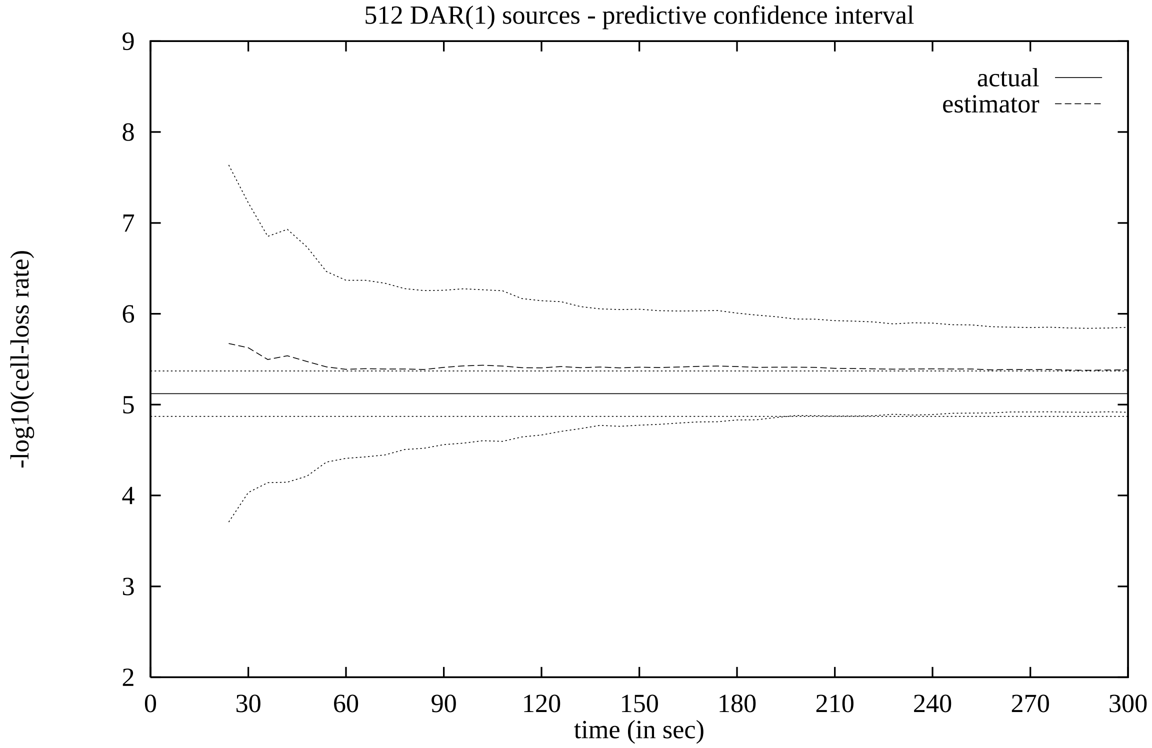

Figure 4: Convergence of cell-loss rate estimator, for 512 video DAR(1) sources

5 Summary and Directions for Further Research

An important issue in ATM networks is the accurate and timely estimation of the cell-loss probabilities that occur at various switches in the network, so that correct decisions can be made about whether or not to accept more traffic or how it should be routed. In this paper, we present a methodology for estimating the cell-loss rate of an ATM link in a reasonable short time. The method uses information that is obtained from the actual traffic flows and it is effective when there is a large degree of statistical multiplexing. The method can be used by Call Admission Control and Routing procedures in ATM networks. Also, it can be used for speeding up the simulation time for large systems.

Simulations were conducted which demonstrated the practicality of the method using the ON/OFF fluid, the AR(1), the AR(2) and the DAR(1) models as input traffic sources. A more exhaustive simulation study should be conducted using more realistic traffic sources, long sequences of real video data and heterogeneous input traffic. Another important issue that should be further investigated is the use of different approximations of formula 1 in ours on-line estimation technique.

References

[1] 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*).

[2] C. Courcoubetis and R. Weber. Buffer Overflow Asymptotics for a Switch Handling Many Traffic Sources. submitted for publication, 1994.

[3] C. Courcoubetis and R.R. Weber. Effective bandwidths for stationary sources. to appear in *Probability in the Engineering and Informational Sciences*, 1995.

[4] G. De Veciana and J. Walrand. Effective bandwidths: call admission, traffic policing and filtering for ATM networks. to appear in *Queueing Systems*, 1994.

[5] N.G. Duffield. Economies of scale in queues with sources having power-law large deviation scalings. Preprint, 1994.

[6] A.I. Elwalid and D. Mitra. Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. *IEEE/ACM Trans. Networking*, **1**:329–343, 1993.

[7] G. Fouskas. Experimental Evaluation of Call-Acceptance Strategies in ATM Networks. Technical report, Institute of Computer Science, FORTH, 1993. Available only in greek.

[8] M. Frater, P. Tan, and J. Arnold. Variable Bit Rate Video Traffic on the Broadband ISDN: Modelling and Verification. In *proceedings of ITC'94*, 1994.

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

[10] D. Heyman and T. Lakshman. Source Models for VBR Broadcast Video Traffic. In *proceedings of INFOCOM'94*, 1994.

[11] D. Heyman, A. Tabatabai, and T. Lakshman. Statistical Analysis and Simulation Study of Video Teleconference Traffic in ATM Networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 2(1):49, March 1992.

[12] P. Jacobs and P. Lewis. Stationary Discrete Autoregressive Moving Average Time Series Generated by Mixtures. *Journal of Time Series Analysis*, 4(1), 1983.

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

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

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

[16] M. Nomura, T. Fujii, and N. Ohta. Basic Characteristics of Variable Rate Video Coding in ATM Environment. *IEEE JSAC*, 7(5):752, June 1989.

[17] A. Simonian and J. Guilbert. Large deviations approximation for fluid queues fed by a large number of on-off sources. Preprint, 1994.

[18] K. S. Trivedi. *Probability & Statistics with Reliability Queueing and Computer Science Applications*. Prentice-Hall, 1982.

[19] J. Walrand. *An Introduction to Queueing Networks*. Prentice Hall, 1988.

[20] W. Whitt. Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues. *Telecommunication Systems*, **2**:71–107, 1993.