An Efficient Bandwidth Assignment Algorithm for Real-Time Traffic in ATM Networks

Alamin A. Belhaj, László Pap
Department of Telecommunications, Technical University of Budapest
Sztoczek 2, 1111 Budapest – Hungary
Email: belhaj@hit.bme.hu
Fax : (+36)-1-463-3266 Phone : (+36)-1-463-3266

Abstract

Asynchronous Transfer Mode is the chosen transport mechanism for future broadband networks. Although there is a huge number of mathematical models and experimental results, there are still many problems to be investigated. One of these problems is the way to provide guaranteed performance for real time traffic. In this paper we propose an algorithm for call admission control based on a simple and accurate estimate of the cell loss probability considering both cell scale and burst scale components. The loss probability is estimated from the asymptote of the tail probability by introducing a correction factor. Extensive numerical evaluations and simulations are made to evaluate the accuracy of the proposed algorithm.

Keywords

ATM, Real-Time Traffic, Cell Loss Probability, Tail Estimation, Bandwidth Assignment.

1 INTRODUCTION

Due to its great flexibility, Asynchronous Transfer Mode (ATM) is widely considered as the suitable technique to realise the full integration of transmission and switching for different kinds of services, such as voice, video, and data. These services are known to have different traffic characteristics (peak rate, mean rate, burst length, etc.), each having their own specific performance requirements (real-time, non-real-time, loss sensitive, etc.), which results into

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: 10.1007/978-0-387-35353-1_28

D. Kouvatsos (ed.), *Performance Analysis of ATM Networks*

conflict Quality of Service (QoS) requirements (cell loss, transfer delay, cell delay variation, etc.).

One of the main areas of ongoing research is the way to carry real-time traffic applications such as voice and video over ATM networks. In addition to the specific cell loss probability (CLP), real-time traffic requires a *strict limit* on the maximum end-to-end cell delay, beyond which arriving cells may be considered lost. Because of the this delay limit the size of the used buffer is usually very small.

In ATM networks, cells from different connections interact with each other at each switch. Without proper control, these interactions may adversely affect the network performance. One of the most important issues in providing guaranteed performance services is the choice of the cell service discipline at ATM switches (Zhang 1995). First-In-First-Out (FIFO) service discipline is the most used queuing discipline. FIFO can provide only an *average* performance for the aggregate traffic. For guaranteed Quality of Service (QoS) traffic, performance should be on a *per-connection* basis. Therefore, we emphasise using non-FIFO scheduling algorithms such as Weighted Round Robin (WRR) (Kang *et al.* 1995, Rampal *et al.* 1995). Along with scheduling, Call Admission Control (CAC) is required to provide guaranteed QoS.

An important goal behind ATM is the efficient utilisation of the resources due the ability of ATM to provide statistical multiplexing. Because of the small buffer size used with real-time traffic applications, many authors assume no statistical multiplexing gain can be achieved and therefore, they propose very conservative CAC algorithms such as peak rate allocation (Mitrou *et al.* 1996), or bufferless model schemes (Hsu *et al.* 1996). In this paper we insist that a reasonable statistical gain can be achieved although the buffer size is small. This is done by introducing an efficient CAC which is based on an accurate CLP estimation.

To develop an accurate estimate of the CLP of small finite buffer we used the idea of adjusting the tail distribution so as to provide a simple and accurate closed form formula for CLP. This is done through the development of a correction factor that makes the necessary adjustment. This method depends mainly on simple analytical formulation combined with simulation and numerical evaluations.

The algorithm that is being accepted in the literature to provide efficient link utilisation is by making best-effort traffic utilises the unused bandwidth left over by the real-time traffic applications (Tsang *et al.* 1996). In this scheme, traffic is classified into different classes, some of them for real-time traffic applications and others for best-effort applications.

In our belief, efficient utilisation of the link capacity means that : not only each slot is being utilised by the incoming traffic, but it means that : *each slot is utilised by the cell that should be served.* Because real-time traffic; 1) provides more revenue than data traffic, and, 2) is expected to comprise a large percentage of the network load in the future broadband networks, the

issue of efficient utilisation of resources assigned to real-time traffic becomes an important issue. This matter will be more crucial if the link is solely used to transport real-time traffic (i.e. no best effort traffic) (Tsang *et al.* 1996).

Therefore, we argue that the above mentioned algorithm is not optimum from the efficiency point of view unless provided with efficient algorithms for CAC and bandwidth assignment.

2 SYSTEM MODEL

2.1 Source Modelling

We model a single source using On-Off source model, where each source is modelled as a two-state Markov chain. On-Off source model can be considered as the output of the shaping device at the user interface, therefore it may represent an actual traffic characterisation.

To describe the On-Off sources we assume the source switches from On state to Off state with probability $1 - P_{11}$ and switch from Off state to On state with probability $1 - P_{00}$. Both On and Off duration are assumed to be independent and geometrically distributed. Therefore, the average On period (burst length) T_a is given by $T_a = 1/(1 - P_{11})$ and average Off period T_s is given by $T_s = 1/(1 - P_{00})$. The source activity factor p is given by $p = T_a/(T_a + T_s)$, so the burstiness β is given by $\beta = 1/p$. During On period the sources generates information with peak rate R_p bps while during Off period no information is generated. Accordingly, the mean rate m is given by $m = R_p p$. The ratio of the service rate to the peak rate of the source is denoted by M. For numerical experiments we consider five classes of real-time traffic applications of different traffic characteristics. Table 1 provides the parameter values of these applications.

2.2 Overall System Model

Our simplified model for a statistical multiplexer of an ATM network node consists of output buffered switches with nonblocking switch fabrics. Switch output nodes are organised as parallel FIFO's which share the output link's capacity V via Weighted Round Robin (WRR) scheduling algorithm. Scheduling have the effect of providing access to share of bandwidth, as if each service class had its own server at its given rate. Therefore, we assume that traffic is classified into K different classes according to the QoS requirements and traffic characteristics such as peak rate, mean rate and mean burst length.

Application	Parameter		
	Peak rate (Mbps)	Burstiness	Burst length (cell)
Voice	0.064	2.9	58
Videotel	2	2.5	212
MPEG1	1.856	4	2570
HDTV	30	4.6	12264
Image	2	23	2604

Table 1 Parameter values for different applications

Each class will be assigned its own buffer and individual share of the total link capacity to support its QoS. There are K admission controllers, one associated with each traffic class. A bandwidth assignment controller is linked to all admission controllers and the multiplexer. The bandwidth assignment controller, according to the network state, allocates a fraction C_i (we call it service rate of class i) from the total link capacity to each class i such that it satisfies the required QoS. This capacity will be modified over time as connections are dynamically set up and torn down. In this way, the analysis of heterogeneous traffic is simplified into the case of homogeneous traffic. The CAC problem then reduces to the analysis of the *single-class single queue* with its specific service rate.

In our analysis we assume that the delay constraint is provided by limiting the buffer size, therefore, only the cell loss probability will be the measure of performance, hence we assume that each class has a specific requirement cell loss probability denoted by ϵ_i. For each class i, we assume that N_i independent and identically distributed (i.i.d.) On-Off sources share a buffer of finite size B_i cells so that the maximum queuing delay is of τ_i seconds. The total load of each class is given by $\rho_i = N_i p_i / M_i$.

According to the non-work conserving WRR scheduling algorithm, let the cycle length to be W cells, and each queue has a quota of q_i cells. Then the service rate of each queue is given by:

$$C_i = \left(\frac{q_i}{W}\right) V. \tag{1}$$

In this paper we make the analysis for one class so we omit the subscript i.

3 TAIL ESTIMATION

The tail of the queue length distribution is defined as the probability that the steady state content of the infinite queue exceeds certain value B, i.e., $P_{tail} = Q(B) = \Pr(Q \geq B)$, where Q denotes the steady state buffer content. Generally it is difficult to evaluate the tail probability exactly, therefore, the tail distribution itself is usually approximated. To develop an approximate estimate of the tail we use the decomposition approach where we model the queuing system as the superposition of two separate components, namely the cell scale, $Q_{cell}(B)$, and burst scale, $Q_{burst}(B)$, components (Mignault *et al.* 1996). Therefore, we can write the tail as:

$$P_{tail} = Q_{cell}(B) + Q_{burst}(B). \tag{2}$$

3.1 Burst Scale Component

The burst scale component $Q_{burst}(B)$ is usually estimated using large buffer approximation, i.e. the burst scale component is approximated by the tail of an infinite queue. Because the evaluation of the tail is not simple, we develop a simple approximation to the tail for the burst scale component using discrete time queuing model. For discrete-time queuing systems, it has been observed that the steady state queue length distribution exhibits a geometrically distributed tail (Bruneel *et al.* 1996, Ishizaki *et al.* 1995, Sohraby 1993). That is, for sufficiently large buffer size B, we have:

$$Q_{burst}(B) \approx A \cdot z_o^{-B}. \tag{3}$$

Where z_o is called the dominant root and A is called the leading factor to be determined. The dominant root is relatively simple to evaluate. Sohraby (Sohraby 1993) gives several approximations for z_0 for different On-Off source models.

(a) The Leading Factor

The evaluation of the leading factor A is not as simple and needs some mathematical analysis. Different proposals have been introduced in the literature, here we review some of them and then we propose an approximation for the leading factor A.

- **Effective Bandwidth Approximation**
  In what is called effective bandwidth approximation, the leading factor is put equal to 1, i.e.,

$$Q_{burst}(B) = z_0^{-B}. \tag{4}$$

- **Sohraby Approximation**
  Based on the heavy traffic assumption, Sohraby simply put A equal to the total load $\rho = Np/M$, i.e.,

$$Q_{burst}(B) = \rho z_0^{-B}.$$

- **Ishizaki *et al.* Approximation**
  In a try to decrease the conservatism of Sohraby's method, Ishizaki *et al.* (Ishizaki *et al.* 1995) proposed a heuristic approximation of the leading factor A as follows:

$$A = A_{Soh}^{1/2} \cdot A_{Geo}^{1/2}, \tag{5}$$

  where A_{Soh} is the leading factor of Sohraby's proposal, and A_{Geo} is the geometric mean of the leading factors of the lower and upper bounds of the tail probability which are given in (Ishizaki *et al.* 1995) through considerably complex formulas.
- **Proposed Approximation**
  We thought of the leading factor A as the value of the tail, $\Pr(Q > B)$, when $B = 0$, that is, $A = \Pr(Q > 0)$. The probability $\Pr(Q > 0)$ can be related to the saturation probability that the input rate exceeds the service rate, i.e., the bufferless saturation probability $P_{zero,sat}$ as follows (Artiges *et al.* 1996):

$$P_{zero,sat} \leq \Pr(Q > 0). \tag{6}$$

  Hence, the probability $P_{zero,sat}$ can be considered as an approximation of $\Pr(Q > 0)$. Accordingly, in the next section we review the bufferless approximation, where we develop an estimate of the leading factor A using Bahadur-Rao formula.

3.2 Different Bufferless Approximations

For the bufferless model the saturation probability $\Pr(X > C)$ is given by:

$$P_{zero,sat} = \Pr(X > C) = \sum_{k=\lceil M \rceil}^{N} P_k, \tag{7}$$

where P_k is the binomial distribution given by:

$$P_k = \binom{N}{k} \cdot p^k \cdot (1-p)^{N-k}. \tag{8}$$

The average cell loss probability $P_{zero,loss}$ is given by:

$$P_{zero,loss} = \frac{1}{pN} \sum_{k=\lceil M \rceil}^{N} P_k \cdot (k - M). \tag{9}$$

It is clear that $P_{zero,sat}$ is easier to obtain than $P_{zero,loss}$. Equation(9) is usually called the Virtual Cell Loss Probability (VCLP).

While equation (7) and (9) can be exactly evaluated, they become too difficult for on-line calculations for large N, therefore, they are usually approximated. Several approximations have been used such as Gaussian approximation and large deviation approximations. Gaussian approximation may be very inaccurate, it may overestimate or underestimate the saturation probability. An other well known approximation is to apply large deviations approximations using Chernoff bound. In this case for the non-negative random variable X and positive constant C, we have for θ positive:

$$\Pr(X > C) \leq E[e^{\theta(X-C)}]. \tag{10}$$

where $E[x]$ is the expected value of x.

There is an optimum choice of the parameter θ which minimises (10) for given C. Using simple mathematical formulation the Chernoff bound for N i.i.d On-Off sources can be written as:

$$\Pr(X > C) \leq \left(\frac{p}{a}\right)^{Na} \left(\frac{1-p}{1-a}\right)^{N(1-a)}, \tag{11}$$

where, $a = \frac{C}{NR_p} = \frac{M}{N}$.

3.3 Bahadur Rao Bound

Chernoff's bound as given by the upper bound (11) often overestimates the saturation probability $\Pr(X > C)$. A better refined approximation can be obtained using Bahadur-Rao theorem (Hsu *et al.* 1996). If θ is given, then we can write the inequality in (11) with *equality* as follows:

$$\Pr(X > C) = \left(\frac{p}{a}\right)^{Na} \left(\frac{1-p}{1-a}\right)^{N(1-a)} \cdot \xi(\theta), \tag{12}$$

where $\xi(\theta)$ is a correction factor.

Bahadur and Rao used the idea to shift the most accurate point of the estimation from the region of the original mean value to the interesting region of

very small saturation probabilities. The shifted distribution can be approximated accurately by Gaussian distribution around its own mean. Accordingly, the correction factor $\xi(\theta)$ is given by:

$$\xi(\theta) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{N}\sigma\theta}, \tag{13}$$

where $\theta = \frac{1}{R_p} \ln\left(\frac{a\,(1-p)}{(1-a)\,p}\right)$ and $\sigma^2 = a(1-a)$.

Accordingly, we have the following approximation for the saturation probability using Bahadur-Rao theorem, $P_{BR,sat}$:

$$P_{BR,sat} = \frac{1}{\sqrt{2\pi}\sqrt{N}\sigma\theta} \left(\frac{p}{a}\right)^{Na} \left(\frac{1-p}{1-a}\right)^{N(1-a)}. \tag{14}$$

Bahadur-Rao approximation that is equivalent to the virtual cell loss probability equation (9) is given by:

$$P_{BR,loss} = \frac{P_{BR,sat}}{\theta N p R_p}. \tag{15}$$

Equations (14) and (15) are applicable to any system with number of independent sources. They are also accurate and relatively easy to calculate.

Accordingly, we propose to approximate the leading factor A by the Bahadur-Rao loss probability, $P_{BR,loss}$, as given by equation (15).

3.4 Cell Scale Component

One proposal to model the cell scale component is to use $M/D/1$ model. A more accurate model proposed is using $N * D/D/1$ queuing system (Fiche *et al.* 1994, Mignault *et al.* 1996).

Although the exact solution of $N * D/D/1$ queue is relatively straightforward, it may be not simple for fast calculations. Based on heavy traffic assumption using Brownian Bridge approximation method an approximation is provided as given in (Pitts *et al.* 1996). This approximation underestimates the cell loss for low utilisation. In (Fiche *et al.* 1994) a much better approximation is derived which is good for all traffic intensities. For the cell scale component $Q_{cell}(B)$, we use this approximation which is given by ;

$$Q_{cell}(B) \approx \frac{1-\rho}{ln(\rho)} \cdot exp\left(-B \cdot \left(\frac{2B}{N} + 1 - \rho - ln(\rho)\right)\right). \tag{16}$$

4 ACCURACY OF THE PROPOSED ESTIMATE OF THE TAIL

Now we will investigate the accuracy of the proposed tail by comparison with some exact and approximate tail distributions. Some results are shown in Figure 1 and Figure 2. For the exact tail calculations we present two different algorithms. The first is the algorithm of Anick, Mitra and Sohadni (Anick *et al.* 1982) which is based on fluid flow model, we denote it by AMS exact tail. The other algorithm is the one developed by (Choudhury *et al.* 1996) which is based on $\sum G/G/1$ queuing model, we call this $G/G/1$ exact tail. For $\sum G/G/1$ exact tail we do not make the numerical calculations, instead we reproduced the results directly from (Choudhury *et al.* 1996).

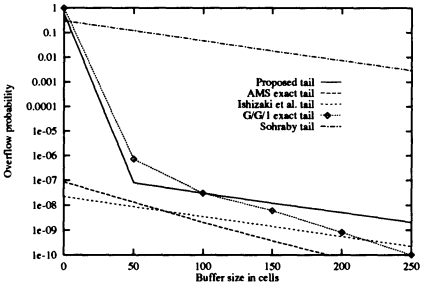

Figure 1 Accuracy of the proposed tail estimation with small load and small buffer size, $\rho = 0.3$, $M = 18.2$

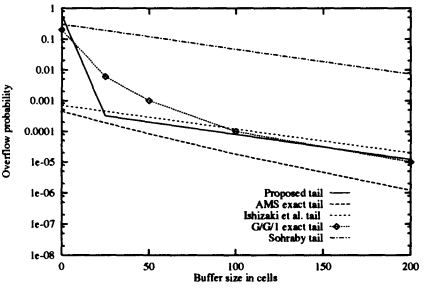

Figure 2 Accuracy of the proposed tail estimation with small load and small buffer size. $\rho = 0.3$, $M = 7.3$

In Figure 1 and Figure 2 we compare, for small load and small buffer sizes, the proposed approximation of the tail with the exact tail of AMS and G/G/1. The results for G/G/1 are respectively reproduced from Figure 7 and Figure 1 in (Choudhury *et al.* 1996). Also in these figures we include the tail approximations as proposed by Sohraby and Ishizaki *et al.* The source parameters for Figure 1 are as follows: $N = 60$, $T_a = 60$ cells, $T_s = 600$ cells, $\rho = 0.3$ (hence, $M = 18.2$). For Figure 2 the only change is that the number of sources is now 24 (hence, $M = 7.3$). From these two figures we see that the proposed tail approximately follows the G/G/1 exact tail in the buffer range indicated. Ishizaki *et al.* tail approximation does not differ much from the proposed tail. Sohraby tail approximation is very conservative compared to the all other tail estimations. Results with large buffer size have also been done and the proposed tail showed to be very accurate.

5 VALIDITY OF USING THE TAIL DISTRIBUTION AS AN ESTIMATE OF THE CELL LOSS PROBABILITY OF FINITE BUFFER

Usually the tail of the infinite buffer is used to approximate the CLP of the respective finite buffer queueing system. The question to be answered is how much accurate is that approximation (Belhaj *et al.* 1997)? The well known idea is that the tail distribution $P_{tail} = Pr(Q \geq B)$ of an infinite buffer configuration is *always* an upper bound to the respective finite buffer one, P_{loss}, see for example (Mitrou *et al.* 1996). On the other hand it is stated in (Bisdikian *et al.*, Mignault *et al.* 1996) that the above statement is true only for heavy traffic cases. Therefore the relation between the tail probability and CLP of the corresponding finite buffer should be investigated. We checked this matter by comparing the exact tail and other tail approximations with the exact CLP obtained by simulation.

In Figure 3 we give an example from a set of numerical examples that compare the cell loss approximated with the exact tail evaluated using AMS exact tail with simulation for different number of sources (i.e. load). We see clearly that the tail overestimates the cell loss at high loads and underestimates it at small loads.

A contradictory result is obtained in Figure 4 (the source parameter is same with Figure 2) where the tail overestimates the cell loss even the load is very small ($\rho = 0.3$). This result indicates that not only the load that affects the accuracy of the tail as an estimate for the cell loss probability, there may be other parameters which have also some effect. This matter will be investigated more in section 6.

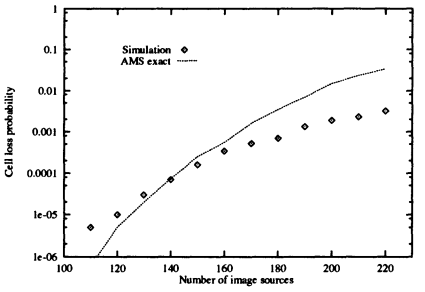

Figure 3 Comparison of the exact cell loss probability produced by simulation and the exact tail overflow probability. Image source, $M = 15.1$, $B = 100$

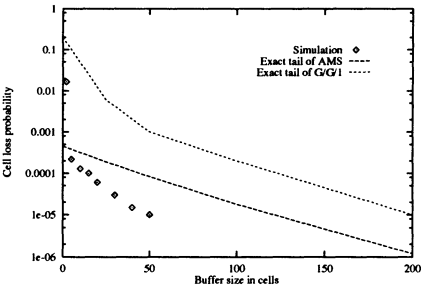

Figure 4 The tail overestimates the CLP even with small load. $\rho = 0.3$

6 CELL LOSS PROBABILITY OF FINITE BUFFER SYSTEM

To establish an accurate estimate of the cell loss probability of a finite buffer we start with the derivation of a Correction Factor (CF) to adjust the tail approximation that has been developed. Therefore, we write the CLP estimate, P_{loss}, as follows:

$$P_{loss} = CF(.) \cdot P_{tail}. \tag{17}$$

where CF(.) is a general function of the source and network parameters.

Since we want to relate the infinite buffer analysis to the respective finite buffer, we start with the simplest situation that is well known in the literature, namely the Markovian $M/M/1$ and $M/M/1/B$ systems, where the results governing them are well established.

We denote the correction factor that relates the infinite queue to the corresponding finite queue for the case of Markovian queues by Basic Correction Factor, CF_{Basic}, which can be approximated, for large buffer size, as follows:

$$CF_{Basic} = \frac{(1-\rho)}{\rho}. \tag{18}$$

In (Bisdikian *et al.* 1993), a similar result is provided also for $M/G/1/K$, and $Geo^{[x]}/D/1/K$ queuing models. The basic correction factor given by equation (18) is used in (Bruneel *et al.* 1996, Mignault *et al.* 1996) as the correction factor to adjust the tail. On the other hand, Takine *et al.* (Takine *et al.* 1994) proved that the CLP for discrete finite buffer is related to the tail probability as :

$$P_{loss} = \frac{(1-\rho)}{\rho} \cdot \frac{P_{tail}}{(1-P_{tail})}. \tag{19}$$

For small P_{tail} , equation (19) provides the same correction factor as CF_{Basic}.

6.1 Proposed Correction Factor

To have some insight to the effect of all parameters on the correction factor we investigated the effect of the source and network parameters on the accuracy of the CLP estimation. We considered, burst length, burstiness, load, buffer size, and ratio of service rate to the peak rate of the sources. In all situations we investigated the effect of each parameter on the CLP as estimated from the tail modified by CF_{Basic} compared to the simulation results.

In Figure 5 to Figure 8 we plot the cell loss probability using the tail approximation which is adjusted by CF_{Basic} compared to simulation. In these

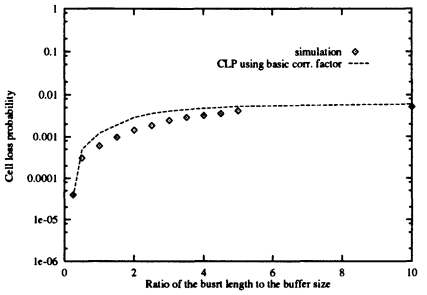

Figure 5 Effect of burst length on the CLP using CF_{Basic}. Videotel sources. $M = 20$, $\rho = 0.63$

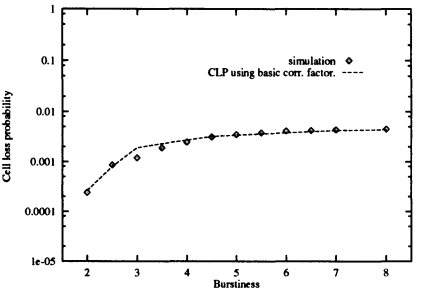

Figure 6 Effect of burstiness on CLP using CF_{Basic}. Videotel sources, $M = 20$, $\rho = 0.88$

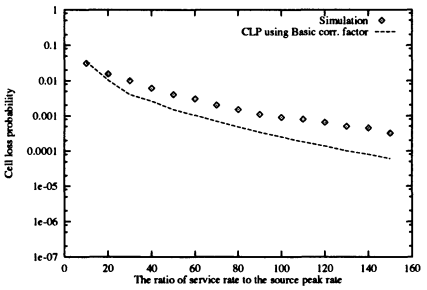

Figure 7 Effect of the ratio of service rate to the source peak rate on CLP using CF_{Basic}. Voice sources, $M = 40$, $B = 100$ $\rho = 0.88$

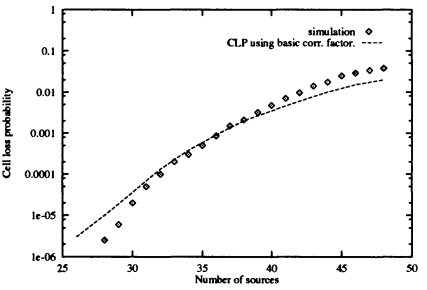

Figure 8 Effect of load on CLP using CF_{Basic}. Videotel sources, $M = 20$,

figures we checked the effect of burst length, burstiness, ratio of service rate to the peak rate and load. In each case the other parameters are kept fixed. With a careful investigation of these figures we see that for the case of burst length and burstiness the CLP estimate is approximately accurate. This means that the burst length and burstiness do not have a *clear* effect on the correction factor. On the other hand, for the load and ratio of service rate to peak rate ratio we see clearly that the cell loss estimate deviates from simulation as the parameter changes. From this we concluded that the load and the ratio of service rate to source peak rate both have some effect on the correction factor and therefore they should be put under more investigation.

After a wide number of trials we arrived to a conclusion that a suitable correction factor can be put in the form $(1 - \rho)M^{f(\rho)}/3.5\rho$, where $f(\rho)$ is a general function of the load. Refer to (Belhaj 1998) for a detailed derivation of this correction factor.

In order to get accurate estimate of the CLP, the exact values of $f(\rho)$ must be determined for different source and network parameters. One way is to

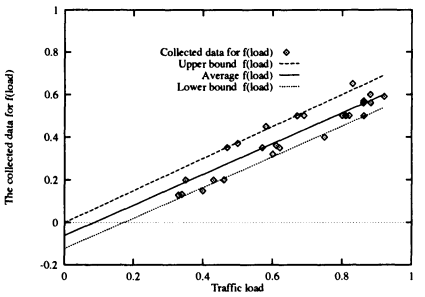

Figure 9 Collected data for $f(\rho)$ versus load

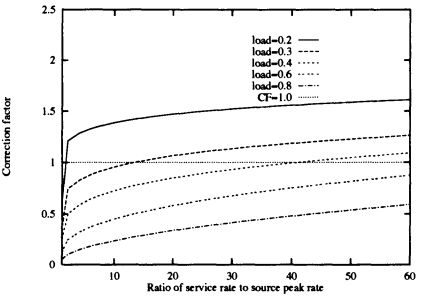

Figure 10 The proposed correction factor as a function of M for different load values

use tabulated values for different cases. Using tabulated values would require a considerable amount of memory capacity and some decision rule to select the value of $f(\rho)$ because all possible values could not be stored in the table. Another possibility is to calculate a large range of values of $f(\rho)$ and fit the data describing it using some criterion. This approach is the one we want to follow. Using comparison with simulation we made numerical experiments of more than 30 settings. In all these experiments we found the value of $f(\rho)$ that produces *exact* CLP estimate. The collected data is shown in Figure 9. To fit the data we made a linear regression (to simplify the calculation for the CLP) of the collected data and found that $f(\rho)$ can be represented by a linear function of the load as follows:

$$f(\rho) = 0.72\rho - .06. \tag{20}$$

Beside the function $f(\rho)$ given by equation (20), which we call average $f(\rho)$ we presented two other estimates of $f(\rho)$. The first we call lower bound $f(\rho)$, given by $f(\rho) = 0.72\rho - 0.12$, which is in tangent to the lower values of the collected data. The other proposal of $f(\rho)$ we call it upper bound $f(\rho)$ given by $f(\rho) = .72\rho$, which is in tangent to the upper values of the collected data. By comparison with simulation we found that using either the average $f(\rho)$ or the other two estimations does not change the cell loss estimation significantly. That is from the network dimensioning point of view using any of the estimations of $f(\rho)$ will give same results approximately.

Therefore, we can write the overall Proposed Correction Factor, $CF_{Proposed}$, as follows:

$$CF_{Proposed} = (1-\rho) \cdot \frac{M^{f(\rho)}}{3.5\rho}, \tag{21}$$

where $f(\rho)$ is given by equation (20) or one of the other two proposals.

By a closer look to into the developed correction factor as given by equation

(21) we could reach a conclusion that explains the relation of the tail and the CLP. Here, we have the result which necessities the correction to the statement that says the tail underestimates the loss for small load and overestimates it for high load as it has been emphasised by Bisdikian *et al.* (Bisdikian *et al.* 1993) and Mignault et al. (Mignault *et al.* 1996). We have now the result that states: *the tail may overestimate or underestimate the cell loss probability depending on the load and the ratio of the service rate to the source peak rate. For small values of load the tail usually underestimates the cell loss, and for high values of the load the tail usually overestimates the cell loss, but for moderate loads the tail may overestimate or underestimate the cell loss depending on the ratio of the service rate to the peak rate of the sources.* See (Belhaj 1998) for more explanation of this result. Finally, according to the correction factor given by equation (21), the overall estimate of cell loss probability is given by:

$$P_{loss} = (1-\rho) \cdot \frac{M^{f(\rho)}}{3.5\rho} \cdot P_{tail}. \tag{22}$$

7 SIMULATION AND NUMERICAL EVALUATIONS

In this section we give more results from the simulation and numerical evaluations that have been made to test the accuracy of the proposed estimate of the cell loss probability as given by equation (22). In this investigation, we compare with simulation, the proposed algorithm and four other on-line algorithms proposed in the literature. These are the algorithm proposed by Sohraby which is based on an approximate of the tail of a discrete time queuing model (Sohraby 1993), the algorithm of Ishizaki *et al.* which is based on the evaluation of the CLP using discrete time model (Ishizaki *et al.* 1995) , the algorithm of Lee and Mark which is based on an approximate tail derived using fluid flow model (Lee *et al.* 1995), and the bufferless approximation model as proposed by Hsu and Warland which is based on large deviation approximation (Hsu *et al.* 1996).

First we start with voice sources which are mainly characterised by the small peak rate and small burst length. In Figure 11 and Figure 12 we show the CLP versus buffer size and number of sources respectively. In these two figures the ratio of service rate to the source peak rate is 40. To check the results at higher values of M we present Figure 13 where $M = 156$. To test the accuracy of the algorithms when changing the parameter M, we include Figure 14, where the load is kept constant at 0.86 by changing the number of sources. From these figures related to voice sources we see clearly that the proposed algorithm is very accurate and predicts approximately the same results as that is given by simulation.

The second set of experiments are done with videotel sources. In Figure 15 and Figure 16 we present respectively the CLP versus buffer size and number of sources. Here, also with videotel sources we get very amazingly accurate

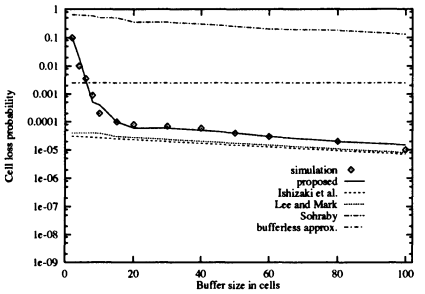

Figure 11 Cell loss probability versus buffer size. Voice sources. $M = 40$, $\rho = 0.7$.

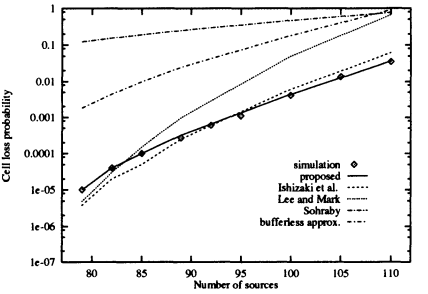

Figure 12 Cell loss probability versus number of sources. Voice sources. $B = 100$, $M = 40$

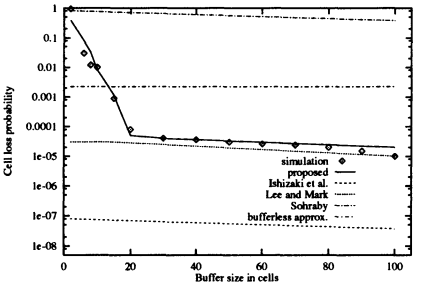

Figure 13 Cell loss probability versus buffer size. Voice sources, $M = 156$, $\rho = 0.82$.

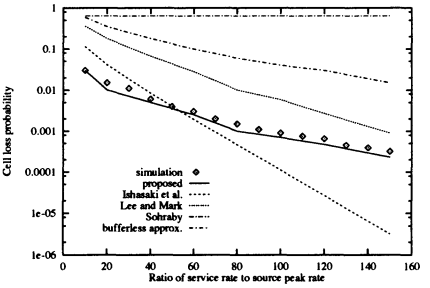

Figure 14 Cell loss probability versus M. Voice sources. $B = 100$, $\rho = 0.86$.

results using the proposed algorithm. Special look should be given to Figure 15 where M is small, we see here that all other algorithms provide very conservative prediction of the cell loss probability.

In the third set of experiments we consider image sources which are char-

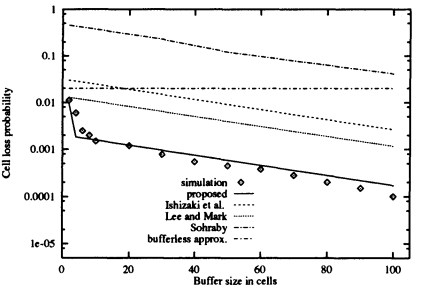

Figure 15 Cell loss probability versus buffer size. Videotel sources, $M = 5$, $\rho = 0.6$.

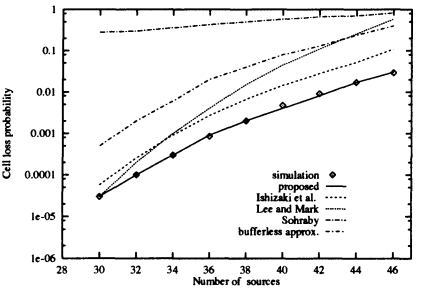

Figure 16 Cell loss probability versus number of videotel sources. $B = 100$, $M = 20$

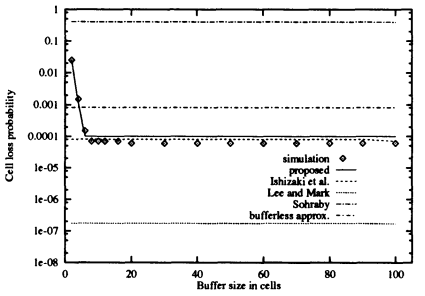

Figure 17 Cell loss probability versus buffer size. Image sources. $M = 15$, $\rho = 0.58$.

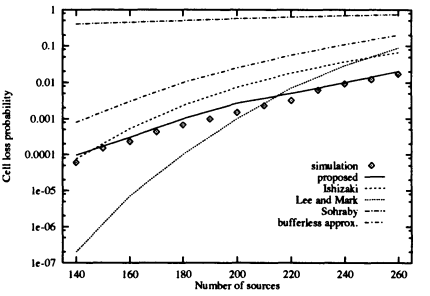

Figure 18 Cell loss probability versus number of sources. Image sources. $B = 100$, $M = 15$

acterised mainly with the very high burstiness (β=23). In Figure 17 and Figure 18 we see the CLP versus buffer size and number of sources respectively. From these figures we find that the proposed algorithm is also accurate. Other results for HDTV and MPEG are given in (Belhaj 1998).

8 SERVING REAL-TIME TRAFFIC EFFICIENTLY

The usual approach for the bandwidth assignment is that according to the state of the system each class is assigned a fixed bandwidth during the *whole* duration of the assignment period, and that bandwidth is made *available* to be used by the present connections of that class whatever the resulted QoS, i.e. the connections present of any class can use *all* the bandwidth assigned to that class. Now due to the idea of ATM, which is clear from the word *"Asynchronous"*, the connections of any class that are present at any time may consume bandwidth much higher than that is required to satisfy their QoS if they allowed to do so. This happens whenever the number of connections in any class falls below the number that should be served by the assigned bandwidth. This is an other possible cause of the inefficient utilisation of the bandwidth assigned to any class. The *ideal* solution to this problem is to allow the present connections at any time consume *just* the amount of bandwidth assigned to them according to their QoS requirements. It may be argued that this can be achieved by making the bandwidth reassignment whenever a connection is terminated or a new connection is accepted. This is not suitable because it makes the assignment too frequent, so that the time from the previous assignment is not enough to make the necessary calculations required for the overall bandwidth assignment operation. The scheme we propose to solve this problem is based on the idea to differentiate between the assigned bandwidth and the available bandwidth.

- **The Assigned Bandwidth**

We mean by the assigned bandwidth, the maximum bandwidth that the connections of a certain class can use during the assignment period. This bandwidth can not be assigned to any other class during this assignment period. The assigned bandwidth is determined according to an optimisation of some cost function. A connection will be accepted only if the assigned bandwidth is enough to serve this new connection with the present connections of the class, i.e the assigned bandwidth determines the CAC action. The assigned bandwidth is updated at the start of a new assignment period.

- **The Available Bandwidth:**
  By the available bandwidth we mean the bandwidth that can be used by the present connections of the class, and this should be just the necessary bandwidth to satisfy the required performance. The available bandwidth is determined when ever a new connection requests admission or a connection terminates. Therefore it is updated according to the connection activity (belhaj 1998). When a connection is accepted (according to the assigned bandwidth) the available bandwidth is increased by the amount just to satisfy all the present connections of the class. When a connection terminates the available bandwidth is re-evaluated accordingly. Therefore, connections of each class will not be served with over satisfied QoS. The difference between the assigned bandwidth and the available bandwidth is made available to be used by the best-effort traffic, and hence bandwidth utilisation is maximised.

9 NUMERICAL EVALUATIONS

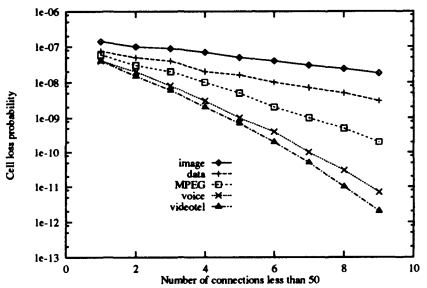

Figure 19 Cell loss probability versus number of connections less than 50

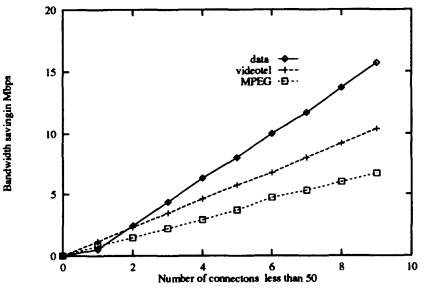

Figure 20 Bandwidth saving versus number of connections less than 50

Here we give an example to see how much the dynamics of the connections affect the CLP that the connections are served with. We assume that different classes have $\epsilon = 10^{-7}$. We assign a bandwidth to each class such 50 connections are served by the required CLP. Then we keep the assigned bandwidth constant and decrease the number of connections present in each class. For

the data source the parameters are, $R_p = 10Mbps$, $\beta = 10$ and $T_a = 330$. In Figure 19 we plot the CLP that the connection are served with versus the number of connections less than 50. We see clearly that the connections are served with over satisfied service and this increases as the connections are less bursty. For example if the number of videotel connections is decreased by 6 then the present connections will be served with 10^{-10} instead of 10^{-7}. In Figure 20 we see the saved bandwidth for the same situation as in Figure 19.

10 CONCLUSIONS

In this paper we introduced a very simple CAC that is suitable to provide guaranteed QoS for real time traffic. Scheduling disciplines other than FIFO together with CAC can provide the required performance. Peak rate allocation although is very simple it considerably limits the number of real time traffic connections. This may not be desirable for the network provider because real time traffic is expected to provide more revenue than best effort traffic. Therefore, bandwidth allocation for real time traffic should be based on algorithms that take into account the effect of statistical multiplexing for small buffer sizes. In this paper we proposed such a simple and efficient algorithm for CAC. Using simulation and numerical evaluations we compare several algorithms available in the literature and validate the accuracy of the proposed algorithm.

REFERENCES

D. Anick, D. Mitra, and M. Sondhi (1982) Stochastic theory of a data-handling system with multiple sources *Bell System Tech. J.*, 61:191–214, 1882.

D. Artiges and P. Nain (1996) Upper and lower bounds for multiplexing of multiclass Markovian On/Off sources, *Performance Evaluation*, 27-28:673–698, 1996.

A. A. Belhaj (1998) An Efficient Approach of Call Admission Control for Real-Time Traffic in ATM Networks. *Ph.D. Thesis, Telecommunication Department, Budapest Technical University, Hungary, 1998.*

A. A. Belhaj, L. Pap and T. V. DO. (1997) Statistical Call Admission Control for Real-Time Traffic in ATM Networks. In *Proc. ICCC'97, paper 50*, November 1997, France.

C. Bisdikian, J.S. Lew, and A.N. Tantawi, (1993) On the tail approximation of the blocking probability of single server queues with finite buffer capacity. In R.O. Onvural and I.F. Akyildiz, editors, *Queueing Networks with Finite Capacity*, pages 267–280. Elsevier, 1993.

H. Bruneel and S. Wittevrongel, (1996) An approximate analytical technique for the performance evaluation of ATM switching elements with burst routing. *Computer Networks and ISDN Systems*, 28:325–343, 1996.

G. L. Choudhury, D. Lucantoni, and W. Whitt, (1996) Squeezing the most out of ATM. *IEEE Transactions on communications* Vol. 22, No.2, pp. 203-217, February 1996.

G. Fiche, W. Lorcher, R. Veyland, and F. Oger (1994) Study of multiplexing for ATM traffic sources. *ITC'14* pp. 441–451, 1994.

I. Hsu and J. Walrand (1996) Admission control for multi-class ATM traffic with overflow constrains. *Computer Communication and ISDN Systems*, 28:1739–1751, 1996.

F. Ishizaki, T. Tetsuya, H. Terada, and T. Hasegawa (1995) Loss probability approximation of a statistical multiplexer and its application to call admission control in high-speed networks. In *Proc. IEEE Globecom'95*, pp. 417-421, 1995.

K. Kang, Y. Yoon, and C. Kim. (1995) A CAC scheme for heterogeneous traffic in ATM networks to support multiple QoS requirements. In *Proc. IEEE Globecom'95*, pp. 422-427, 1995.

H. Lee and J. W. Mark (1995) Capacity allocation in statistical multiplexing of ATM sources. *IEEE/ACM transactions on Networking*, 3(2):139–151, April 1995.

J. Mignault, A. Gravey, and C. Rosenberg (1996) A survey of straightforward statistical multiplexing models for ATM networks. *Telecommunication Systems*, 5:177–208, 1996.

N. Mitrou, K. Kontovasilis, and E. Protonotarios (1996) ATM traffic engineering for ABR service provisioning. *Telecommunication Systems*, 5(1-3):135–157, May 1996.

J. Pitts and J. Schormans, (1996) *Introduction to ATM, Design and Performance*. Jhon Wiley, 1996.

S. Rampal, D. Reeves, and D. Agrawa (1995) End-to-end guaranteed QoS with statistical multiplexing for ATM networks. In D. Kouvatsos, editor, Volume 1, *Performance Modelling and Evaluation of ATM networks*, Chapman and Hall, 1995.

K. Sohraby (1993) On the theory of general on-off sources with applications in high-speed networks. In *Proc. IEEE Infocom.*, pages 401–410, 1993.

T. Takine, B. Sengupta, and T. Hasegawa (1994) An analysis of a discrete-time queue for broadband ISDN with priorities among traffic classes. *IEEE Transactions on Communications*, 42(2/3/4):1837-1845, Feb./Mar./Apr. 1994.

R. Tsang, P. Keattithananant, T. Chang, J. Hsieh, and D. Du, (1996) Dynamic resource control for continuous media traffic over ATM networks. *Computer Communication*, 19:1092–1111, 1996.

H. Zhang (1995) Service disciplines for guaranteed performance service in packet-switching networks. *Proceeding of the IEEE*, 83(10):1374-1396, October 1995.