Diffusion Based Statistical Call Admission Control in ATM *

Erol Gelenbe
Department of Electrical and
Computer Engineering
Duke University
Durham, NC 27708-0291
erol@ee.duke.edu

Xiaowen Mang
Cascade Communications Corp.
Westford, MA 01886
xmang@casc.com

Raif Önvural
Allied Telesyn International
ATM System Development Laboratory
900 Perimeter Park, Suite B
Morrisville NC 27560

Abstract

We propose a call admission control (CAC) procedure for asynchronous transfer mode (ATM) networks using statistical bandwidth as the decision criterion, based on closed-form expressions that use diffusion models. This approach is computationally very efficient and easily implementable. It is related to Gaussian approximations previously proposed for CAC, though our expressions are derived from a more detailed representation of traffic and the buffer length process. The statistical bandwidth expressions we use take into consideration the users' cell loss requirements, their aggregate traffic characteristics, the available buffer size at the statistical multiplexers, and capture the interaction between individual traffic streams at the ATM multiplexer. Extensive numerical and simulation results are presented to evaluate the efficiency and adequacy of this CAC procedure. Comparison with existing methods such as the Equivalent Bandwidths and Gaussian Approximations indicate that our approach remains conservative with respect to cell loss compared to previously proposed CAC schemes, yet that it is more ecomomical in bandwidth allocation leading to larger admission regions both for homogenous and heterogenous traffic.

1 Introduction

ATM technology is expected to support a wide variety of services and applications, and to satisfy a range of user quality needs and network performance objectives. This technology allows flexibility in the choice of connection bit rates and enables the statistical multiplexing of variable bit rate traffic streams. Thus ATM provides a universal bearer service for B-ISDN networks, which can carry voice, data and video with the same cell transport arrangement. It is thought that much of the traffic in B-ISDN will be bursty, and this can be lead to poor performance. However, if the burstiness is adequately reflected in network management, considerable economy of network resources can be achieved. In a bursty and dynamic traffic environment, all users will not send traffic at the peak data rate at the same time. Therefore, one of the major challenges in traffic control is to achieve a statistical multiplexing gain while satisfying users' *Quality of Service (QoS)*.

An important functionality of traffic control in ATM is *Call Admission Control* (CAC). A connection can only be accepted if sufficient network resources are available to establish the connection end to end at its required quality of service. Also, the agreed QoS of pre-established connections in the network must not be adversely influenced by the new connection. As indicated in [Fel93] resource allocation in high speed networks will be an important

*This research was supported by IBM Network Products Division, but it only represents the views of the authors. The work was completed while Dr. Mang was a Ph.D. candidate in the Electrical and Computer Engineering Department at Duke University, and while Dr. Önvural was with IBM.

issue at least for the next few years and we need to design appropriate resource allocation schemes which maximize network income while guaranteeing users' desired quality of service. A uniform call admission control framework which not only includes an ideal resource allocation but also is robust and simple, is needed for ATM. The robustness requires that the control functions should be capable of accommodating future emerging and as yet unpredictable services. The need for simplicity means that CAC must be applicable in real-time, and implementable from a practical standpoint.

For the past several years various call admission control schemes have been proposed. Those that use bandwidth as the decision criterion include the equivalent bandwidth based schemes [RGN91, Reg94], the Gaussian approximation approach [Onv93, GG92, RGN91, Reg94] and heavy traffic approximations [Reg94, Soh92]. Comparisons of these schemes have been conducted by the authors of [Reg94, KME]. Their results indicate that they are all simple enough to be used in call admission control, but that they are all too conservative due to the limitations of their underlying assumptions. Realizing this conservatism, researchers have offered some refined schemes in the literature [GG92, Reg94]; however there is still much room for improvement. In [Onv93], it is pointed out that the bandwidth needed by a connection at a statistical multiplexer does not only depend on its own stochastic characteristics, but also on the characteristics of existing connections in the network.

We propose a call admission control procedure using statistical bandwidth as the decision criterion, based on closed-form expressions that use diffusion models. This approach is computationally very efficient and easily implementable. The statistical bandwidth expressions we use take into consideration the users' cell loss requirements, their aggregate traffic characteristics, the available buffer size at the statistical multiplexers, and capture the interaction between individual traffic streams at the ATM multiplexer. Extensive numerical and simulation results are presented to evaluate the efficiency and adequacy of this CAC procedure. We also thoroughly evaluate its sensitivity to traffic characteristics. Comparison with existing methods such as the Equivalent Bandwidths and Gaussian Approximations indicate that our approach remains conservative with respect to cell loss, yet that it is more ecomomical in bandwidth allocation leading to larger admission regions both for homogenous and heterogenous traffic.

The rest of this paper is organized as follows. In Section 2 we survey some related work on bandwidth allocation and CAC methods reported in the literature. We particularly identify some underlying constraints of existing schemes. In Section 3 we describe the diffusion approximation model, discuss the finite buffer approximation in detail and summarize the analysis for the infinite buffer approximation. We also report numerous simulation experiments used to evaluate the accuracy of its cell loss ratio predictions. In Section 4, we motivate and then present a new CAC procedure using two closed-form statistical bandwidth derived from the diffusion model with the finite buffer and infinite buffer approximation. In Section 5, we conduct extensive numerical and simulation experiments to validate our approach. Conclusions and suggestions for further work are presented in Section 6.

2 Related Work on Bandwidth Allocation

Although various approaches have been proposed to solve the problems related to bandwidth allocation and call admission control, the following two principles are the most popular ones - equivalent bandwidth and Gaussian approximation.

The equivalent bandwidth of a source is defined as the minimum bandwidth needed to carry the traffic generated by that source *in isolation* without violating the QoS requirements. The appeal of CAC schemes based on equivalent bandwidth concepts lies in their inherent simplicity where determining whether a given set of traffic sources can be accommodated without violating their QoS requirements reduces to comparing the sum of the equivalent bandwidths of individual sources to the link capacity.

However, despite its simplicity, the equivalent bandwidth principle is highly conservative when the buffer size is small or moderate [RGN91, Reg94, GG92]. The causes of this conservatism are that it is derived under the asymptotic regime where the product of buffer size and cell loss probability tend to zero, and it uses buffer overflow probability as the QoS requirement. Rege [Reg94] pointed out that simulations with various traffic models and combinations of link capacities and buffer sizes indicated that the buffer overflow probability is normally larger than the corresponding cell loss probability. Moreover, the ratio of the buffer overflow probability to the cell loss ratio was usually larger for higher link utilization levels. Typically it has been determined that the admissible region can be significantly larger than the equivalent bandwidth estimate [RGN91, Reg94, KME].

One of the most appealing equivalent bandwidth methods is derived in [RGN91] and further explored in [GG92].

It gives the equivalent bandwidth of source u for the buffer size B as:

$$c_u^e = R_u \frac{y_u^e - B + \sqrt{[y_u^e - B]^2 + 4Ba_u y_u^e}}{2y_u^e}, \qquad y_u^e = (-ln\epsilon)(\frac{1}{\beta_u})(1 - a_u)R_u$$

Where R_u is the source peak rate, a_u is the source activity factor and $1/\beta_u$ is the burst length.

Another well recognized approach is the Gaussian approximation based on the zero-buffer assumption. If the number of sources being multiplexed, N, is sufficiently large, the aggregate traffic can be approximated by a Gaussian process with mean rate $\lambda = \sum_{u=1}^{N} \lambda_u$ and variance $\sigma^2 = \sum_{u=1}^{N} \sigma_u^2$.

In contrast to the equivalent bandwidth concept, the statistical multiplexing gain that can be realized in a zero-buffer system is well studied. However for a non-zero-buffer system, the buffer's capacity to absorb traffic bursts is ignored in this Gaussian approximation. Also when N is small, the Gaussian approximation will not be valid. The resulting bandwidth can then be excessively conservative [GG92, Reg94].

Two important estimates have been calculated using the Gaussian approximation:

- Overflow Probability [Onv93, GG92]

$$Pr(overflow) = Pr(R(t) \geq C) \approx \frac{1}{\sqrt{2\pi}} e^{-\frac{(\lambda - C)^2}{2\sigma^2}} \tag{1}$$

- Upper Bound to Cell Loss Probability [Reg94]

$$Pr(loss) = \frac{E[(R(t) - C)^+]}{\lambda} \leq \frac{\sigma}{\lambda\sqrt{2\pi}} e^{-\frac{(\lambda - C)^2}{2\sigma^2}} \tag{2}$$

where $R(t)$ is the instantaneous cell arrival rate.

2.1 Hybrid Schemes

Because of the conservatism of the pure equivalent bandwidth based admission criterion in non-asymptotic regimes where they fail to account for the statistical multiplexing gain, in [RGN91, GG92] a compromise was made in such a way that the required bandwidth for N VBR sources equals:

$$\min\{C_e, C_g\}$$

where $C_e = \sum_{u=1}^{N} c_u^e$, $\quad C_g = \lambda + \alpha'\sigma$, $\quad \alpha' \simeq \sqrt{-2ln\epsilon - ln2\pi}$. C_g is the required bandwidth calculated from the Gaussian approximation of Equation (1).

Another refinement described in [Reg94] uses a highly non-linear function of the individual equivalent bandwidths to determine the admissibility of the given set of sources. The following underlying assumption was imposed in this refined scheme: a traffic stream with peak rate R_u, mean rate λ_u and equivalent bandwidth c_u^e feeding into a buffer of size B is equivalent to an On-Off stream with peak rate c_u^e and mean rate λ_u feeding into a buffer of size zero. This approach does yield a larger admissible region; however experimental results [Reg94] show that it can be over optimistic under certain conditions when buffer size is small leading to higher cell loss ratios than the user's desired QoS requirement. Thus in this paper we focus on an efficient and conservative CAC scheme.

3 Finite Buffer Diffusion Model for ATM Statistical Multiplexers

ATM performance prediction is difficult because the aggregate arrival process is a superposition of traffics with diverse characteristics, and the arrival rates among successive time intervals are highly correlated, and also because the measures of interest include very small probabilities or cell loss ratios. Current techniques are approximations whose validity is limited to certain classes of traffic and which are accurate for a limited range of parameters (including asymptotic cases). Various approaches which have been proposed to evaluate the performance of ATM multiplexers include direct studies of the discrete processes, and their continuous approximations such as flow fluid and diffusion approximations. However, beyond the issue of accuracy of models with respect to specific traffic characteristics, another important factor which limits analytical techniques is the numerical complexity of calculations

for performance measures such as cell loss ratio estimates. The advantage of diffusion approximations is that they can provide very simple formulae for performance metrics which can be computed in quasi closed form with very limited numerical computation. This makes them particularly attractive candidates for use in CAC where a fast real-time decision is needed. In the next section we shall see that these diffusion estimates can be remarkably accurate for commonly used traffic models using "On-Off" sources and their generalizations.

Diffusion approximations are continuous approximations to the discontinuous arrival and service processes in queuing models. They have long been used in queueing theory to model traffic and services. The advantage is that they will generally result in computationally more tractable models of performance for more detailed traffic representations, than what can be obtained from a direct study of the corresponding discrete processes. In the past, two different approaches to diffusion approximations for queueing models have been proposed. In both cases whenever the queue length is non-zero and the maximum buffer capacity has not been attained, the queue length distribution is approximated by solving a partial differential equation. However the two methods differ according to the choice of boundary conditions. The simpler one uses reflecting boundaries [Kob74a, Kob74b, KR93] so that no probability mass accumulates at the boundaries. Clearly this approach will not be totally satisfactory if the boundaries themselves are very important to the process being modeled. The more sophisticated approach is based on the "instantaneous return process" [Gel75, GP76, Dud86] which combines the partial differential equation formulation for the process *strictly* inside the boundaries, with a discrete state-space model at the boundaries themselves [Gel75]. This leads to a more accurate model of the queueing behavior of the system when the load is low, or when the queue length is close to the maximum value allowed by a finite buffer.

Diffusion approximations require that the first two moments of the interarrival and service times be known. These can be directly deduced from measurements or from traffic models, such as the "On-Off" model often used in the literature [HL86] or superpositions of homogenous or heterogenous traffic. The diffusion approximation approach we take for an ATM multiplexer buffer of size B, considers a random process $\{X(t),\ t \geq\ 0\}$ to represent the buffer contents. In the open interval]0,B[(excluding the two boundaries) it is a continuous random variable with probability density function $f(x,t)$ defined as:

$$f(x,t)dx = Pr[x \leq X(t) < x + dx], x \in]0, B[, \tag{3}$$

while at the boundaries we have:

$$m(t) = Pr[X(t) = 0], M(t) = Pr[X(t) = B]. \tag{4}$$

The parameters for the diffusion process inside in]0,B[are the "drift" or instantaneous average rate of change:

$$\mu = \lim_{\Delta t \to 0} \frac{E[X(t+\Delta t) - X(t) | X(t) \in]0, B[\,]}{\Delta t} \tag{5}$$

and the instantaneous variance of the change in X(t):

$$\alpha = \lim_{\Delta t \to 0} \frac{Var[X(t+\Delta t) - X(t) | X(t) \in]0, B[\,]}{\Delta t} \tag{6}$$

Since the service time is constant due to the fixed length of the cells being transmitted, α will only depend on the variance of interarrival times. Assuming time-independent traffic characteristics, let the mean aggregate cell arrival rate to the buffer be λ and the multiplexer cell transmission rate be C, both given in cells per second, so that $\mu = \lambda - C$.

In the instantaneous return process model, when queue length reaches the lower boundary of the interval at $x = 0$, it remains there for a random length of random time which we denote h. This time clearly represents a period when the buffer is empty, and it ends as soon as a cell arrives to the multiplexer. At that time, say some time τ, the process $X(t)$ will jump from $X(\tau)\ = 0$ to $X(\tau^+)\ =\ +1$. Similarly for the upper boundary at $x = B$ where the random time spent at the boundary will be denoted by H, while the jump of the queue length process will be from the value B to the value $B-1$ representing the end of a service or transmission epoch for a cell, resulting in a decrease of buffer length by 1. This behavior results in the following system of equations for the ATM multiplexer queue length process as derived in [Gel75] for $t \geq 0$:

$$-\frac{\partial f(x,t)}{\partial t} - \mu \frac{\partial}{\partial x} f(x,t) + \frac{\alpha}{2} \frac{\partial^2}{\partial x^2} f(x,t) + \frac{m(t)}{E[h]} \delta(x-1) + \frac{M(t)}{E[H]} \delta(x-B+1) = 0 \tag{7}$$

$$\frac{dm(t)}{dt} - \frac{m}{E[h]} + \lim_{x \to 0+} \int f(x,t)dx = 0, \tag{8}$$

$$\frac{dM(t)}{dt} - \frac{M}{E[H]} - \lim_{x \to B-} \int f(x,t)dx = 0, \tag{9}$$

where $\delta(x)$ is the Dirac Delta function. Also the probabilities must sum to 1:

$$m(t) + M(t) + \int_{0+}^{B-} f(x,t)dx = 1. \tag{10}$$

These equations have a simple interpretation. Equation (7) represents the motion of the queue length process in the interval $]0, B[$, and the effect of the jumps of the process $X(t)$ from 0 and B into the interval. On the other hand (8) represents the depletion of the probability mass $m(t)$ at the lower boundary due to the jumps to $+1$ at the end of the holding time at the lower boundary, as well as the flow of probability mass from inside the interval $]0, B[$ towards the lower boundary. Equation (9) has a similar interpretation. The above equations may be solved directly in steady-state by dropping the dependence on t [Gel75], to obtain:

$$f(x) = \begin{cases} \Phi\ [1\text{-}e^{\gamma x}], & 0 < x \leq 1 \\ \Phi\ [e^{-\gamma}\text{-}1]e^{\gamma x}, & 1 \leq x \leq B-1 \\ \Phi\ [e^{\gamma(x-B)}\text{-}1]e^{\gamma(B-1)}, & B-1 \leq x \leq B \end{cases} \tag{11}$$

with m and M the probability masses at 0 and at B, respectively, in stationary state being:

$$m = -\mu E[h]\Phi, \qquad M = -\mu E[H]\Phi e^{\gamma(B-1)} \tag{12}$$

where $\gamma = \frac{2\mu}{\alpha}$, and

$$\Phi = \frac{1}{(1-\mu E[h]) - (1+\mu E[H])e^{\gamma(B-1)}} \tag{13}$$

In order to make use of these expressions we need to determine the parameters μ, α, $E[h]$ and $E[H]$ from the arrival and service characteristics of the ATM multiplexer.

3.1 Calculation of $E[h]$ and $E[H]$ for the Finite Buffer Model

In general the distributions for the residence times of moderately complex finite capacity queueing models at the upper and lower boundaries 0 and B are unknown. Their characterization can be quite complex and depends on both the arrival process, the buffer size, and the service process. Thus we will have to calculate $E[h]$ and $E[H]$ in a heuristic but plausible manner. If the arrival process can be approximated by a Poisson process with arrival rate λ it follows that $E[h] = \lambda^{-1}$. Since the arrival traffic to an ATM multiplexer is made up of many superposed sources, when the number of sources is large this approximation may be acceptable. In our simulations it turns out that this heuristic for $E[h]$ slightly underestimates the actual value for superposed "On-Off" sources.

Recall that the time for transmitting one cell is C^{-1}. Now assume that at instant t the transmission of a cell begins and that $X(t) = B-1$. At some instant $t+Z$ before $t+C^{-1}$ another arrival occurs so that now $X(t+Z) = B$. Then H, the random variable representing the holding time at the upper boundary, has the following distribution:

$$Pr[H \leq v] = Pr[\frac{1}{C} - Z \leq v | Z \leq \frac{1}{C}] = \frac{Pr[\frac{1}{C} - Z \leq v \ and \ Z \leq \frac{1}{C}]}{Pr[Z \leq \frac{1}{C}]} \tag{14}$$

We make the simplifying approximation that the arrival process is Poisson of rate λ so as to complete the computation, on the basis that it is justified when the arriving traffic results from the superposition of many independent sources. Then $Pr[Z \leq \frac{1}{C}] = 1 - e^{-\frac{\lambda}{C}}$, and

$$Pr[\frac{1}{C} - Z \leq v \ and \ Z \leq \frac{1}{C}] = Pr[\frac{1}{C} - v \leq Z \leq \frac{1}{C}] = e^{-\frac{\lambda}{C}}[e^{\lambda v} - 1]. \tag{15}$$

Thus

$$Pr[H \leq v] = \frac{e^{\lambda v} - 1}{e^{\frac{\lambda}{C}} - 1}, \tag{16}$$

with density function $f_H(v) = \frac{\lambda e^{\lambda v}}{e^{\frac{\lambda}{C}} - 1}$, for $0 \leq v \leq \frac{1}{C}$, and $f_H(v) = 0$ elsewhere. We can now derive the estimate for the average holding time at the upper boundary:

$$E[H] = \int_0^{\frac{1}{C}} v f_H(v) dv = \frac{\frac{1}{C}}{1 - e^{-\frac{\lambda}{C}}} - \frac{1}{\lambda}. \tag{17}$$

Of course, the fist and second moments of the interarrival times are also needed in order to compute the density function $f(x)$ and the probability masses m and M. However, these moments will be available from the precise traffic characteristics we use are discussed in Section 3.

3.2 Estimating the cell loss ratio

The long run cell loss ratio L is the proportion of cells lost at the entrance to the multiplexer due to buffer overflow, to total cells arriving to the multiplexer. It is the primary measure of interest in this study and it needs to be estimated both *accurately* and in a *conservative* manner. Thus what is needed is in fact a tight upper bound, rather than a relatively accurate value which may underestimate L. Clearly cells will be lost only when the buffer is full, i.e. when buffer length has attained size B, in which case all the arriving cells will be lost. Thus the cell loss ratio in steady state may be written as $L = \lim_{t \to \infty} M(t) Pr[\, N(t, t+H) \geq 1 \mid X(t) = B \,]$, where $N(t, t+H)$ is the number of arrivals in the open interval $(t, t+H)$. If the arrival process is stationary in time and independent of buffer size, in steady state the cell loss ratio is: $L = M.Pr[\, N(t, t+H) \geq 1 \,]$.

There are several difficulties with using this expression when one deals with real traffic, including the issue of estimating H and the probability of the number of arrivals in the interval when the buffer is full. However we do know that $H \leq \frac{1}{C}$. Thus we have found that L^+ given below is a useful upper bound which yields cell loss ratio values which are within the same order of magnitude as the value measured from simulation with various forms of "On-Off" traffic:

$$L \leq L^+ = M.Pr[N(t, t + \frac{1}{C}) \geq 1]. \tag{18}$$

where we have used the fact that H is always less than $\frac{1}{C}$. The quality of this estimate is tested by simulation with a wide variety of "On-Off" traffic models, as shown in the simulation results we present.

The expression (18) is based on modeling an ATM multiplexer as a finite capacity queue. This leads to our Finite Buffer Diffusion Cell Loss Estimate (FBDCLE):

$$L_{FB} = \Psi e^{\frac{2(B-1)}{\alpha}\mu} Pr[R(t) \geq C] \tag{19}$$

where $R(t)$ is the instantaneous arrival rate and $\Psi = \frac{-\mu E[H]}{(1-\mu E[h]) - (1+\mu E[H]) e^{\frac{2(B-1)}{\alpha}\mu}}$. Note that $[R(t) \geq C]$ and $[N(t, t + \frac{1}{C}) \geq 1]$ are equivalent for time independent (stationary) traffic.

Another approach can be taken where the analysis assumes that buffer size is infinite, but that the cell loss ratio is given by:

$$L^+ = Pr[X \geq B] \frac{E[(R(t) - C)^+]}{\lambda} \tag{20}$$

where X is the stationary diffusion approximation queue length for the infinite buffer model and (as before) λ is the aggregate average arrival rate. We will not go into the details of the analysis but just point out that $Pr[X \geq B]$ can be computed via a system of equations similar to (7), (8), without equation (9) for the probability mass at the upper boundary, and without the last term in (7) related to the jump to $x = B - 1$. In addition, we use the following lower bound from the GI/G/1 queue to estimate $E[h]$ [Med91]: $E[h] \geq \frac{1}{\lambda} - \frac{1}{C}$

This leads to the Infinite Buffer Diffusion Cell Loss Estimate (IBDCLE):

$$L_{IB} = \Upsilon e^{\frac{2B}{\alpha}\mu} \frac{E[(R(t) - C)^+]}{\lambda} \tag{21}$$

where

$$\Upsilon = \frac{1}{1 - \mu E[h]} [1 - e^{-\frac{2\mu}{\alpha}}] \frac{\alpha}{2\mu}. \tag{22}$$

3.3 Accuracy of the Diffusion Cell Loss Estimate for Superposed "On-Off" Traffic

Much of the work on ATM traffic analysis and cell loss estimates is based on the "On-Off" traffic model and on the superposition of such traffic streams [HL86]. Thus it is of particular interest to evaluate the accuracy of our DCLE (diffusion estimate) for this specific class of practically useful models. In order to do so, we will first derive the appropriate traffic parameters to be used in the diffusion approximation. Then the distribution of the number of arrivals in a given time duration will be computed so as to be used in calculating the DCLE value L.

Consider first a single user u whose traffic follows a simple "On-Off" behaviour. This user u either sends traffic into the network at a constant peak rate R_u during the "On" period, or it sends no traffic at all during the "Off" period. The following notation describes this traffic model:

- R_u – peak traffic rate during the "On" period, $T_u = 1/R_u$;
- θ_u^{-1} – average length of the "Off" period;
- β_u^{-1} – average length of the "On" period;
- $a_u = \theta_u/(\beta_u + \theta_u)$ – source activity.

The duration of the successive on and off periods are assumed to be independent, so that the cell arrival process from a single such source is a *renewal process*. The cell interarrival time will be denoted by Y_u, and let $F_u(x) = Pr[Y_u \leq x]$ so that [HL86]:

$$F_u(x) = [(1 - \beta_u T_u) + \beta_u T_u (1 - e^{-\theta_u (x - T_u)})]U(x - T_u) \tag{23}$$

where $U(x)$ is the unit step function. The Laplace-Stieltjes transform (LST) of the interarrival time density is given by:

$$\tilde{f}(s) = \int_0^\infty e^{-sx} dF_u(x) = [1 - \beta_u T_u + \beta_u T_u \theta_u/(s + \theta_u)]e^{-sT_u} \tag{24}$$

The mean cell arrival rate of cells from source u is then:

$$\lambda_u = -1/\tilde{f}'(0) = 1/(T_u + \beta_u T_u/\theta_u) = a_u/T_u = a_u R_u \tag{25}$$

Let $A_u(t)$ denote the number of arrivals of cells of user stream u in the interval [0,t). Then the squared coefficient of variation of the interarrival time from source u is [CL66, HL86]:

$$c_u^2 = \frac{Var[Y_u]}{E^2[Y_u]} = \frac{Var[A_u(t)]}{E[A_u(t)]} \tag{26}$$

which leads to [HL86]:

$$c_u^2 = \frac{1 - (1 - \beta_u T_u)^2}{(\beta_u T_u + \theta_u T_u)^2}. \tag{27}$$

Since $E[A_u(t)] = \lambda_u t$, we can write(26) as:

$$\lim_{t \to \infty} \frac{Var[A_u(t)]}{t} = \lambda_u \frac{Var[Y_u]}{E^2[Y_u]} = \lambda_u c_u^2 \tag{28}$$

Now if the total arrival process to the ATM multiplexer results from the superposition of N uncorrelated "On-Off" sources of renewal type as discussed above, $A(t)$ the resulting counting process $A(t) = \sum_{u=1}^N A_u(t)$ has the obvious properties:

$$E[A(t)] = \sum_{u=1}^N E[A_u(t)], Var[A(t)] = \sum_{u=1}^N Var[A_u(t)] \tag{29}$$

and $E[A(t)] = \sum_{u=1}^N \lambda_u t,\ Var[A(t)] = \sum_{u=1}^N \lambda_u c_u^2 t$.

Let $D(t, t+\tau)$ denote the number of departures in an interval $[t, t+\tau)$ when the queue is non-empty. Note that if the multiplexer queue is non-empty, then the service or emptying process at the queue is independent of the arrival process. Thus we have:

$$E[X(t + \Delta t) - X(t) | X(t) > 0] = E[A(t + \Delta t) - A(t)] - E[D(t + \Delta) - D(t)] \tag{30}$$

and

$$Var[X(t+\Delta t) - X(t) | X(t) \in]0, B[\,] = Var[A(t+\Delta t) - A(t)] + Var[D(t+\Delta) - D(t)] \tag{31}$$

so that

$$\mu = \lim_{\Delta t \to 0} \frac{E[X(t+\Delta t) - X(t) | X(t) \in]0, B[\,]}{\Delta t} = \sum_{u=1}^{N} \lambda_u - C, \tag{32}$$

$$\alpha = \lim_{\Delta t \to 0} \frac{Var[X(t+\Delta t) - X(t) | X(t) \in]0, B[\,]}{\Delta t} = \sum_{u=1}^{N} \lambda_u c_u^2. \tag{33}$$

We now have all the parameters needed by the diffusion model described in Section 2 when it is used for superposed "On-Off" traffic sources, and can use it to calculate the DCLE formula given in (17).

In order to calculate the DCLE, the probability distribution $Pr[N(t, t+\frac{1}{C}) = l], \quad l \geq 0$ must be obtained. In order to do so, we will consider the general case of arrival traffic composed of multiple "On-Of" sources of K different *types*. Each source of the same type will have the same set of parameters, and N_k will be the number of k-type sources, each with the same peak traffic rate R_k, activity a_k. Here we use the subscript k to denote a user type, rather than the subscript u to denote an individual user. The total number of users or sources is then $N = \sum_{k=1}^{K} N_k$. The average arrival rate of cells will then be:

$$\lambda = \sum_{k=1}^{K} a_k N_k R_k \tag{34}$$

Now let $Z_k(t)$ be the random variable denoting the number of sources of type k which are "On" at some time t. Since the sources are independent and stationary we have for large enough t that:

$$Pr[Z_1(t) = n_1, ..., Z_K(t) = n_K] = \prod_{k=1}^{K} \binom{N_k}{n_k} a_k^{n_k} (1-a_k)^{N_k - n_k} \tag{35}$$

On the other hand for small enough $1/C$:

$$N(t, t+\frac{1}{C}) = [Z_1(t)R_1 + ... + Z_K(t)R_K]/C, \tag{36}$$

so that:

$$Pr[N(t, t+\frac{1}{C}) \geq 1] = Pr[Z_1(t)R_1 + ... + Z_K(t)R_K] \geq C], \tag{37}$$

which can be computed from the distribution (35).

For homogeneous traffic, i.e. when all sources are of just one type, we simply have $K = 1$ and:

$$Pr[N(t, t+\frac{1}{C}) \geq 1] = 1 - \sum_{n_1=1}^{int(C/R_1)} \binom{N_1}{n_1} a_1^{n_1} (1-a_1)^{N_1 - n_1}. \tag{38}$$

3.4 Comparison of the DCLE with simulations

In this section we present the numerical and simulation results to evaluate the accuracy of our approach. The validation of our new diffusion model is focused on the comparison of the cell loss probability predicted by the DCLE and that obtained by simulations for a wide variety of "On-Off" traffic models. In our simulations, the runs were independently replicated 20 times, and each run included the transmission of 10^7 cells. Confidence intervals are calculated using the *Student–t* distribution with a 98% confidence level, so that the simulation results are of sufficiently high statistical quality.

Analytical results from the diffusion model for DCLE and simulation results are compared in Figures 1 to 4. Figures 1 and 2 show cases with homogeneous sources. In Figure 1 the cell loss ratio, as calculated from the DCLE, is plotted and compared with the value measured from the simulations, as a function of buffer size B for different loads. Here the load is defined as the ratio of total incoming traffic rate to outgoing link capacity: $\sum_{u=1}^{N} \lambda_u / C$. The high speed On-Off sources we use here are very bursty and the ATM multiplexer link is of high capacity ($C = 150\ Mb/s$).

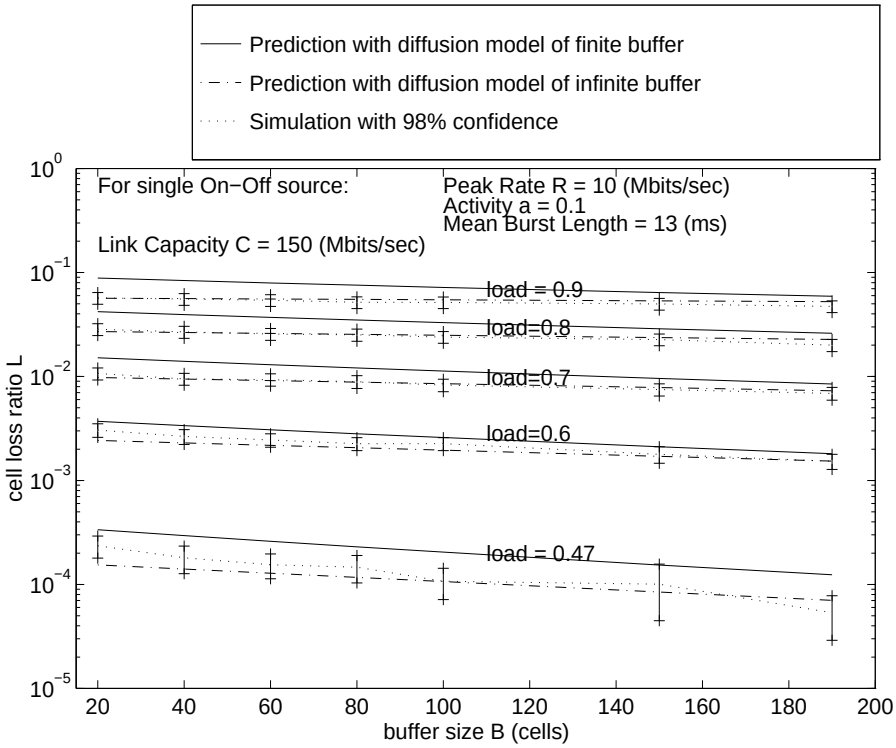


Figure 1: Cell loss probability vs. buffer size: comparision of simulation and DCLE for homogeneous sources under varying load (load = aggregate mean arrival rate /link capacity).

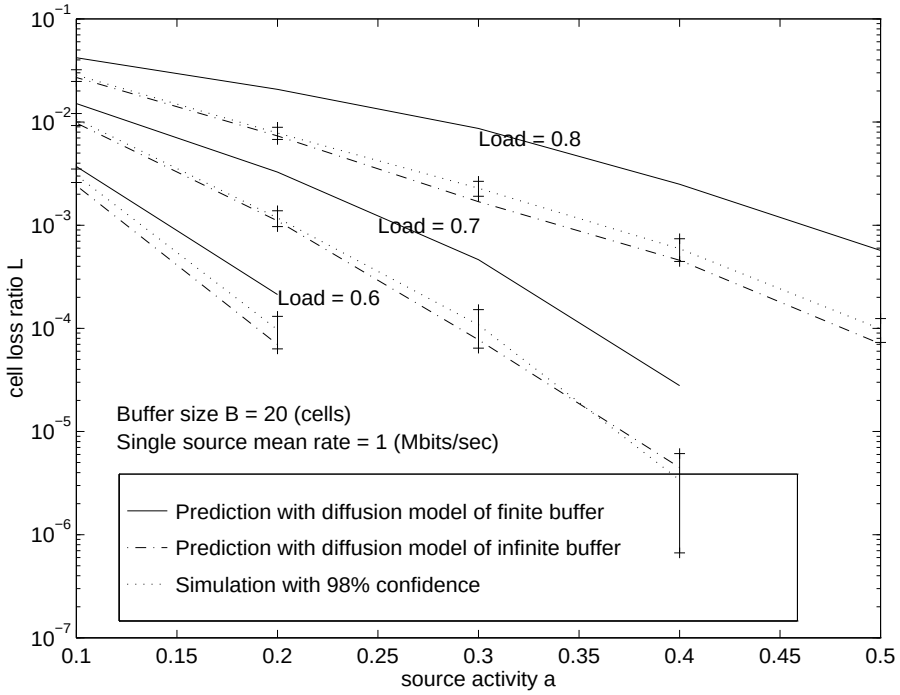


Figure 2: Cell loss probability vs. source activity (burstiness): comparision among simulations and analytical approach using DCLE for the homogeneous sources under variant load (load = aggregate mean rate /link capacity).

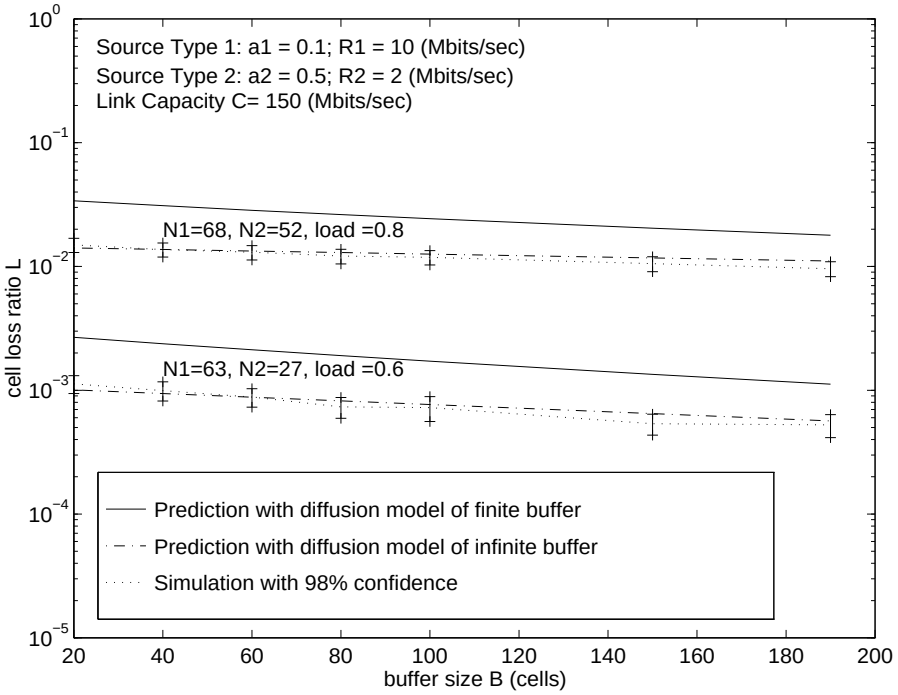


Figure 3: Cell loss probability versus buffer size: comparison between simulation and DCLE for heterogeneous sources with varying load (load = aggregate mean rate /link capacity).

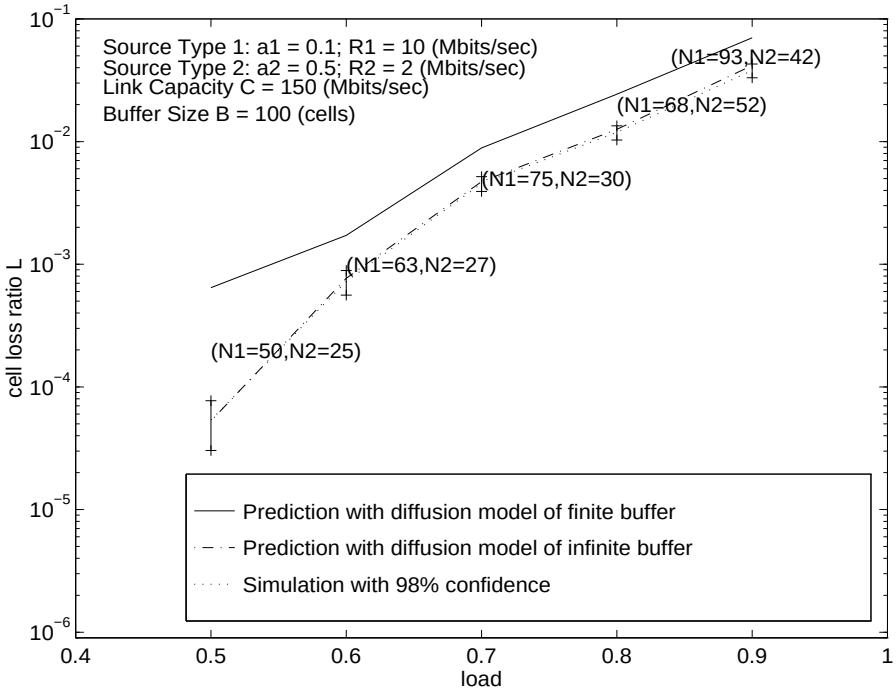


Figure 4: Cell loss probability versus load: comparision between simulation and the DCLE for heterogeneous sources (load = aggregate mean rate /link capacity).

Excellent agreement between the model predictions and the simulations is also observed in Figure 2 where we vary the source burstiness while keeping the mean source arrival rate constant at $\lambda_u = 1(Mbits/sec)$ with fixed buffer size of $B = 20$ cells.

Figures 3 and 4 compare DCLE with simulation under heterogeneous traffic where we have chosen two types of sources – more bursty sources with $a_u = 0.1$ and less bursty sources with $a_u = 0.5$. Let N_1 and N_2 denote the number of sources with $a_1 = 0.1$ and $a_1 = 0.5$ respectively, and $N = N_1 + N_2$. In Figure 3 we show matched results of simulations and the DCLE diffusion model under different combinations of N_1 and N_2. In Figure 4 the cell loss probability is plotted versus traffic load for a given buffer size. The simulation results, together with their confidence intervals, show a very promising agreement with our DCLE cell loss value. We see that generally the DCLE slightly overestimates the cell loss ratio, which means that if we use this approach in Call Admission Control it will tend to provide accurate and conservative results, which is the first requirement in any CAC procedure.

4 CAC Using the Diffusion Model: A Statistical Bandwidth Approch

In this section we introduce our CAC Algorithm based on the Diffusion Cell Loss Estimates. We will refer to it as a Statistical Bandwidth based algorithm because it uses information of the multiplexed traffic sharing a common link in the derivation of diffusion approximation based cell loss estimates.

In order to apply (19) and (21) to CAC, it is highly desirable to have a closed-form expression of the statistical bandwidth. The objective of having a closed-form expressions are two-fold:

- They require little computation and are therefore suitable for real-time decisions for acceptance or rejection of call/connection requests.
- They also simplify bandwidth management procedures. From the network management viewpoint, it is very useful to know how much bandwidth is needed for current users and how much is left for potential new calls. Using statistical bandwidth the network manager can easily monitor the bandwidth status on the link so as to accommodate other services such as ABR.

Since Equations (19) and (21) are composed of both rational terms and exponential terms, it is not possible to obtain the value of C that yields a prespecified cell loss ratio in closed form. Simplifications must be made which are acceptable only if they are conservative with respect to the resulting bandwidth allocation, and this can be achieved by using upper bounds for some terms in the original equations. Equations (19) and (21) are the steady-state solution of a stationary process under the condition that $\lambda < C$ or $\mu < 0$. It is easy to prove that:

$$\Psi e^{\frac{-2}{\alpha}\mu} < 1, \qquad \Upsilon < 1.$$

From (1), (2), (19) and (21) we have:

$$L_{FB_{bound}} = e^{\frac{2B}{\alpha}(\lambda - C)} \frac{1}{\sqrt{2\pi}} e^{-\frac{(\lambda - C)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}} e^{\frac{2B}{\alpha}(\lambda - C)} e^{-\frac{(\lambda - C)^2}{2\sigma^2}} \tag{39}$$

$$L_{IB_{bound}} = e^{\frac{2B}{\alpha}(\lambda - C)} \frac{\sigma}{\lambda\sqrt{2\pi}} e^{-\frac{(\lambda - C)^2}{2\sigma^2}} = \frac{\sigma}{\lambda\sqrt{2\pi}} e^{\frac{2B}{\alpha}(\lambda - C)} e^{-\frac{(\lambda - C)^2}{2\sigma^2}} \tag{40}$$

Then:

$$L_{FB_{bound}} > L_{FB}, \qquad L_{IB_{bound}} > L_{IB}$$

Also if B=0, Equations (39) and (40) become Equations (1) and (2).

Let L^* be the desired cell loss ratio. We therefore can derive following two statistical bandwidths C_{df_1} and C_{df_2}.

C_{df_1} denotes the statistical bandwidth obtained from the diffusion model of a finite capacity queueing system - FBDCLE. From (39) we can easily obtain the following quadratic equation:

$$\mu^2 - 2\delta\mu + 2\sigma^2\varpi_1 = 0$$

where $\delta = \frac{2B}{\alpha}\sigma^2$ and $\varpi_1 = ln(L^*\sqrt{2\pi})$. Moreover since $\mu = \lambda - C < 0$ we have:

$$\mu = \delta - \sqrt{\delta^2 - 2\sigma^2\varpi_1}$$

and

$$C_{df_1} = \lambda - \delta + \sqrt{\delta^2 - 2\sigma^2\varpi_1} \tag{41}$$

$\mathbf{C_{df_2}}$ denotes the statistical bandwidth obtained from the diffusion model of an infinite capacity queueing system - IBDCLE. Similarly to C_{df_1}, from (40) we have:

$$C_{df_2} = \lambda - \delta + \sqrt{\delta^2 - 2\sigma^2\varpi_2} \tag{42}$$

where $\varpi_2 = ln(L^*\lambda\sqrt{2\pi}) - ln(\sigma)$

We notice that C_{df_1} and C_{df_2} have the same expression except for the difference of $ln(\lambda) - \frac{1}{2}ln(\sigma^2)$ between ϖ_1 and ϖ_2. Since α and μ can be calculated using the peak rate R_u, the mean rate λ_u, and the burst length $1/\beta_u$, it follows that C_{df_1} and C_{df_2} are functions of these three standard parameters. For the given buffer size B, the bandwidth requirement can then be easily determined by Equations (41) or (42).

Call Admission Control Procedure The procedure we propose is based on the fact that a connection will only be established if there is enough bandwidth available on every intermediate link along the selected path which will carry the traffic of this connection. To do so, for each link along the path we use an information vector

$$\mathbf{I} = \{\sum_{u=1}^{N}\lambda_u, \sum_{u=1}^{N}\sigma_u^2, \sum_{u=1}^{N}\lambda_u c_u^2\}$$

which contains the status of current connections on each corresponding link. Let C_{df} denote the statistical bandwidth we will use in the CAC function, C_l be the link capacity and

$$\mathbf{U} = \{\lambda_U, \sigma_U^2, \lambda_U c_U^2\}$$

be the information vector of a new request. We thus propose the following CAC procedure:

For each link along the selected path:

1. Update **I⇐I+U**
2. Calculate C_{df}
3. If $C_{df} \leq C_l$, *Accept*
4. If $C_{df} > C_l$, *Reject* and *Restore* **I⇐I-U**

5 Evaluation of Statistical Bandwidths

In this section we will evaluate our statistical bandwidths by comparing them to the hybrid scheme proposed in [RGN91, GG92]. In that scheme, the required bandwidth is determined by $\min\{C_e, C_g\}$.

In general, bursty traffic (such as VBR traffic) can be characterized by a bit rate which changes randomly between different constant values according to the activation and disactivation of different services. Thus for most purposes, the combination of such services can be considered as a superposition of simpler "On-Off" type of sources. An "On-Off" source is alternatively active and inactive. An active period is also known as a burst and an inactive period a silence. Our numerical and simulation studies in this section will use three well known "On-Off" traffic models. Let u stand for a particular user or connection u.

- **Two-state MMPP** is a Poisson process whose rate is determined by a two-state Markov chain. In state S_1 the mean holding time is $1/\beta_u$ with mean arrival rate Λ_{1_u}, while in state S_2, the mean holding time is $1/\theta_u$ with mean arrival rate Λ_{2_u}, and $\Lambda_{1_u} > \Lambda_{2_u}$.

  The mean arrival rate of cells is then:

$$\lambda_{u_{(MMPP)}} = \frac{\beta_u}{\beta_u + \theta_u}\Lambda_{1_u} + \frac{\theta_u}{\beta_u + \theta_u}\Lambda_{2_u}$$

The variance of cell arrival rate is:

$$\sigma^2_{u_{(MMPP)}} = \frac{\beta_u}{\beta_u + \theta_u}\Lambda^2_{1_u} + \frac{\theta_u}{\beta_u + \theta_u}\Lambda^2_{2_u}$$

The squared coefficient of variation of the interarrival times of cells in Two-state MMPP model is [Onv93]:

$$c^2_{u_{(MMPP)}} = 1 + \frac{2\beta_u\theta_u(\Lambda_{1_u} - \Lambda_{2_u})^2}{(\theta_u\Lambda_{1_u} + \beta_u\Lambda_{2_u} + \Lambda_{1_u}\Lambda_{2_u})(\beta_u + \theta_u)^2}$$

- **IPP** is a special case of two-state MMPP where $\Lambda_{2_u} = 0$.
  The mean arrival rate of cells is:

$$\lambda_{u_{(IPP)}} = \frac{\beta_u}{\beta_u + \theta_u}\Lambda_{1_u}$$

  The variance of cell arrival rate is:

$$\sigma^2_{u_{(IPP)}} = \frac{\beta_u}{\beta_u + \theta_u}\Lambda^2_{1_u}$$

  The squared coefficient of variation of the interarrival times of cells in Two-state IPP model is:

$$c^2_{u_{(IPP)}} = 1 + \frac{2\beta_u\Lambda_{1_u}}{(\beta_u + \theta_u)^2}$$

- **Two-state IDP** is similar to IPP except that a constant traffic intensity $\Lambda_{1_u} = R_u$ (peak rate) is generated at state S_1. The mean arrival rate of cells is:

$$\lambda_{u_{(IDP)}} = \frac{\beta_u}{\beta_u + \theta_u}R_u$$

  The variance of cell arrival rate is:

$$\sigma^2_{u_{(IDP)}} = \frac{\beta_u}{\beta_u + \theta_u}R^2_u$$

  The squared coefficient of variation of the interarrival times of cells is then [CL66, HL86]:

$$c^2_{u_{(IDP)}} = \frac{1 - (1 - \beta_u T_u)^2}{(\beta_u T_u + \theta_u T_u)^2}$$

The corresponding statistical bandwidth are represented by $C_{df_{1_{(DP)}}}/C_{df_{2_{(DP)}}}$ for IDP, $C_{df_{1_{(PP)}}}/C_{df_{2_{(PP)}}}$ for MMPP and $C_{df_{1_{(IP)}}}/C_{df_{2_{(IP)}}}$ for IPP respectively.

The traffic parameters we use in our numerical and simulation studies are given in Table 1. For fairness and convenience of comparison, we choose $1(Mbits/sec)$ as the mean traffic rate for all types (Type 1 to Type 4) and all kinds (IDP, IPP and MMPP) of traffic. For any given type of traffic the mean rate, peak rate and burst length are the same for all traffic models: IDP, IPP and MMPP. Also they all have an average burst length of 306 cells.

5.1 Conservatism and Efficiency

In order to show that the bounds we derive do not lose too much efficiency with respect to the original FBDCLE/IBDCLE in call admission control, we compare the resulting admission regions of original estimates and the corresponding statistical bandwidth. Here we use the term "admission region" to refer to the maximum number of users which can be carried on a link with given capacity, buffer size and desired QoS requirement. The admission regions identified by FBDCLE/IBDCLE are calculated through exhaustive search. The comparisons are shown in Figure 5 where the IDP traffic model is used. In addition, the admission regions obtained by using the known [RGN91, GG92] $min\{C_e, C_g\}$ approach are also shown in the figures. Various parameters have been used in order to demonstrate the consistency of comparison outputs. The ordinate represents the maximum number of burstier sources (a=0.1), and the abscissa represents the maximum number of less bursty sources.

The results of comparisons (see Figure 5) are very favorable. The admission regions identified by FBDCLE/IBDCLE and C_{df_1} and C_{df_2} are almost the same. Moreover for given buffer size of 1 or 2 times source burst length, they are

Table 1: Traffic types (classes) used in numerical and simulation studies

			Type 1	Type 2	Type 3	Type 4
IDP/IPP	a_u		0.1	0.2	0.4	0.5
	λ_u	(Mbs/sec)	1	1	1	1
	R_u/Λ_{1_u}	(Mbs/sec)	10	5	2.5	2
	Λ_{2_u}	(Mbs/sec)	0	0	0	0
	$1/\beta_u$	(ms)	13	26	52	65
	$1/\theta_u$	(ms)	117	104	78	65
MMPP	a_u		0.053	0.111	0.25	0.333
	λ_u	(Mbs/sec)	1	1	1	1
	Λ_{1_u}	(Mbs/sec)	10	5	2.5	2
	Λ_{2_u}	(Mbs/sec)	0.5	0.5	0.5	0.5
	$1/\beta_u$	(ms)	13	26	52	65
	$1/\theta_u$	(ms)	234	208	156	130

also larger than that of the $min\{C_e, C_g\}$ policy which is regarded as the state of the art, and the saving is significant. The larger the buffer size, the more saving. And of a particular note that when the sources are less bursty, the allocated bandwidths are closer to the sum of the mean rate of sources than that of burstier sources. For instance the link capacity we use here is $150(Mbits/sec)$; and the mean rate of each individual source is $1(Mbits/sec)$. The mean rate allocation scheme will result in 150 connections. From Figure 5, the maximum number of less bursty sources (X axis) is more than 100 and that of burstier sources (Y axis) is less than 80. In a previous study [GMF96], we have shown that FBDCLE is more conservative than IBDCLE in terms of cell loss predictions. The numerical results shown here indicate that the corresponding statistical bandwidth C_{df_1} is also more conservative than C_{df_2} in terms of bandwidth allocation. However they are closer when the required cell loss ratio L^* is lower. More thorough examinations regarding the efficiency of C_{df_1} and C_{df_2} will be pursued below.

5.2 Sensitivity to Traffic Burstiness and to the Number of Users Being Multiplexed

It is well known that VBR sources generate bursty traffic. However the definition of the term burstiness is not unique in the literature [Onv93]. Most frequently used definitions include: (1) The ratio of the peak bit rate to the average bit rate; for the cases where IDP and IPP traffic models are used it equals the inverse of the "activity" we defined previously. (2) The average burst length, i.e. the mean active period when a source generates traffic at its peak rate. (3) The squared coefficient of variation of the interarrival times of cells, c_u^2, where c_u^2=[variance of interarrival time]/[average interarrival time]^2.

Though the ITU (formerly CCITT) defines the burstiness as the ratio of the peak-to-average traffic generation rate, researchers have found that c_u^2 is a more meaningful representation of traffic burstiness [Reg94, KME]. Our experiments also show that it is more appropriate to use c_u^2. For example, in Figure 6, by changing burst length for given activity and peak rate, we observe that c_u^2 is an increasing function of burst length. In particular, the c_u^2 parameter of IPP and IDP models is more sensitive to burst length than that of MMPP. The bandwidth requirement for a group of this kind of sources is also an increasing function of bursty length. The longer the burst length, the more required bandwidth. These observations confirm the fact that although IPP and IDP are the special cases of the two-state MMPP traffic model, they produce significantly different behaviour within the network. For given peak rate, mean rate and burst length, IPP and IDP yield larger squared coefficient of variation of interarrival times of

cells c_u^2 and thus are much burstier. Therefore, they represent the worst case in terms of burstiness and require larger buffers or more bandwidth to meet users' QoS requirements. Figure 7 shows that the admission regions identified by MMPP in all cases are much larger than that of IPP and IDP. So if we use either of these two models to perform call admission control, the network will be more robust to statistical fluctuations of source traffic. Thus we recommend that the IDP traffic model be used to determine the bandwidth requirement in CAC and in the following numerical and simulation experiments. Figure 8, where the IDP traffic model is used, obviously shows that burstier sources need more bandwidth than that of less bursty ones. Both our statistical bandwidth and $min\{C_e, C_g\}$ are shown in Figure 8 for comparison. Compared to $min\{C_e, C_g\}$ the saving by C_{df_1} and C_{df_2} at most cases is significant.

Due to the inherent Gaussian approximation which we use, when the number of users being multiplexed is small and the ratio of burst length (in cells) to buffer size is significantly long, the statistical bandwidth C_{df_1} and C_{df_2} tend to over estimate the required bandwidth. We thus propose that the following refined statistical bandwidth be used in the CAC procedure:

$$C_1^* = min\{C_{df_2}, C_p\}, \quad C_2^* = min\{C_{df_2}, C_p\},$$

where C_p is the sum of the connections' peak rate: $C_p = \sum_{u=1}^{N} R_u$. Because when the burst length is significantly long, the required bandwidth for a set of connections will tend to be the sum of their peak rates. Figure 8 shows that both $min\{C_{df_2}, C_p\}$ and $min\{C_e, C_g\}$ converge to the peak rate allocation scheme when only 5 users are multiplexed and burst length is longer than 150 (ms) or 1744 (cells) which is more than 5 times of buffer size.

Another test of the sensitivity of control schemes to traffic burstiness is to examine the influence of the activity parameter. In Figure 9, for given mean rate, the activity a_u varies in the range from 0.1 to 0.5. Here the higher activity a_u induces lower peak-rate-to-mean-rate ratio which means lower burstiness. The results once again show that required bandwidth for a set of users being multiplexed is a function of traffic burstiness. Burstier traffic demands more bandwidth to absorb the statistical fluctuation of source traffic so as to guarantee users' desired QoS. In Figure 10 the activity a_u also varies in the range from 0.1 to 0.5, however the peak rate keeps constant. Thus the higher activity a_u introduces higher traffic load, and higher required bandwidth. Moreover we observe that for given source peak rate and a small set of users, when the activity factor a_u increases, both $min\{C_e, C_g\}$ and $min\{C_{df}, C_p\}$ converge to the peak rate allocation.

The required bandwidth increases when the number of users increases; however it does not increases linearly as the equivalent bandwidth approach predicts. The plots of Figure 11 show the nonlinear increasing behavior of bandwidth versus number of users. This nonlinear increasing behaviour implies that when more users are multiplexed together, we can obtain more statistical gain. We also note that the $min\{C_{df}, C_p\}$ formula successfully overcomes the inefficiency of the original C_{df} approach when the number of users is small. The overall performance of $min\{C_{df}, C_p\}$ is better than that of $min\{C_e, C_g\}$, in particular, when the sources are burstier.

5.3 Sensitivity to Buffer Size and Desired Cell loss Ratio

The admission region concept used above provides an effective tool to compare the efficiency of different bandwidth allocation schemes when multiple users belonging to two types of traffic are statistically multiplexed. In this section, we will examine the impact of buffer size on the required bandwidth for a given cell loss ratio.

In Figure 12 we fix the link capacity, and search for the maximum number of users which can be multiplexed on that particular link while changing the buffer size. The improvement of bandwidth allocation efficiency by statistical bandwidth is dramatic when the buffer size is small or moderate. When we increase the buffer size to more than 100 times source burst length, both statistical bandwidth and the $min\{C_e, C_g\}$ schemes converge to the mean rate allocation scheme. When the buffer size is very large, i.e., large enough to absorb more than burst length $\sum_{u=1}^{N} R_u/\beta_u$, as long as the allocated bandwidth is not less than mean rate the cell loss ratio can be controlled to be less than the desired threshold value. Here the mean rate of sources is $1(Mbits/sec)$. Thus if we use the mean rate allocation scheme, the maximum number of users which can be multiplexed on a $150(Mbits/sec)$ link will be 150, as shown in Figure 12. Figure 12 also shows that the lower the desired cell loss ratio, the more the required bandwidth. Figure 13 leads to the same conclusions as those found from Figure 12, except that here we fix the number of users and calculate the required bandwidth.

6 Robustness and Fairness of Statistical Bandwidths

In the previous section we evaluated the statistical bandwidth which we propose for CAC. The comparisons to the $\min\{C_e, C_g\}$ policy indicate that CAC based on our statistical bandwidth can lead to more efficient usage of bandwidth. In this section we will evaluate the robustness and fairness of statistical bandwidth with respect to groups and types of users, and for individual users, through extensive simulations with various source traffic characteristics. Here by robustness we mean the ability of our CAC approach to conservatively respect the required cell loss ratio. By fairness we will refer to its ability to avoid leading to penalize less bursty traffic in the presence of burstier traffic.

The traffic parameters we use in our simulations are given in Table 1. In this section the source traffic is generated by using the IDP source model only. The runs were independently replicated 20 times, and each run included the transmission of 2×10^7 cells. Confidence intervals are calculated using the $Student-t$ distribution with 98% confidence so that the simulation results are of sufficiently high statistical quality. In order to perform the simulations with sufficient statistical quality and also within reasonable time, we choose the value of desired of cell loss ratio $L^* = 10^{-4}$ even though this is higher than what could be required in applications. For convenience of presentation, we define the following abbreviations which will be used in Tables 2 to 6:

- GCLR (Group Cell Loss Ratio): cell loss ratio measured for aggregate traffic.
- CCLR (Class Cell Loss Ratio): cell loss ratio measured for aggregate traffic of each class (type).
- UCLR (User Cell Loss Ratio): cell loss ratio measured for each individual user.
- UEACLR: Number of Users Exceeding Allowable Cell Loss Ratio – this will determine if fairness to individual users is respected.

Table 2 to 5 summarize the simulation results for homogeneous traffic. The cell loss statistics are obtained by using statistical bandwidth allocation. However the amount of required bandwidth calculated from $min\{C_e, C_g\}$ is also shown here for comparison. The simulation results show that with a statistical bandwidth which is less than $min\{C_e, C_g\}$, both GCLR and UCLR are successfully controlled to be less than the desired QoS requirement 10^{-4}. Note that UEACLR, the number of users which exceed the allowable cell loss ratio requirement, is zero for all the cases from Table 2 to 5. In Table 6, we present the simulation results for heterogeneous traffic. We observe that the burstier traffic which has the higher peak rate experiences higher cell loss ratio. In a certain sense, we may say that this is "Fair". The QoS requirement cannot be guaranteed for each individual user. However the overall cell loss ratio for the whole set of users is well under $L^* = 10^{-4}$.

We thus conclude that even though our proposed CAC appears to be fair in that it does not penalize the less bursty traffic in the presence of among heterogeneous users, it would be worth studying additional schemes that will not penalize burstier traffic either.

7 Conclusions

In this paper we have reviewed some of the traffic based bandwidth allocation schemes proposed in the literature for CAC in ATM. Then we have suggested the use of diffusion approximations with the instantaneous return model to estimate cell loss ratios to be used in CAC. The finite buffer diffusion model has been presented in detail, and the infinite buffer model has been briefly sketched. Predictions of both models have been presented and validated with extensive simulations with traffic resulting from the superposition of "On-Off" traffic models with varying degrees of burstiness. Simulation results have been used to show that these diffusion estimates are conservative: they tend to slightly overestimate the measured cell loss ratios. Both homogenous and heterogenous mixtures of traffic have been considered in these evaluations.

We have then derived two closed-form formulae for statistical bandwidth which have been simplified using upper bounds so that they can be easily computed from simple first and second moment traffic characteristics and from the total buffer size. Based on this analysis we have proposed a simple call admission control (CAC) procedure. In order to validate our approach, we have reported extensive numerical and simulation experiments. Compared to the existing state–of–the–art, the statistical bandwidth approach to CAC which we propose is conservative, but is able to accomodate significantly more user traffic while respecting the required cell loss ratio. It also appears to work well for traffic with different characteristics under both small and large buffer conditions.

Simulations have been used to verify that in most cases individual users' cell loss ratios requirements are respected in particular for the less bursty traffic ("fairness"), in addition to respecting the total cell loss ratio requirement. Our approach does appear to penalize the very bursty traffic. Thus we think that further research is needed to design better policies where even burstier sources are not penalized by the CAC. We are currently evaluating the proposed CAC scheme with real traffic in a simulated environment related to ATM switch architectures which are close to certain existing commercial products.

Acknowledgments

We gratefully acknowledge Dr. Levent Gün's help in correcting an inconisitency in one of our results, Yutao Feng's kind assistance with the calculations of FBDCLE and IBDCLE reported in Figure 5, and Dr. Gerald Marin's comments and suggestions.

References

[CL66] D. R. Cox and P. A. W. Lewis. *The Statistical Analysis of Series of Events.* Methuen, London, 1966.

[Dud86] A. Duda. Diffusion approximations for time-dependent queueing systems. *IEEE J. SAC*, SAC-4(6):905–918, September 1986.

[Fel93] D. C. Feldmeier. A framework of architecture concepts for high-speed communication systems. *IEEE J. on Selected Areas in Communications*, 11(4):480–488, May 1993.

[For95] The ATM Forum. *Traffic Management Specification.* ATM Forum Technical Committee, version 4.0 edition, 1995.

[Gel75] E. Gelenbe. On approximate computer system models. *J. ACM*, 22:261–263, 1975.

[Med91] J. Medhi. *Stochastic Models in Queueing Theory*, Academic Press, San Diego and London, 1991.

[GG92] R. Guerin and L. Gun. A unified approach to bandwidth allocation and access control in fast packet-switched networks. In *Proc. INFOCOM'92*, pages 0001–0012, 1992.

[GMF96] E. Gelenbe, X. Mang, and Y. Feng. A diffusion cell loss estimate for atm with multiclass bursty traffic. *in Performance Modelling and Evaluation of ATM Networks*, Chapman and Hall, London, Volume II, 1996.

[GP76] E. Gelenbe and G. Pujolle. An approximation to the behaviour of a single queue in a network. *Acta Informatica*, 7:123–136, 1976.

[HL86] H. Heffes and D. M. Lucantoni. A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE J. SAC*, SAC-4(6):856–867, September 1986.

[KME] H. Perros and K. M. Elsayed. A comparison of call admission control schemes in ATM networks. *Submitted for publication.*

[Kob74a] H. Kobayashi. Application of the diffusion approximation to queueing networks: Parts i. *Journal ACM*, 21:316–328, 1974.

[Kob74b] H. Kobayashi. Application of the diffusion approximation to queueing networks: Parts ii. *Journal ACM*, 21:459–469, 1974.

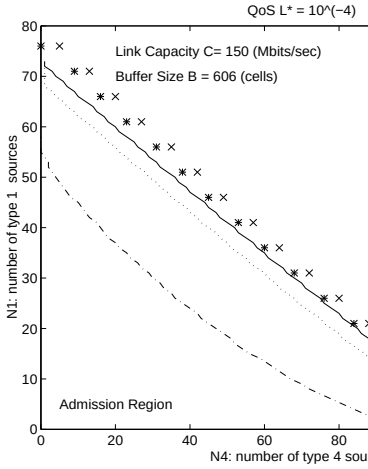
[KR93] H. Kobayashi and Q. Ren. A diffusion approximation analysis of an ATM statistical multiplexer with multiple state solutions: Part I: Equilibrium state solutions. In *Proc. ICC'93*, pages 1047–1053, 1993.

[Onv93] R. O. Onvural. *Aysnchronous Transfer Mode Networks: Performance Issues.* Artech House, Boston, 1993.

[Reg94] K. M. Rege. Equivalent bandwidth and related admission control criteria for ATM systems - a performance study. *International Journal of Communication Systems*, 7:181–197, 1994.

[RGN91] H. Ahmadi, R. Guerin and M. Naghshineh. Equivalent capacity and its application to bandwidth allocation in high speed networks. *IEEE J. on Selected Areas in Communications*, 9:968–981, 1991.

[Soh92] K. Sohraby. Heavy traffic multiplexing behaviour of highly-bursty heterogeneous sources and their admission control in high-speed networks. In *Proc. INFOCOM'92*, pages 1518–1523, 1992.



94B306.ps.bb

Figure 5: Comparison of admission regions created by $C_{df\,1}(DP)$, $C_{df\,2}(DP)$, $min\{C_e, C_g\}$, and exhaustive search of FBDCLE and IBDCLE.

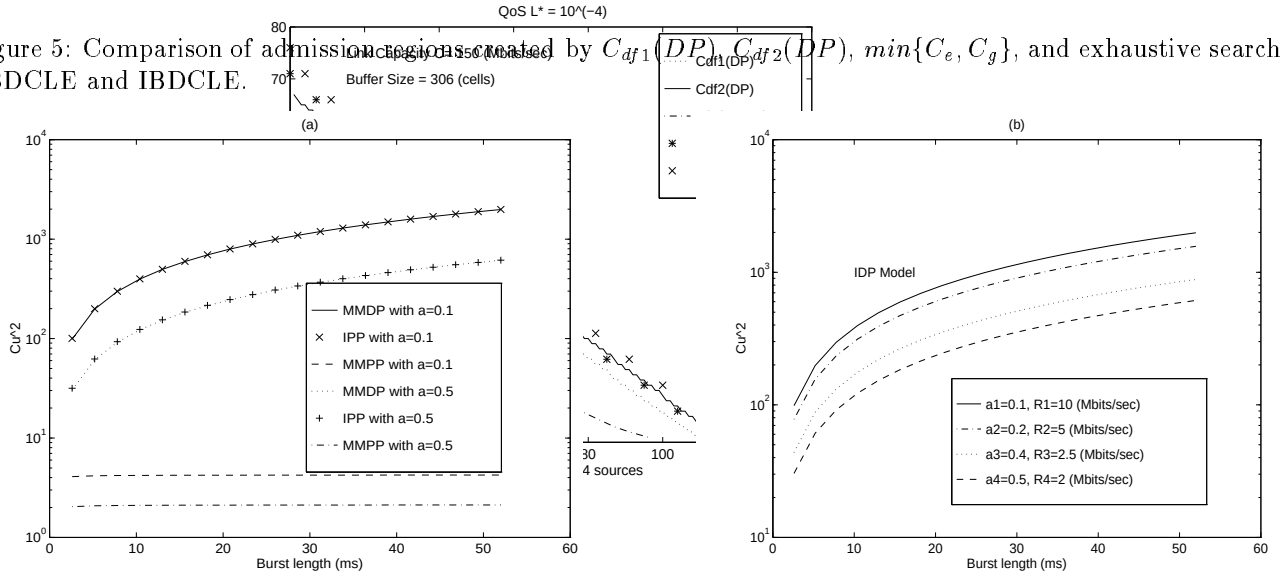


Figure 6: c_u^2 vs burst length: c_u^2 is an increasing function of burst length for given activity a_u

Table 2: Cell loss ratio measured via simulations I

Homogeneous sources of Type 1 QoS $L^* = 10^{-4}$, Buffer B=306				
N	20	30	40	50
$min\{C_e, C_g\}$	74.63	96.92	117.27	136.39
$C_{df}(Mbs/sec)$	63.35	84.29	103.43	121.44
GCLR$\times 10^{-4}$	0.5745	0.5402	0.3761	0.3239
	$\pm$0.1236	$\pm$0.1666	$\pm$0.1503	$\pm$0.1704
UCLR$_{max}$	0.7123	0.9517	0.8855	0.7577
# of UEACLR	0	0	0	0

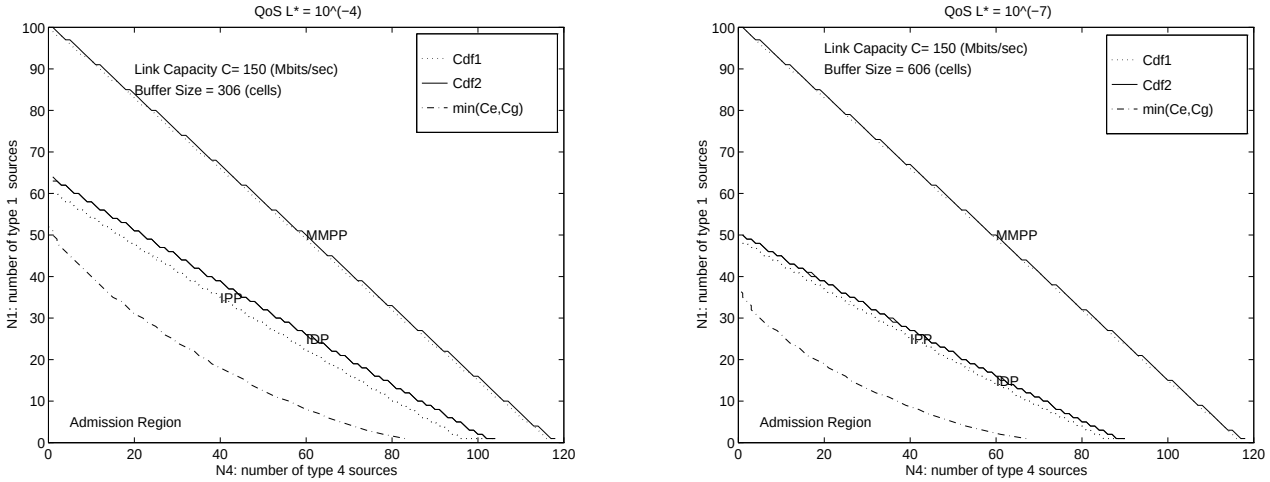


Figure 7: Comparison of admission regions created by $C_{df1}(DP)$, $C_{df2}(DP)$, $C_{df1}(IP)$, $C_{df2}(IP)$,$C_{df1}(PP)$, $C_{df2}(PP)$.

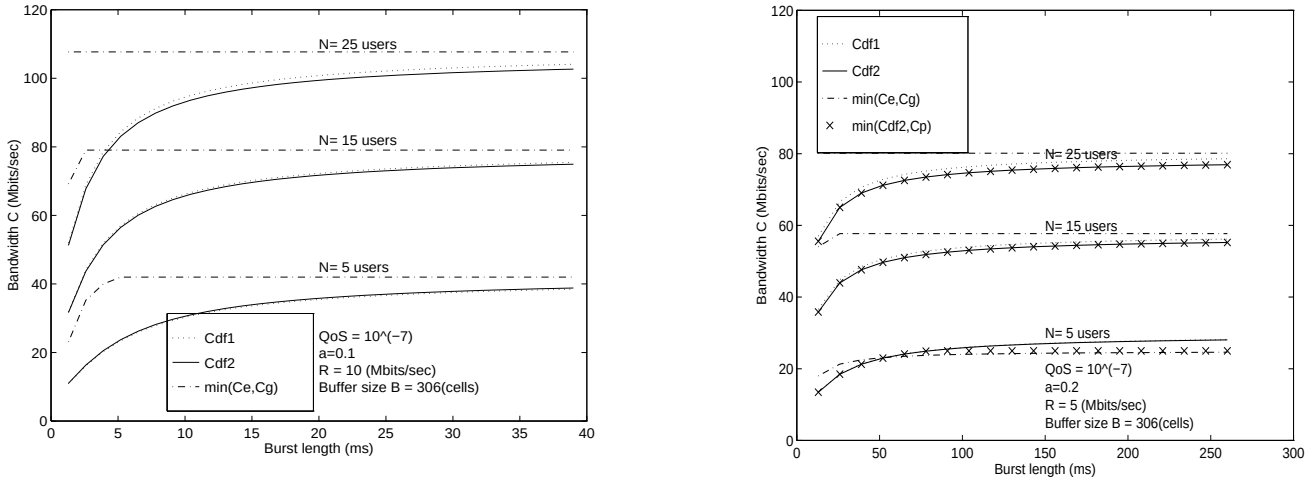


Figure 8: Required bandwidth vs burst length: keep activity factor constant

Table 3: Cell loss ratio measured via simulations II

Homogeneous sources of Type 2 QoS $L^* = 10^{-4}$, Buffer B=306				
N	20	30	40	50
$min\{C_e, C_g\}$	56.43	74.61	91.51	107.59
$C_{df}(Mbs/sec)$	48.93	66.02	81.96	97.15
GCLR$\times 10^{-4}$	0.2182	0.1782	0.1530	0.2360
	$\pm$0.0862	$\pm$0.0876	$\pm$0.0942	$\pm$0.1957
UCLR$_{max}$	0.3743	0.3979	0.3436	0.6155
# of UEACLR	0	0	0	0

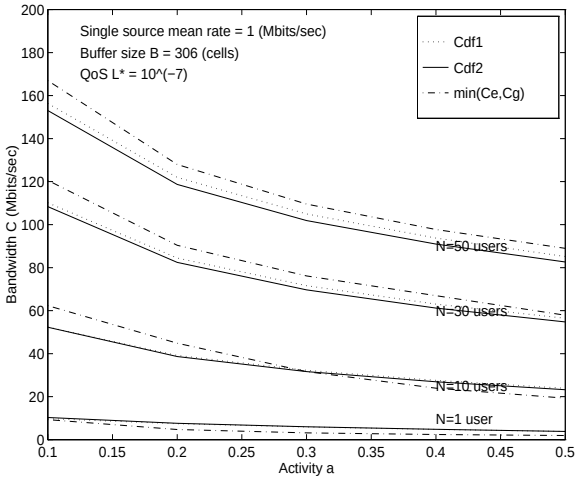


Figure 9: Bandwidth vs activity: keep mean rate constant

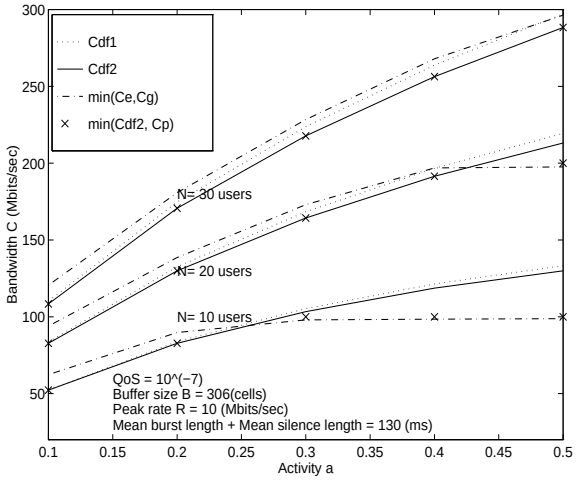
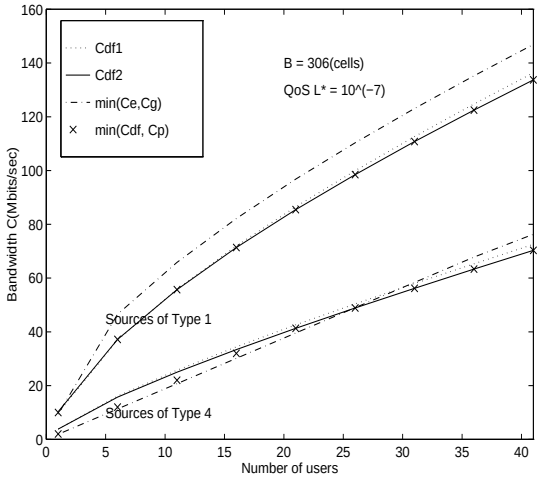


Figure 10: Bandwidth vs activity: keep peak rate constant



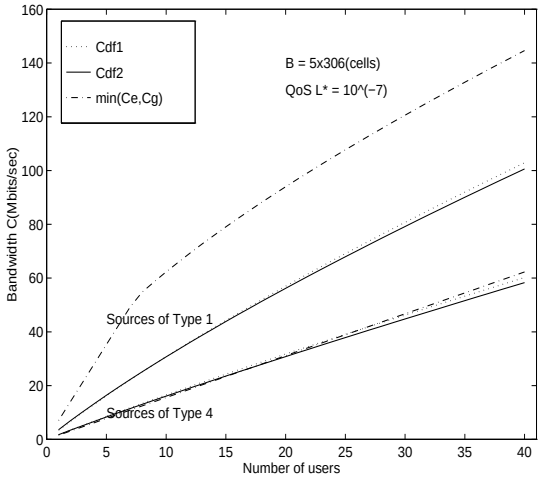


Figure 11: Bandwidth vs number of users being multiplexed

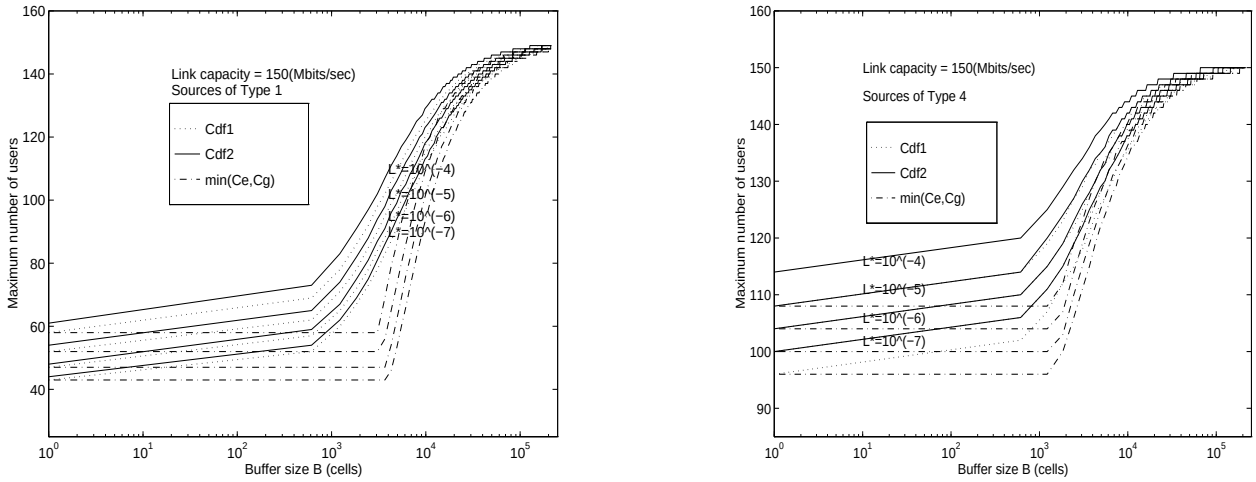


Figure 12: Maximum admissible number of users vs buffer size

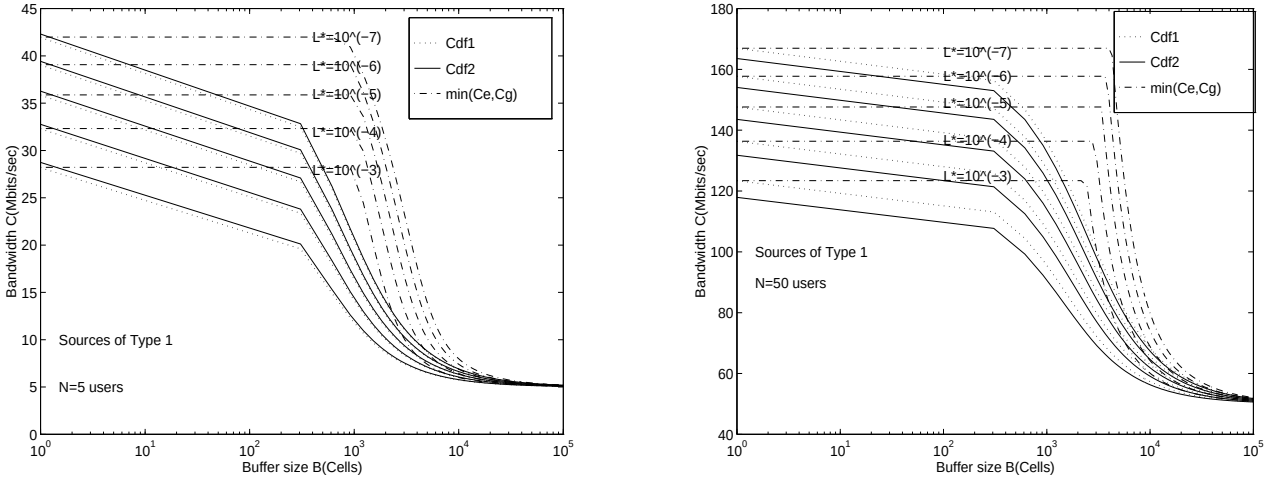


Figure 13: Bandwidth vs buffer size

Table 4: Cell loss ratio measured via simulations III

Homogeneous sources of Type 3 QoS $L^* = 10^{-4}$, Buffer B=306				
N	20	30	40	50
$min\{C_e, C_g\}$	42.31	57.32	71.55	85.27
$C_{df}\,(Mbs/sec)$	36.75	56.91	64.39	77.43
GCLR$\times 10^{-4}$	0.0208	0.0230	0.0307	0.0341
	$\pm$0.0181	$\pm$0.0213	$\pm$0.0325	$\pm$0.0050
UCLR$_{max}$	0.0573	0.0758	0.0862	0.1952
# of UEACLR	0	0	0	0

Table 5: Cell loss ratio measured via simulations IV

Homogeneous sources of Type 4 QoS $L^* = 10^{-4}$, Buffer B=306				
N	20	30	40	50
$min\{C_e, C_g\}$	36.13	52.31	65.76	78.80
$C_{df}(Mbs/sec)$	32.97	46.29	59.07	71.49
GCLR$\times 10^{-4}$	0.0007	0.0036	0.0110	0.0161
	$\pm$0.0018	$\pm$0.0064	$\pm$0.0160	$\pm$0.017
UCLR$_{max}$	0.0050	0.0281	0.0619	0.0966
# of UEACLR	0	0	0	0

Table 6: Cell loss ratio measured via simulations V

QoS $L^* = 10^{-4}$, Buffer B=306					
Heterogeneous N	10	20	30	40	50
$min\{C_e, C_g\}$	35.60	61.43	77.23	95.62	111.09
$C_{df}(Mbs/sec)$	28.72	52.10	67.23	84.45	99.02
GCLR$\times 10^{-4}$	0.9310	0.7348	0.4965	0.3207	0.3920
	$\pm$0.1639	$\pm$0.1196	$\pm$0.1938	$\pm$0.1272	$\pm$0.1925
GCLR$_{max} \times 10^{-4}$	1.1996	1.4227	1.2120	0.7953	0.9524
# of UEACLR	2	6	1	0	0
N1	2	8	9	13	16
CCLR1$\times 10^{-4}$	1.981	1.150	0.7969	0.4899	0.6018
	$\pm$0.3774	$\pm$0.1958	$\pm$0.3533	$\pm$0.1989	$\pm$0.2975
UCLR1$_{max} \times 10^{-4}$	1.1996	1.4227	1.2120	0.7953	9.5240
# of UEACLR	2	6	1	0	0
N2	4	6	10	13	14
CCLR2$\times 10^{-4}$	0.8140	0.5555	0.4421	0.2594	0.3644
	$\pm$0.1454	$\pm$00.0972	$\pm$0.1663	$\pm$0.1232	$\pm$0.1735
UCLR2$_{max} \times 10^{-4}$	0.9100	0.6545	0.5500	0.3510	0.5550
# of UEACLR	0	0	0	0	0
N3	3	3	5	7	10
CCLR3$\times 10^{-4}$	0.5424	0.3793	0.3527	0.2321	0.2566
	$\pm$0.1338	$\pm$0.1040	$\pm$0.1435	$\pm$0.1053	$\pm$0.1510
UCLR3$_{max} \times 10^{-4}$	0.5713	0.4235	0.3823	0.2837	0.3461
# of UEACLR	0	0	0	0	0
N4	1	3	6	7	10
CCLR4$\times 10^{-4}$	0.4651	0.3426	0.2545	0.2075	0.2313
	$\pm$0.1069	$\pm$0.1001	$\pm$0.1097	$\pm$0.0855	$\pm$0.1291
UCLR4$_{max} \times 10^{-4}$	0.4651	0.3828	0.3287	0.2672	0.3586
# of UEACLR	0	0	0	0	0