Call Admission Control Schemes : A Review *

Harry G. Perros
Department of Computer Science
North Carolina State University
Raleigh, NC 27695-8206, U.S.A.
hp@csc.ncsu.edu

Khaled M. Elsayed
Nortel Inc.
P.O.Box 833871
Richardson, TX 75083-3871, U.S.A.
khalede@nortel.com

Abstract

Over the last few years, a substantial number of call admission control (CAC) schemes have been proposed for ATM networks. In this paper, we review the salient features of some of these algorithms. Also, we quantitatively compare the performance of three of these schemes.

1 Introduction

The current infrastructure for public networking comprises two different realms: circuit-switched telephone networks and packet-switched data networks. The need for the integration of services resulted in the introduction of the Narrowband Integrated Services Digital Networks (N-ISDN) in the 1980's. The benefits of introducing N-ISDN included common user-network interfaces for a variety of services, improved signalling capabilities, and enhanced integrated services.

Broadband Integrated Services Digital Networks (B-ISDN) were envisioned as a provider of higher bit rates to the user than N-ISDN. One of the key design objectives of B-ISDN is "The provision of a wide range of services to a broad variety of users utilizing a limited set of connection types and multi-purpose user-network interfaces" [1]. The two prominent enabling technologies for the deployment of B-ISDN are fiber optics, and the Asynchronous Transfer Mode (ATM) network architecture.

*Supported in part by BellSouth, GTE Corporation, and NSF and DARPA under cooperative agreement NCR-8919038 with the Corporation for National Research Initiatives and in part by a gift from BNR INC.

ATM has been the hottest topic in the networking community for the past few years. ATM has been proposed by the International Telecommunications Union (ITU), formerly known as CCITT, as the transport mechanism of choice for B-ISDN [1]. "Transport" here refers to ATM switching and multiplexing techniques at the data link layer of the 7-layer ISO model used to convey user traffic from source to destination. ATM is the first scheme to provide a unified interface which can be used by a variety of services with drastically different requirements. It is a blend of circuit-switching and packet-switching technologies. It borrows the notion of connection-oriented services from circuit-switched networks. However, in ATM, resources may or may not be reserved for the whole duration of the connection. ATM is based on packet-switching in the sense that all traffic is transported via fixed-size packets (called cells in ATM terminology). Traffic is relayed and routed by means of information contained within the cell.

ATM cells have been standardized by ITU to be 53 octets long. The length has been chosen so that it would be possible for ATM to transport traffic from interactive communication services (e.g. voice and video) efficiently. A cell consists of a 5-octet header and a 48-octet information payload. The format ATM uses is universal for any network, be it a local or wide area, public or private. This has a potential for not only providing a uniform scheme for integrating various types of services but also for a seamless integration of local and wide area networking.

One area of paramount importance in ATM networks is congestion control. The primary role of a network congestion control procedure is to protect the network and the user in order to achieve network performance objectives and optimize the usage of network resources. In ATM-based B-ISDN, congestion control should support a set of ATM quality of service classes sufficient for all foreseeable B-ISDN services.

Congestion control procedures can be classified into *preventive* control and *reactive* control. In preventive congestion control, one sets up schemes which prevent the occurrence of congestion. In reactive congestion control, one relies on feedback information for controlling the level of congestion. Both approaches have advantages and disadvantages. In ATM networks, a combination of these two approaches is currently used in order to provide effective congestion control. For instance, CBR and VBR services use preventive schemes and ABR service is based on a reactive scheme.

Preventive congestion control involves the following two procedures: *call admission control*

(CAC) and *bandwidth enforcement*. As mentioned above, ATM is a connection oriented service. Before a user starts transmitting over an ATM network, a connection has to be established. This is done at *call set-up* time. The main objective of this procedure is to establish a path between the sender and the receiver. This path may involve one or more ATM switches. On each of these ATM switches, resources have to be allocated to the new connection.

The call set-up procedure runs on a resource manager, which is typically a workstation attached to the switch (see figure 1). The resource manager controls the operations of the switch, accepts new connections, tears down old connections, and performs other management functions. If a new connection is accepted, bandwidth and/or buffer space in the switch is allocated for this connection. The allocated resources are released when the connection is terminated.

Call admission control deals with the question as to whether a switch can accept a new connection or not. Typically, the decision to accept or reject a new connection is based on the following two questions:

1. Does the new connection affect the quality-of-service of the connections that are currently being carried by the switch?
2. Can the switch provide the quality-of-service requested by the new connection?

Call admission control schemes may be classified as a) non-statistical allocation, or peak bandwidth allocation, and b) statistical allocation. Below, we examine these two case. As will be seen, it is difficult to design good call admission schemes for statistical allocation. For presentation purposes, let us consider a non-blocking ATM switch, as the one shown in figure 1. In a non-blocking switch, the point of congestion occurs at the output ports. In view of this, as we can see in figure refatm-outbuff, each output port is provided with a finite buffer. We will assume that each output port has its own dedicated buffer, rather than several output ports sharing a common output buffer. Also, we make the obvious assumption that the existing traffic currently going through an output port is such that it can be handled by the output port at the required quality-of-service. Let us assume that the output port provides a cell loss probability of 10^{-8} for the existing traffic. Assuming that the new connection is accepted, would the cell loss probability be also of the order of 10^{-8} for the total traffic carried by the port?

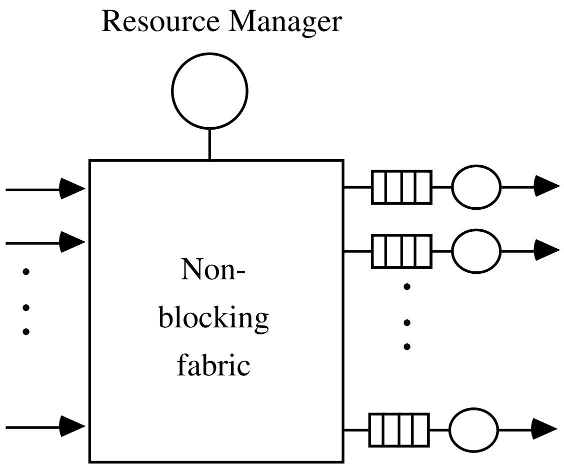


Figure 1: An ATM switch with output buffering

1.1 Non-statistical allocation (Peak bandwidth allocation)

Suppose a source has an average bandwidth of 20 Mb/s and a peak bandwidth of 45 Mb/s. Peak bandwidth allocation, otherwise known as non-statistical allocation, requires that 45 Mb/s be reserved at the output port for the specific source, independent of whether the source transmits continuously at 45Mb/s or not. Peak bandwidth allocation is used in CBR services, which are suitable for applications such as: PCM-encoded voice and other fixed rate applications, unencoded video, and very low bandwidth applications such as telemetry.

The advantage of peak bandwidth allocation is that it is easy to decide whether to accept a new connection or not. This is because only knowledge of the peak rate of the new connection is required. The new connection is accepted if the sum of the peak rates of all the existing connections plus the peak rate of the new connection is less than the capacity of the output link. (We note here that it is possible that cells belonging to a connection may be interleaved with cells from other connections. In view of this, cells belonging to a connection may momentarily arrive faster than expected. That is, the peak rate may be momentarily exceeded. To avoid this problem, one should allocate at a peak rate slightly higher than the one requested.)

The disadvantages of peak allocation is that unless connections transmit at peak rates, the output port link will be grossly under-utilized.

1.2 Statistical allocation

In statistical allocation, bandwidth for a new connection is not allocated on per peak rate basis. Rather, the allocated bandwidth is less than the peak rate of the source. As a result, the sum of all peak rates may be greater than the capacity of the output link. Statistical allocation makes economic sense when dealing with bursty sources, but it is difficult to carry out effectively. This is because of difficulties in characterizing an arrival process and lack of understanding as to how an arrival process is shaped deep in the ATM network.

Another difficulty in designing a call admission control algorithm for statistical allocation is that decisions have to be done on the fly, and therefore they cannot be CPU intensive. Typically, the problem of deciding whether to accept a new call or not may be formulated as a queueing problem. For instance, let us consider the non-blocking switch shown in figure 1. The call admission control algorithm has to be applied to the buffer of each output port. If we isolate an output port and its buffer from the rest of the switch, we will obtain the queueing model shown in figure 2. This type of queueing structure is known as an ATM multiplexer. It represents a number of ATM sources feeding a finite capacity queue which is served by a server (the output port). The service time is constant equal to the time it takes to transmit an ATM cell. Now, assuming that the quality of service of the existing connections is satisfied, the question arises whether the quality of service will still be maintained if the new connection is added. This can answered by solving this ATM multiplexer with the existing and new connections. However, the solution to this problem is very difficult and CPU intensive (see for example Elsayed and Perros [15] and Li [41]). It gets even more difficult, if we assume complicated arrival processes. Certainly this is not something that can be done on the fly. In view of this, a variety of different bandwidth allocation algorithms have been proposed which are based on different approximations, or different types of schemes which do not require the solution of such a queueing problem.

Another issue that has not been addressed adequately so far, is call admission control for video sources. One can safely opt for peak bandwidth allocation for un-encoded video or CBR encoded video. However, given the trends in video encoding, it is reasonable to assume that video-based applications will make use of VBR encoded video. Characterizing the behaviour of the output process of an encoder is still an open research question (see Magalaris et al. [44], Heyman,

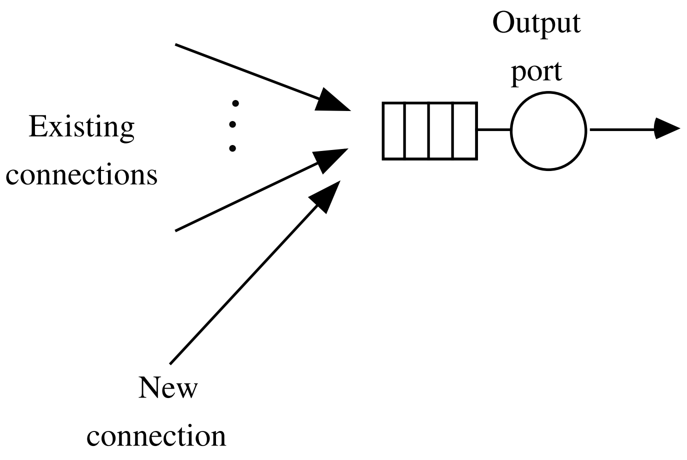


Figure 2: An ATM multiplexer

Tabatabai and Lakshman [34], and Lucantoni, Neuts, and Reibman [43]).

In this paper, we will review some of the call admission control algorithms that have been proposed for statistical allocation. Before we proceed, however, we examine briefly the problem of traffic characterization.

2 Characterization of an arrival process

Prior to the advent of ATM networks, performance models of telecommunication systems were typically developed based on the assumption that arrival processes are Poisson distributed. That is, the time between successive arrivals is exponentially distributed. In some cases, such as in public switching, extensive data collection actually supported the Poisson assumption. In early performance studies of ATM networks, arrival processes were also assumed to be Poisson distributed. Alternatively, they were assumed to be Bernoulli distributed. This is due to the fact that an ATM cell has a fixed length. Therefore, one can model cell arrival by dividing the time axis into slots. Each slot is assumed to be long enough so that it can accommodate a complete transmission of a cell. Now looking at the slotted time axis, each slot may or may not contain a cell. Assume that a slot contains a cell with a probability $p < 1$, or it is empty with probability $1 - p$. Then, the time between two successive arrivals has a geometric distribution, and the number of arrivals per unit time is Bernoulli distributed. The Bernoulli arrival is the discrete-time equivalent of the Poisson process.

Over the last few years, we have gone through several paradigm shifts regarding our understanding of how to model an ATM source. Following the first performance models which were based on the Poisson assumption or the Bernoulli assumption, it became apparent that these traffic models did not capture the notion of burstiness that is present in traffic resulting from applications such as moving a data file and packetized encoded video. Thus, there was a major shift towards using distributions of the on/off type, such as the Interrupted Poisson Process (IPP) or its discrete-time counterpart the Interrupted Bernoulli Process (IBP). In an IPP, there is an active period during which arrivals occur in a Poisson fashion, followed by an idle period during which no arrivals occur. These two periods are exponentially distributed, and they alternate continuously. An IBP is defined similarly, only the arrivals during the active period are Bernoulli distributed, and the two periods are geometrically distributed. An IPP or an IBP, however, does not capture the notion of correlation since successive inter-arrival times are independent of each other (that is the inter-arrival time is a renewal process). Another way of describing a source is using the fluid approach. Here arrivals occur with a continuous rate during the active period. This defines an on/off fluid source or equivalently an Interrupted Fluid Process (IFP).

Early traffic characterization of ATM traffic showed that the inter-arrival times of cells from a specific source may well be correlated. As a result, more complex distributions were introduced for modeling ATM traffic. These distributions are in the form of a Markov Modulated Poisson Process (MMPP), its discrete-time counterpart a Markov Modulated Bernoulli Process (MMBP), or a Markov Modulated Fluid Process (MMFP). An MMPP is a Markov process that can find itself in several different states. In each state, arrivals occur in a Poisson fashion at a rate which is state-dependent. An MMBP/MMFP is similarly defined, only in each state arrivals occur in a Bernoulli/continuous fluid fashion at a state-dependent rate. An IPP/IBP/IFP is a special case of an MMPP/MMBP/MMFP. In general, the more complex the distribution, the harder it is to incorporate it into analytic performance models of ATM networks.

One of the underlying assumption of an MMPP/MMBP/MMFP is that the time the arrival process spends in each state is exponentially (or geometrically) distributed. This assumption is made for mathematical convenience. There was not much concern about this assumption, since these distributions captured the notion of burstiness and correlation, two factors that were deemed more important than the exponentiality assumption. However, the current thinking is that this may not be a realistic assumption for applications such as file transfer. It seems that a bursty data source should be characterized by an on/off process, like an IBP, but the on and off periods should have arbitrary distributions. In fact, an ATM traffic study of VISTAnet (see Perros, Nilsson, and Kuo [48]). clearly points out to an on/off traffic model with constant on period. The off period seems to be best described by a mixture of two constants. Analyzing the behaviour of an ATM multiplexer under on/off periods with arbitrarily distributed on and off periods is very difficult (see Elsayed [14] and Guibert [32].

Finally, we should mention that several auto-regressive type of models have been proposed to characterize the traffic due to video (see for example Magalaris et al. [44], Heyman, Tabatabai and Lakshman [34], and Grünenfelder et al. [29]). This is an area of active research. Also, more recently, a different approach has been used to characterize traffic based on the notion of long-term correlations. This approach is based on the theory of self-similarity (see Leland et al. [40], Erramilli, Gordon and Willinger [20] and Duffield, Lewis and O'Connel [12] and references therein).

To compound the problem of choosing an appropriate model for ATM traffic, the ATM forum decided to standardize the following parameters: peak rate, average rate, cell delay variation for the peak rate, and maximum burst length. Using the peak rate and the cell delay variation, one can effectively police the peak rate. Also, using the maximum burst length, one can estimate a cell delay variation that can be used to police the average rate. These parameters are fairly inadequate when it comes to bandwidth allocation. For, it can be easily shown that there are different distributions with the same peak, average rate, and maximum burst length, but with different burstiness and inter-arrival correlations. Burstiness and correlation are two parameters that can grossly affect QoS measures such as cell loss probability.

Finally, assuming that the arrival process can be adequately characterized by a traffic model, the next question that arises is how does the burstiness and the correlation of the inter-arrival time are affected as the source goes through several switches, multiplexers and demultiplexers? If the source gets less bursty as it proceeds through the network, then it is easier to decide how much bandwidth to allocate. However, this decision gets more difficult if the source becomes burstier as it goes through the network. This is an open problem that has not as yet been adequately addressed.

3 Classification of call admission schemes

A variety of different call admission schemes have been proposed in the literature. Some of these schemes require an explicit traffic model and some only require traffic parameters such as the peak and average rate. In this tutorial we review some of these schemes. For presentation purposes, the schemes have been classified into the following groups:

1. Equivalent capacity
2. Heavy traffic approximation
3. Upper bounds of the cell loss probability
4. Fast buffer/bandwidth allocation
5. Time windows

This classification was based on the underlying principle that was used to develop the scheme. Below, we discuss the salient features of each group and review some of the proposed schemes.

3.1 Equivalent capacity

The equivalent capacity of a source (or sources) is a popular notion in call admission, and it has also given rise to some interesting queueing problems. Let us consider a single source feeding a finite capacity queue. Then, the equivalent capacity of the source is the service rate of the queue that corresponds to a cell loss of ϵ.

The equivalent capacity for a single source can be derived as follows, see Guérin, Ahmadi, and Naghshineh [30]. Each source is assumed to be an IFP. Let R be its peak rate, r the fraction of time the source is active, and b the mean duration of the active period. Then, an IFP source can be completely characterized by the vector (R, r, b). Let us now assume that the source feeds a finite capacity queue with constant service time. Let K be the capacity of the queue. Then using the technique of Anick, Mitra, and Sondhi [2], one can obtain the queue-length distribution. From this distribution, it is possible to determine a service rate c that corresponds to a give cell loss ϵ. The equivalent capacity c can be found to be in the form:

$$\epsilon = \beta \;\; exp\{-\frac{K(c - rR)}{b(1-r)(R-c)c}\},$$

where

$$\beta = \frac{(c - cR) + \epsilon r(R-c)}{(1-r)c}.$$

The equivalent capacity c can then be obtained by solving the above equation for c. No closed-form solution, however, can be obtained from the above equation, and the solution has to be calculated numerically. A simplification can be obtained when b is set equal to 1 (typically $\beta < 1$). In this case, we obtain:

$$c = \frac{a - K + \sqrt{(a-K)2 + 4Kar}}{2a} \; R \tag{1}$$

where $a = ln(1/\epsilon)b(1-r)R$.

In the case of N sources, and given that the buffer has a capacity K, the equivalent capacity is again the service rate c which ensures that the cell loss is ϵ. The calculation of the equivalent capacity, however, becomes very complicated. In view of this, Guérin, Ahmadi, and Naghshineh [30] proposed the following approximation:

$$c = min\{\rho + a'\sigma, \sum_{i=1}^{N} c_i\} \tag{2}$$

where

- c_i is the equivalent capacity of the ith source calculated using expression (2.1.1), and $\sum_{i=1}^{N} c_i$ is the sum of all the individual equivalent capacities,
- ρ is the total average bit rate, i.e. $r = \sum_{i=1}^{N} \rho_i$, where ρ_i is the mean bit rate of the ith source,
- $\sigma = \sum_{i=1}^{N} \sigma_i$, where σ_i^2 is the variance of the bit rate of the ith source, $\sigma_i^2 = \rho_i(R_i - \rho_i)$, and
- $a' = \sqrt{-2ln(\epsilon) - ln2\pi}$.

This approximation is based on the following two observations. First, the multiplexed N sources may well correspond to an equivalent capacity which is less than the sum of their individual equivalent capacities. Secondly, the stationary bit rate of the N sources has been observed to follow approximately a Normal distribution with mean ρ, and variance σ^2. Assuming that the finite capacity queue has no buffer, i.e. $K = 0$, the equivalent capacity for the N sources is simply a point in the Normal distribution $N(\rho, \sigma^2)$ past which the area under the curve is ϵ. This point expressed in standard deviations is $\rho + a'\sigma$, and a' is obtained by approximately inverting the Normal distribution. Thus, the equivalent capacity of N sources is the minimum of the two different equivalent capacities given by (2). This expression turns out to be an upper bound on the actual bandwidth requirements. The authors, however, mention that this a reasonable upper bound.

Elwalid and Mitra [17] showed that the equivalent capacity of a Markov modulated fluid source is approximately the maximum real eigenvalue of a matrix derived from source parameters, multiplexer resources, and the cell loss probability. Consider a traffic source modeled by L states and let $\boldsymbol{Q}$ be the infinitismal generator of the modulating Markov chain that governs the transition between the states of the arrival process and $\vec{\boldsymbol{\lambda}} = (\lambda_1, \lambda_2, \cdots, \lambda_L)$ be a vector of rate of arrivals at the states of the Markov chain.

The equivalent capacity c of such a source was shown to be the maximal real eigenvalue of the matrix

$$\boldsymbol{\Lambda} - \frac{1}{\xi}\boldsymbol{Q}$$

where $\boldsymbol{\Lambda} = diag(\vec{\boldsymbol{\lambda}})$ and $\xi = ln(\epsilon)/K$. It was also shown that if N such sources are superposed, then their equivalent capacity is asymptotically equal to $c = \sum_{i=1}^{N} c_n$, where c_n is the equivalent capacity of the nth source (computed as if the source is the only source in the system).

Some studies (see Choudhury, Lucantoni, and Whitt [7] and Elsayed and Perros [16]) have clearly indicated the inaccuracy of equivalent capacity methods in some situations. Rege [51] compares various approaches for equivalent capacity and proposes some modifications to enhance the accuracy of the scheme. A recent paper by Elwalid et al. [18] proposes a method combining Chernoff bounds and equivalent capacity approximation to overcome the shortcomings of the equivalent capacity for multiplexers that can achieve a substantial statistical gain even with small or no buffers. This, however, does not solve all the problems with the inaccuracy of equivalent capacity approximation in some other cases.

Kulkarni, Gün, and Chimento [39] considered the equivalent capacity vector for two-priority on/off source. Chang and Thomas [6] introduced a *calculus* for evaluating source equivalent capacity at output of multiplexers and upon demultiplexing or routing. On-line evaluation of equivalent capacity have been proposed by De Veciana, Kesidis and Walrand [63], and Duffield et al. [13] which proposes maximum entropy as a method for characterizing traffic sources and their equivalent capacity. Further relevant references are Gibbens and Hunt [26], Kelly [37], Kesidis, Walrand and Chang [38] and Guérin and Gün [31].

3.2 Heavy traffic approximation

Sohraby [56] proposed an approximation for bandwidth allocation based on the asymptotic behavior of the tail of the queue-length distribution (note that the equivalent capacity method is also based on the asymptotic behavior of the tail of the queue-length distribution).

Let us first consider an infinite capacity queue with constant service time and a MMBP arrival process. Let the probability transition matrix of the modulating Markov chain be given by $\boldsymbol{P} = [P_{ij}]$ and $\vec{\boldsymbol{\lambda}} = (\lambda_1, \lambda_2, \cdots, \lambda_L)$. The probability generating function of the arrival process $\boldsymbol{B}(z)$ is defined as follows: $b_{ij}(z) \stackrel{\triangle}{=} E[z^{A_{n+1}} \mathbf{1}(S_{n+1} = j | S_n = i)]$ where S_n is the state if the underlying Markov chain in slot n and $\mathbf{1}(V)$ is equal to one if the event V is true and equal to zero otherwise. It is known that the steady-state queue-length distribution exhibits a geometrically distributed tail. That is, for sufficiently large i, we have

$$Pr(queue - length > i) \approx \alpha(1/z^*)^i,$$

where z^* is the smallest root outside the unit circle of the determinant $|zI - B(z)|$ and α is an unknown constant. Now, let $\gamma_1, \gamma_2, \cdots, \gamma_L$ be the eigenvalues of $B(z)$. Then, the determinant $|zI - B(z)|$ can be written as follows:

$$|zI - B(z)| = \prod_{i=1}^{n} (z - \gamma_i(z)).$$

Therefore, once the eigenvalues of B(z) are determined, then the zeroes of the above determinant can be easily obtained. The question remains, however, as to which of these n equations $z - \gamma_i(z)$ gives z^*. It can be shown that the root z^* solves equation $z - \gamma(z) = 0$, where $\gamma(z)$ is the Perron-Frobenius (PF) eigenvalue of $B(z)$. For an arrival process which is the superposition of independent arrival processes, the PF eigenvalue is the product of the PF eigenvalues of the individual sources. Assuming a superposition of IBP sources, it can be shown that z^* can be obtained by solving a fixed-point problem. This fixed-point problem has to be solved numerically. Therefore, in order to expedite the calculation of z^* for the superposition of N IBPs, Sohraby proposed the following approximation which is valid under the assumption that the b_i's are very large:

$$z^* \approx 1 + \frac{1 - r}{\sum_{i=1}^{N} r_i R_i (1 - r_i)^2 b_i} \tag{3}$$

where $r = \sum_{i=1}^{N} r_i R_i$. For on/off sources where the on and off periods are characterized by an arbitrary distribution, Sohraby [57] suggested the following approximation. Let the squared coefficient of variation of the lengths of the on and off periods be given by cv_{on}^2 and cv_{off}^2 respectively. In the regime of b and K very large, the following approximation for z^* is valid:

$$z^* \approx 1 + \frac{2(1 - r)}{\sum_{i=1}^{N} r_i R_i (1 - r_i)^2 (cv_{on}^2 + cv_{off}^2) b_i} \tag{4}$$

The tail of the queue-length distribution can be approximated by

$$Pr(queue - length > i) \approx \gamma (1/z^*)^i,$$

where γ is the traffic intensity, and z^* is given by (3) or (4). The author suggested that the approximation is good when the traffic intensity γ is $0.8 < \gamma < 1$. The cell loss probability is approximated by

$$\gamma (1/z^*)^K,$$

where K is the buffer capacity, and z^* is given by (3) or (4). The bandwidth allocation decision is then quite simple. Accept a new connection if the resulting $\gamma(1/z^*)^K$ is small, or when

$$ln[\gamma (1/z^*)^K] < ln(\epsilon).$$

3.3 Upper bounds of the cell loss probability

Several other call admission schemes have been proposed which are based on an upper bound for the cell loss probability. Saito [53] proposed an upper bound based on the average number of cells that arrive during a fixed interval (ANA), and the maximum number of cells that arrive in the same fixed interval (MNA). The fixed interval was taken to be equal to $D/2$, where D is the maximum admissible delay in a buffer. Using these parameters, the following upper bound was derived. Let us consider a link serving N connections, and let $p_i(j)$, $i = 1, 2, \cdots, N$, and $j = 0, 1, \cdots$ be the probability that j cells belonging to the ith connection arrive during the period $D/2$. Then, the cell loss probability CLP can be bounded by

$$CLP \le B(p_1, \cdots, p_N; D/2) = \frac{\sum_{k=0}^{\infty}[k - D/2]^+ p_1 \star \cdots \star p_N(k)}{\sum_{k=0}^{\infty} k p_1 \star \cdots \star p_N(k)}$$

where $\star$ is the convolution operation. Let $\theta_i(j)$ be the following functions:

$$\theta_i(j) = \begin{cases} ANA_i/MNA_i, & j = MNA_i, \\ 1 - ANA_i/MNA_i, & j = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Then it can be shown that

$$\begin{aligned} CLP &\le B(p_1, \cdots, p_N; D/2) \\ &\le B(\theta_1, \ldots, \theta_N; D/2) \\ &= \frac{\sum_{k=0}^{\infty}[k-D/2]^+ \theta_1 \star \cdots \star \theta_N(k)}{\sum_{k=0}^{\infty} k\theta_1 \star \cdots \star \theta_N(k)}. \end{aligned}$$

A new connection is admitted if the resulting $B(\theta_1, \cdots, \theta_{N+1}; D/2)$ is less than the admissible cell loss probability. Saito proposes a scheme for calculating $\theta_1 \star \theta_2 \star \cdots \star \theta_N$ efficiently. He also obtained a different upper bound based on the average and the variance of the number of cells that arrive during $D/2$.

For other upper bounds on the cell loss probability see Rasmussen et al. [50], Castelli, Cavallero, and Tonietti [5], Doshi [11] and the closely related work by Elwalid, Mitra, and Wentworth [19].

3.4 Fast buffer/bandwidth allocation

This scheme was devised for the transmission of bursty sources. The main idea behind this scheme is the following. When a virtual circuit is established, the path through the network is set-up and the routing tables are appropriately updated, but no resources are allocated to the virtual circuit. When a source is ready to transmit a burst, then at that moment the network attempts to allocate the necessary resources for the duration of the burst. Below, we examine two such schemes.

Tranchier, Boyer, Rouaud, and Mazeas [59] proposed a fast bandwidth allocation protocol for VBR sources whose peak bit rate is less than 2% of the link's capacity. A source requests bandwidth in incremental and decremental steps. The total requested bandwidth for each virtual circuit may vary between zero and its peak rate. For a step increase, a virtual circuit uses a special reservation request cell. The requested increase is accepted by a node if the sum of the total requested traffic does not exceed the link's capacity. That is, the decision to accept a step increase or not is based on peak bandwidth allocation. If the step increase is denied by a node on the path of the virtual circuit, the step increase is blocked. Step decreases are announced through a management cell. A step decrease is always accepted. At the cell level, the incoming cell stream of a virtual circuit is shaped, so that the peak cell rate enforced corresponds to the currently accepted bandwidth. A fast reservation protocol (FRP) unit was implemented to handle the relevant management cells. This unit is located at the user network interface points (UNI). The protocol utilizes different types of timers to ensure its reliable operation. The terminal utilizes a timer to ensure that its management cells, such as step increase requests, sent to its local FRP unit are not lost. When the FRP unit receives a step increase request, it forwards the request to the first node in the path, which then sends it to the following node and so on. If the request can be satisfied by each node on the path, the last node sends an ACK to the FRP unit. The FRP unit then informs the terminal that the request has been accepted, updates the policing function, and sends a validation cell to the nodes on the path to confirm the reservation. If the request cannot be satisfied by a node, the node simply discards the request. The upstream nodes, that have already reserved bandwidth, will discard the reservation if they do not receive the validation cell within a fixed period of time, i.e. until a timer expires. This timer is set equal to the maximum round trip between the FRP unit and the furthermost node. If the request is blocked, the FRP unit will re-try to request the step increase after a period set by another timer. The number of attempts is limited.

Turner [60, 61] proposed a fast reservation scheme where buffer space is allocated rather than bandwidth. In this scheme, the sources may have peaks which can be a large fraction of the link's capacity. Each node, maintains a state machine with two states for each virtual circuit. These two states are: active and idle. When a virtual circuit is in the active state, it is allocated a prespecified number of slots in the link's buffer, and it is guaranteed access to these buffer slots until the source becomes idle. Transitions of the state machine occur upon receipt of specially marked start and end cells. A start cell indicates the beginning of a burst and an end cell the end of a burst. All cells in a burst between the start cell and the end cell are marked as middle cells. The scheme also allows for transmission of single cells. These cells are treated as low priority cells with no guarantees of service. That is, they can get discarded if congestion arises. Cells, in general, can also be marked or unmarked. A marked cell has its CLP bit turned on and it can be discarded if a buffer becomes full.

Each node keeps the following information. For each virtual circuit i, it keeps the current state of the virtual circuit (active or idle), the pre-defined number of buffer slots s_i that have to be allocated when the virtual circuit becomes active, and the number of unmarked cells u_i belonging to the ith virtual circuit currently in the buffer. Also, it keeps track of the total number of unused slots in the buffer, K'. Unlike the previous scheme, when a source wants to transmit, it does not go through a request/validation procedure. It simply starts transmitting, having appropriately marked the start cell and the subsequent cells. When a node recognizes the start cell, it verifies whether it can allocate the pre-defined number of buffer slots or not. If the virtual circuit is in the idle state and $s_i > K'$, the start cell and the subsequent cells in the burst are discarded. On the other hand, if the virtual circuit is in the idle state and $s_i \leq K'$, the node accepts the burst. The state of the virtual circuit is changed to active, a timer for that virtual circuit is set, and s_i is deducted from K'. If $u_i < s_i$, then u_i is incremented by one. If $u_i = s_i$, the cell is marked (i.e. its CLP bit is turned on) and it is placed in the buffer. The timer is determined by the cell delay variation. If the timer expires before a middle cell or the end cell arrives, the status of the virtual circuit is changed to idle. We note that marking cells (i.e. set their CLP bit to on) permits the node to accept more than s_i cells from the ith virtual circuit. However, only s_i buffer slots are dedicated to the ith virtual circuit. That is, only s_i cells can be unmarked. The remaining cells are marked, and they can be dropped if new bursts from other virtual circuits arrive and the buffer becomes full. This introduces a form of fair sharing of the buffer. The buffer reservation mechanism can be equally applied to CBR sources.

Let R be the peak rate of a virtual circuit, C be the link's capacity, and K be the available buffer size. Then, the buffer slots allocated to the virtual circuit are given by the expression: $s_i = \lceil KR/C \rceil$. When selecting a route for a new virtual circuit, it is necessary to make sure that the new virtual circuit will be safely multiplexed with the already existing virtual circuits. A call admission procedure is prescribed.

A related work is by Doshi and Heffes [9, 10] that proposed a fast buffer allocation scheme for long file transfers.

3.5 Time windows

Several connection admission schemes have been based on the notion that a source is only allowed to transmit up to a maximum number of bits (or cells) within a fixed period of time. This fixed period of time is known by different names, such as frame and time window. This notion is similar to the jumping window that was proposed as a policing scheme.

Golestani [28] proposed a mechanism whereby for each connection, the number of cells transmitted on any link in the network is bounded. Thus, a smooth traffic flow is maintained throughout the network. This is achieved using the notion of frame, which is equal to a fixed period of time. The frame is not adjustable and it is the same for all links. Each connection can only transmit on a link up to a fixed number of cells per frame. Thus, the total number of cells transmitted by all connections on the same link is upper bounded. On a given switch, time on each incoming and outgoing link is organized into frames. Arriving frames over an incoming link are not synchronized with departing frames over an outgoing switch. A mechanism is proposed so that for each connection, the number of cells per frame transmitted on an outgoing link cannot exceed its upper bound. This mechanism is non work-conserving. However, a cell arriving at an input port in a given frame is guaranteed that it would be transmitted out of the switch at the end of an adjacent frame. This scheme requires buffering. Time windows were also proposed by Faber and Landweber[22].

Vakil and Singh [62] proposed a node to node flow-control mechanism. For each connection, the transmitting node can only transmit up to a certain number of cells every fixed time period. The number of cells it can transmit is specified by the receiving node. This is done using credits. The receiver informs the transmitter how many credits it can use for each connection per fixed period of time. If the credits for a particular connection are exhausted before the time period ends, then no more cells from this connection can be transmitted for the remaining of the time period. The receiver can dynamically modify the number of credits. This method requires buffering.

3.6 Other call admission control schemes

Dynamic bandwidth allocation was investigated by Tedijanto and Gün [58], Saito and Shiomoto [52], and Bolla, Danovaro, Davoli, and Marchese [3]. In this case, bandwidth allocated to a connection is dynamically adjusted every fixed time period. Related to dynamic bandwidth allocation are various reactive congestion control schemes that have been proposed in the literature. Contrary to an initial negative reaction towards these reactive schemes, it has been shown that they can be effective in cases where the source has an on period which is long compared to the round trip propagation delay, see for instance Periyannan [49]. These schemes, though they were developed specifically for cell-level congestion control, lend themselves to an approach for call admission control. See Gersht and Lee [25], Makrucki [45, 46], and Jagannath and Viniotis [36]. Recently, the ATM Forum adopted a feedback-base congestion control scheme referred to as Available Bit Rate (ABR).

Déjean, Dittman, and Lorenzen [8] and Lorenzen and Dittman [42] proposed a multi-path scheme which they referred to as the string mode protocol. The principal idea behind this scheme is that each burst is chopped into sub-bursts and each sub-burst is sent over a different virtual circuit. In view of this, a multi-path protocol can easily handle bursty sources with high peak bit rates compared to the capacity of a link.

Call admission control can be formulated as an optimization problem, where a particular reward function is optimized. See Gün, Kulkarni, and Narayanan [33], Bovopoulos [4], and Evans [21]. Also, neural nets have been used for call admission control. See Hiramatsu [35], Faragó [23], Nordström [47], and Gällmo, Nordström, Gustafsson, and Asplund [24].

An different approach for call admission control has been proposed by Gibbens, Kelly, and Key [27]. They propose using Bayesian decision theory to provide a simple and robust call admission scheme in the existence of uncertainties in the source average rate. A source is characterized by its peak rate and cell delay variation tolerance. Simple load-threshold rules are used for admission control. In this model, buffers are used for cell-scale congestion while burst level congestion is accounted for by a bufferless model.

Finally, call admission control schemes for virtual paths have been examined in Sato and Sato [55], and Sato, Ohta, and Tokizawa [54]. See also Yamamoto, Hirata, Ohta, and Tode [64].

4 Comparison of the performance of some call admission schemes

In this section we provide a numerical comparison of the equivalent capacity, the heavy traffic approximation, and Saito's upper bound of the cell loss probability (hereafter referred to as the CLP upper bound). These schemes were selected since they use the same set/subset of traffic descriptors. Namely, the peak bit rate, mean bit rate, and mean burst length of a call (R, ρ, b). (Note that the CLP upper bound scheme only utilizes the mean and peak bit rate information.) Before presenting the results, let us first define some necessary terms.

We will consider an ATM multiplexer consisting of a finite capacity queue of size K. This queue is served by a server (the outgoing link) of capacity C. The connections handled by this are classified into M classes, namely classes 1 through M (in this work we limit M to 2 for illustration purposes). That is, all the connections in the same class i have the same traffic descriptors (R_i, ρ_i, b_i).

Admission Region: This is the set of all values of $(n_1, n_2, \cdots, n_M)$, for which the cell loss probability is less than a small value ϵ, where n_i is the number of allocated class i connections, $i = 1, 2, \cdots, M$. In other words, this is the set of all combinations of the connections from the M classes for which the required cell loss probability In the numerical results given below with $M = 2$, we obtain the outermost boundary of the region. All points enclosed between the boundary and the axes represent combinations of connections from each class which fall in the admission region. ϵ is achievable.

Statistical gain: Now, let Let $Nmin_i$ be the number of class i connections admitted using peak rate allocation. So $Nmin_i = \lfloor 1/R_i \rfloor$. Likewise, define $Nmax_i$ to be the number of class i connections that can be admitted using mean rate allocation. So $Nmax_i = \lfloor 1/\rho_i \rfloor$. The statistical gain for a particular traffic class is defined as the maximum number, N_i, of connections admitted by a CAC scheme divided by the maximum number of connections that can be accepted using peak rate allocation $(Nmin)$ for a given acceptable bandwidth alloca-

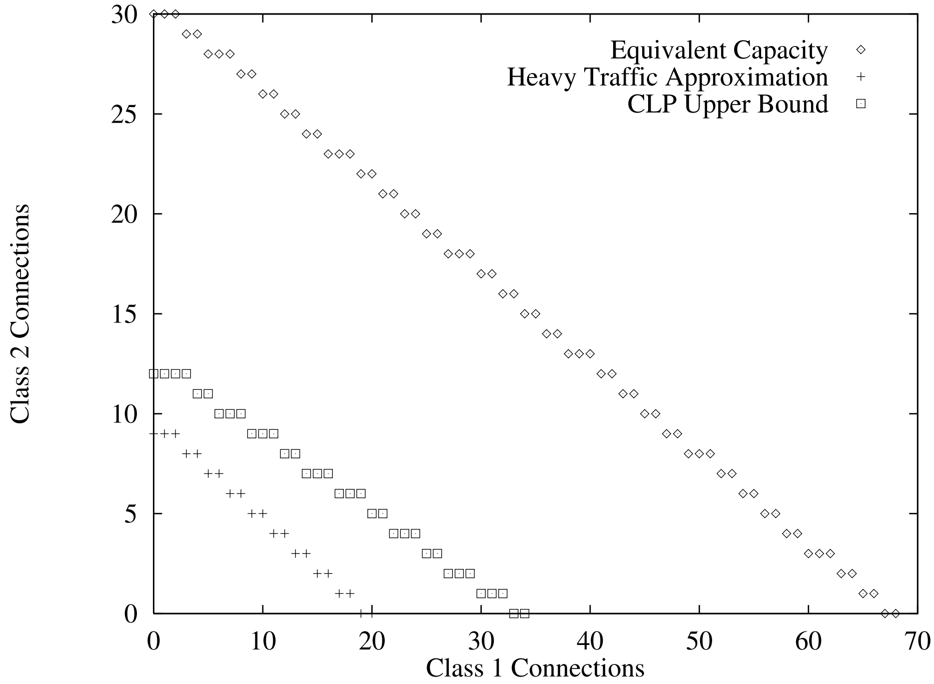


Figure 3: Admission regions for the CAC schemes, K=100, $\epsilon = 10^{-6}$.

tion for connection of the other classes. In the discussion below, we mean by statistical gain $N_i/Nmin_i$ when a single class of calls is exclusively using the multiplexer. In order for a CAC scheme to be effective it should be able to provide some statistical gain when possible.

Each of the three CAC schemes were implemented separately. The performance of these schemes relative to each other for various regions of input traffic parameters, buffer size, and required cell loss probability. Also, operating regions for which a particular scheme provides statistical gain over peak rate allocation were identified.

4.1 Case 1: Relatively Small Buffer Size

We consider the admission control of two classes assuming a relatively small buffer. The system parameters were chosen as follows. We set the required cell loss probability ϵ equal to 10^{-6}, buffer size K equal to 100, class 1 traffic is characterized by $(0.05, 0.01, 80)$, and class 2 traffic is characterized by $(0.1, 0.02, 50)$. This traffic characterization will also be used in the numerical examples given in the following sections. The minimum, $Nmin_i$, and maximum number, $Nmax_i$, of connections for class i, $i = 1, 2$, are respectively: $(Nmin_1, Nmax_1) = (20, 100)$ and $(Nmin_2, Nmax_2) = (10, 50)$.

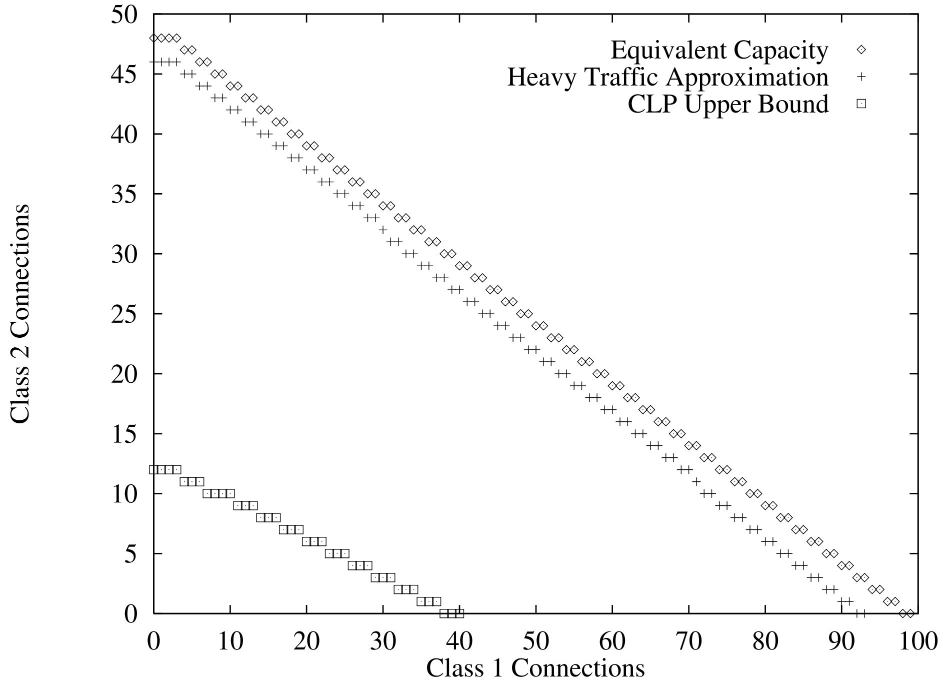


Figure 4: Admission regions for the CAC schemes, K=10,000, $\epsilon = 10^{-6}$.

The admission regions obtained for the three CAC methods are shown in figure 3.

The equivalent capacity scheme provided the largest admission region for this example. When a single class share the multiplexer, the statistical gain for classes 1 and 2 are respectively 3.4 and 3. Since the buffer size is small (relative to the mean burst lengths of each class), the heavy traffic approximation scheme coincides with the peak rate allocation. In order for the heavy traffic scheme to become effective, the ratio of the buffer size to burst length of each class must be large. The CLP upper bound scheme provides a conservative admission regions yielding a statistical gain for classes 1 and 2 of 1.7 and 1.2 respectively. This scheme is in general conservative with respect to the other schemes.

4.2 Case 2: Relatively Large Buffer Size

We assume the same parameters as in case 1, but the buffer size K is now increased by a 100-fold to 10,000. The admission regions for the three schemes are shown in figure 4.

Since the buffer size is increased to 10,000, the admission region of the equivalent capacity scheme grows in size. The statistical gain for classes 1 and 2 increases to 5 and 4.8 respectively. In this case, the equivalent capacity for a class i connection is almost equal to its mean bit rate, i=1,2.

In this example, the buffer size becomes large compared to the mean burst length of connection from class 1 or 2. This causes the admission region of the heavy traffic approximation scheme to grow in size as compared to the admission region when the buffer size is equal to 100. The statistical gain becomes 4.7 and 4.6 for classes 1 and 2 respectively. The admission region of the heavy traffic approximation scheme and that of the equivalent capacity scheme are very close.

For the CLP upper bound scheme, we observe that the maximum number of admitted connections from each class does not increase appropriately when setting the buffer size to 10,000. The maximum number for class 2 remains the same, while that of class 1 increases from 34 to 40. The reason for this is that class 1 has a lower peak rate than class 2. We note that in order for this scheme to provide statistical gain, we need to have traffic sources with small peak rate relative to the link capacity.

4.3 Effect of the Buffer Size

We now study the sensitivity of the selected CAC schemes to changes in the buffer size. Assuming that only class 1 connections are transported, we obtain the maximum number of admitted connections as a function of the buffer size. The buffer size is increased according to a geometric progression from 10 to 100,000 while the required cell loss probability ϵ is fixed at 10^{-6}. The results are plotted in figure 5. The figure indicates that the heavy traffic approximation scheme and the equivalent capacity scheme asymptotically admit the same number of connections as the buffer size approaches infinity.

The CLP upper bound scheme is less sensitive to the increase in buffer size. For this scheme, a strange phenomenon was observed when the buffer size is small. A temporary *drop* occurs to the maximum number of connections that can be admitted as the buffer increases. This is due to the effect of dividing ANA by MNA where MNA, a function of the buffer size and peak rate, must be an integer. So, by increasing the buffer size we get different values of ANA/MNA. We note also that increasing the buffer size above 1000 does not cause any increase in the number of admitted

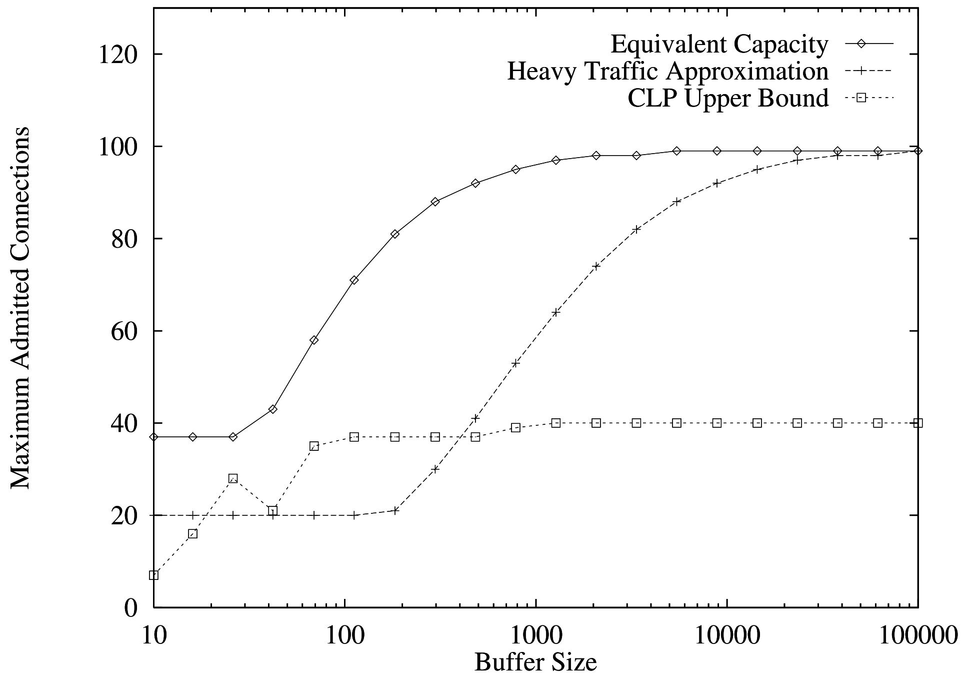


Figure 5: Maximum number of class 1 connections vs. buffer size, $\epsilon = 10^{-6}$.

connections.

4.4 Effect of the Required Cell Loss Probability

In this section, we study the sensitivity of the three CAC schemes to the changes in the required cell loss probability. Assuming that only class 1 connections are transported, we obtain the maximum number of admitted connections as a function of the required cell loss probability. We fix the buffer size at 200 and increase the cell loss probability from 10^{-9} to 10^{-3}. The results are plotted in figure 6.

From this figure, we observe that the equivalent capacity scheme is the least sensitive to the cell loss probability. In this particular case, the buffer size is large enough for the equivalent capacity scheme to admit a large number of connections even for a very small value of the required cell loss probability. The increase in the cell loss probability caused the maximum number of connections for class 1 to only increase from 75 to 91, not even reaching the maximum number of admittable connections, 99.

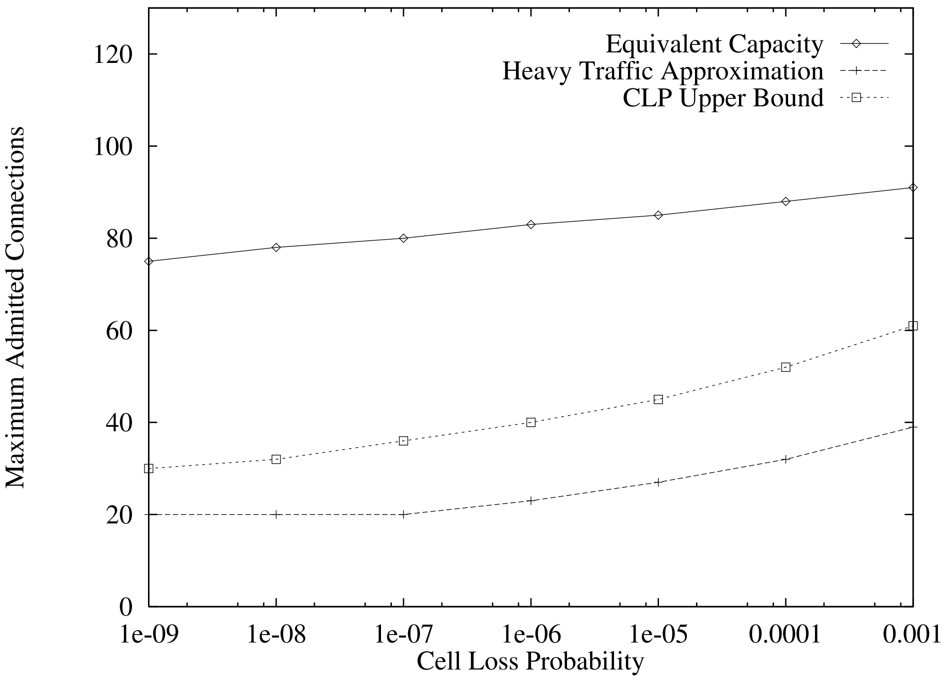


Figure 6: Maximum number of class 1 connections vs cell loss probability, $K = 200$.

The heavy traffic approximation scheme is more sensitive to the required cell loss probability than the equivalent capacity scheme. The maximum number of connections that can be admitted increased from 20 (no statistical gain) to 40, an increase by a factor of two.

The CLP upper bound method is also sensitive to the cell loss probability. In this example the increase in maximum number of connections is of the same magnitude as the heavy traffic method (from 30 to 61). However, the rate of increase is almost uniform while in the heavy traffic method, the cell loss probability started to affect the maximum number of connections admitted when it increased beyond 10^{-7}. We have observed similar sensitivity of the CLP upper bound method to the cell loss probability in other examples. It seems, therefore, that the required cell loss probability can indeed affect the admission region and the statistical gain achieved by the CLP upper bound method. The same can be said to a less extent about the heavy traffic approximation method. This is because in this method, the achieved statistical gain depends more on the ratio of the buffer size to the mean burst length(s).

4.5 Effect of the Ratio of the Buffer Size to the Mean Burst Length

We have already observed that the heavy traffic approximation scheme and the equivalent capacity scheme behave similarly when the buffer size is large. In this section, we study the effect of the ratio of the buffer size to the mean burst length of a connection, while keeping all other parameters fixed. The CLP upper bound scheme is excluded from this comparison since a) it has already been observed that its sensitivity to buffer size is poor, and b) it does not depend on the mean burst length. We consider a multiplexer with a single class of connections with descriptor $(0.04, 0.01, b)$. The mean burst length b is varied to take the values 10, 100, and 1000. For each value of b, the buffer size K is varied so that the ratio K/b varies from 0.1 to 100.

The results for the equivalent capacity and the heavy traffic approximation schemes are shown in figures 7(a) and figure 7(b) respectively. From these figures, it is interesting to note that as long as the ratio K/b is kept constant, the maximum number of admitted connections is almost the same regardless of the value of the mean burst length b. This observation can be used to approximate the solution of a multiplexer with a large buffer size by that of a multiplexer with a smaller buffer. The mean burst length of the source must be scaled down accordingly in order to keep the ratio K/b constant. We also note that the heavy traffic approximation scheme starts to provide a statistical gain when the ratio K/b increases to about 5.

References

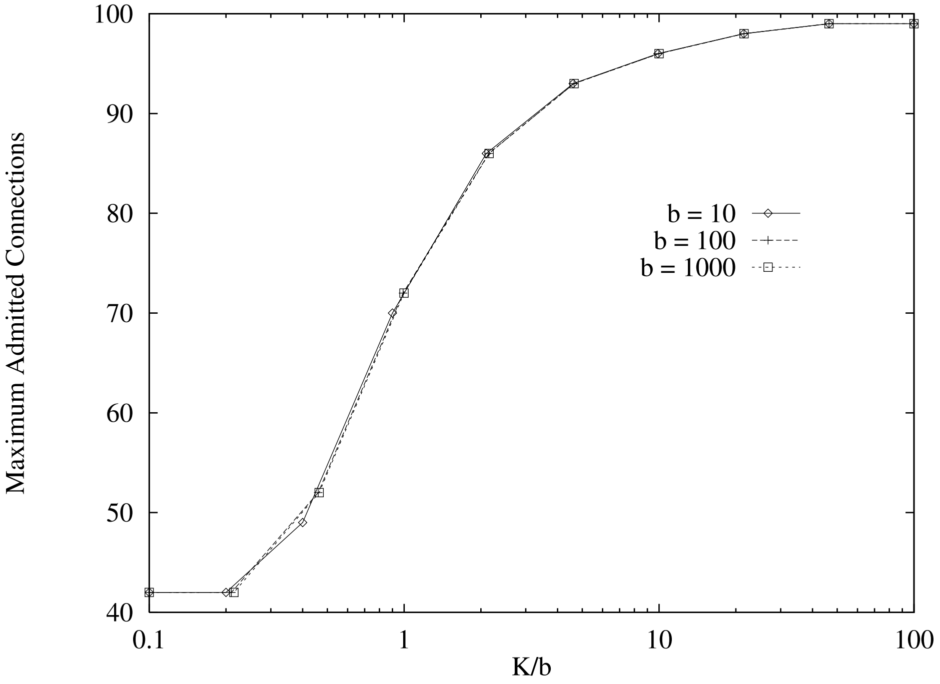
[1] ITU-T. Broadband Aspects of ISDN — ITU-T Recommendation I.121, 1991.

[2] D. Anick, D. Mitra, and M. M. Sondhi. Stochastic Theory of a Data-Handling System with Multiple Sources. *Bell Sys. Tech. J.*, 61:1871–1894, 1982.

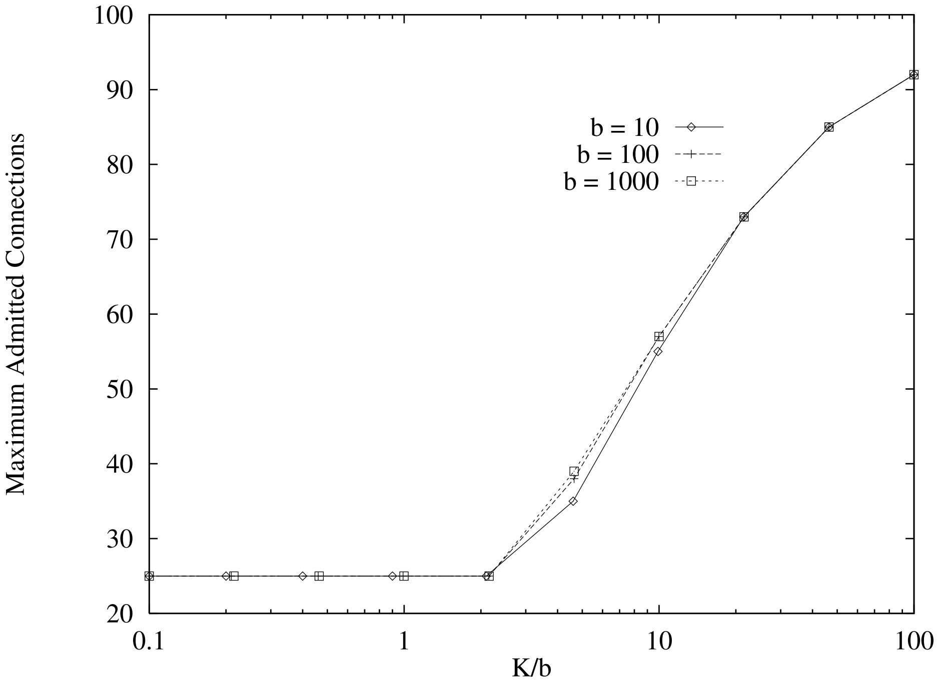
[3] R. Bolla, F. Danovaro, F. Davoli, and M. Marchese, An integrated dynamic resource allocation scheme for ATM networks, *INFOCOM '93*, 1288-1297.

[4] A. D. Bovopoulos, *Optimal burst level admission control in a broadband network*, Tech. Rept., Comp. Sci. Dept., Washington Univ., 1992.

[5] P. Castelli, E. Cavallero, and A. Tonietti, Policing And Call Admission Problems in ATM Networks, in: A. Jensen and V.B. Iversen (Eds.), *Teletraffic and datatraffic in a period of change*, (North-Holland, 1991, 847-852.



(a) Equivalent Capacity



(b) Heavy Traffic

Figure 7: Effect of K/b.

[6] C.-S. Chang and J. A. Thomas. Effective Bandwidth in High Speed Networks. *IEEE Journal on Selected Areas in Communications*, 13:1091–1100, 1995.

[7] G. L. Choudhury, D. M. Lucantoni, and W. Whitt. On the Effective Bandwidths for Admission Control in ATM Networks. In *Proceedings of 14th International Teletraffic Congress (ITC)*, pages 411–420, 1994.

[8] J. H. Déjean, L. Dittman, and C. N. Lorenzen, String mode - a new concept for performance improvement of ATM networks, *IEEE JSAC* **9**, 1991, 1452-1460.

[9] B. T. Doshi and H. Heffes, Performance of an in-call buffer-window reservation/allocation scheme for long file transfers, *IEEE JSAC* **9**, 1991, 1013-1023.

[10] B. T. Doshi and H. Heffes, Overload performance of an adaptive, buffer-window allocation scheme for a class of high speed networks, in: A. Jensen and V.B. Iversens (Eds.), *Teletraffic and datatraffic in a period of change*, North-Holland, 1991 441-446.

[11] B. T. Doshi. Deterministic Rule Based Traffic Descriptors for Broadband ISDN: Worst Case Behavior and Connection Acceptance Control. In *Proceedings of 14th International Teletraffic Congress (ITC)*, pages 591–600, 1994.

[12] N. G. Duffield, J. T. Lewis, and N. O'Connell. Predicting Quality of Service for Traffic with Long-Range Fluctuations. In *Proceedings of the International Conference on Communications (ICC)*, pages 473–477, 1995.

[13] N. G. Duffield, J. T. Lewis, N. O'Connel, R. Russell, and F. Toomey. Entropy of ATM Traffic Streams: A Tool for Estimating QoS Parameters. *IEEE Journal on Selected Areas in Communications*, 13:981–990, 1995.

[14] K. Elsayed, On the Superposition of Discrete-time Markov Renewal Processes and Applications to Statistical Multiplexing of Bursty Traffic Sources, *GLOBECOM '94*.

[15] K. Elsayed and H. G. Perros. An Efficient Algorithm for Characterizing the Superposition of Multiple Heterogeneous Interrupted Bernoulli Processes. In *Proceedings of Second International Workshop on Numerical Solution of Large Markov Chains*, January 1995.

[16] K. Elsayed and H. G. Perros. Analysis of an ATM Statistical Multiplexer with Heterogeneous Markovian On/Off Sources and Applications to Call Admission Control. Submitted for publication and available via the web through the URL: ftp://eceyv.ncsu.edu/pub/papers/stat-mux-n-cac.ps.gz, 1995.

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

[18] A. Elwalid, D. Heyman, T. V. Lakshman, D. Mitra, and A. Weiss. Fundamental Bounds and Approximations for ATM Multiplexers with Applications to Video Teleconferencing. *IEEE Journal on Selected Areas in Communications*, 13:1004–1016, 1995.

[19] A. Elwalid, D. Mitra, and R. H. Wentworth. A new Approach for Allocating Buffers and Bandwidth to Heterogeneous Regulated Traffic in an ATM Node. *IEEE Journal on Selected Areas in Communications*, 13:1115–1127, 1995.

[20] A. Erramilli, J. Gordon, and W. Willinger. Applications of Fractals in Engineering for Realistic Traffic Processes. In *Proceedings of 14th International Teletraffic Congress (ITC)*, pages 35–44, 1994.

[21] S. P. Evans, Optimal resource management and capacity allocation in a broadband integrated services network, in: P.J.B. King, I. Mitrani, and R.J. Pooley (Eds.), *Performance '90*, Elsevier, 1990, 159-173.

[22] T. Faber and L. Landweber, Dynamic time windows: packet admission control with feedback, *SIGCOMM '92*, 124-135.

[23] A. Faragó, A neural structure as a tool for optimizing routing and resource management in ATM networks, *IWANNT '93,* Princeton.

[24] O. Gällmo, E.E. Nordström, M. Gustafsson, and L. Asplund, *Neural networks for preventive traffic control in broadband ATM networks,* Tech. Rept., Comp. Sci. Dept., Univ. of Uppsala, 1993.

[25] A. Gersht and K.L. Lee, A congestion control framework for ATM networks, *IEEE JSAC* **9**, 1991, 1119-1130.

[26] R. J. Gibbens and P. J. Hunt, Effective bandwidths for the multi-type UAS channel, *Queueing Systems* **9** , 1991, 17-26.

[27] R. J. Gibbens, F. P. Kelly, and P. B. Key. A Decision-theoretic Approach to Call Admission Control in ATM Networks. *IEEE Journal on Selected Areas in Communications*, 13:1101–1113, 1995.

[28] S. J. Golestani, Congestion-free communication in broadband packet networks, *IEEE Trans. Comm.* **39**, 1991, 1802-1812.

[29] R. Grünenfelder, J. P. Cosmas, S. Manthorpe, and A. Odinam-Okafor. Characterization of Video Codecs as Autoregressive Moving Average Processes and Related Queueing Systems Performance. *IEEE Journal on Selected Areas in Communications*, 9:284–293, 1991.

[30] R. Guérin, H. Ahmadi, M. Naghshineh, Equivalent capacity and its application to bandwidth allocation in high-speed networks, *IEEE JSAC* **9**, 1991, 968-981.

[31] R. Guérin and L. Gün, A unified approach to bandwidth allocation and access control in fast packet-switched networks, *INFOCOM '92,* 1-12.

[32] J. Guibert. Overflow Probability Upper Bound for Heterogeneous Fluid Queues handling on-off Sources. In *Proceedings of 14th International Teletraffic Congress (ITC)*, pages 65–74, 1994.

[33] L. Gün, V.G. Kulkarni, and A. Narayanan, Bandwidth allocation and access control in high-speed networks, *Annals of Oper. Res.*, to appear.

[34] D. P. Heyman, A. Tabatabai, and T. V. Lakshman. Statistical Analysis and Simulation Study of Video Teleconference in ATM Networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 2:49–59, 1992.

[35] A. Hiramatsu, Integration of ATM call admission control and link capacity control by distributed neural networks, *IEEE JSAC* **9**, 1991, 1131-1138.

[36] S. V. Jagannath and I. Viniotis, A Novel Architecture and Flow Control Scheme for Private ATM Networks, H.G. Perros (Ed.), *High-Speed Communication Networks*, Plenum, 1992, 97-108.

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

[38] G. Kesidis, J. Walrand, and C.-S. Chang. Effective Bandwidths for Multiclass Markov Fluids and Other ATM Sources. *IEEE Transactions on Networking*, 1(4):424–428, August 1993.

[39] V. Kulkarani, L. Gün, and P. Chimento, Effective bandwidth vector for two-priority ATM traffic, *INFOCOM '94*, 1056-1064.

[40] W. E. Leland, M.S. Taqqu, W. Willinger, and D. V. Wilson. On the Self-Similar Nature of Ethernet Traffic (Extended Version). *IEEE Transactions on Networking*, 2:1–15, 1994.

[41] S.-Q. Li. A General Solution Technique for Discrete Queueing Analysis of Multimedia Traffic on ATM. *IEEE Transactions on Communications*, 39:1115–1132, 1991.

[42] C. N. Lorenzen and L. Dittman, Evaluation of the string mode protocol in an ATM networks, in: H.G. Perros, G. Pujolle, and Y. Takahashi (Eds.), *Modelling and performance evaluation of the ATM technology*, North-Holland, 1993, 211-227.

[43] D. M. Lucantoni, M. F. Neuts, and A. R. Reibman. Methods for Performance Evaluation of VBR Video Traffic Models. *IEEE Transactions on Networking*, 2:176–180, 1994.

[44] B. Magalaris, D. Anastassiou, P. Sen, G. Karlson, and J. Robbins. Performance Models of Statistical Multiplexing in Packet Video Communications. *IEEE Transactions on Communications*, 36:834–843, 1988.

[45] B. Makrucki, On the performance of submitting excess traffic to ATM networks, Tech. Rept. BellSouth, Science and Technology, 1990.

[46] B. Makrucki, Explicit forward congestion notification in ATM networks, in: H. G. Perros, (Ed.), *High-speed communication networks,* Plenum Press, 1992, 73-96.

[47] E. Nordström, A hybrid admission control scheme for broadband ATM traffic *IWANNT '93,* Princeton.

[48] H. G. Perros and A. A. Nilsson and H-C Kuo, Analysis of traffic measurement in the Vistanet gigabit networking testbed, Proceedings of the High Performance Networking, North Holland 1994, 313-323.

[49] A. Periyannan, M.S. thesis, Comp. Sci. Dept., NC State Univ., 1992.

[50] C. Rasmussen, J.H. Sørensen, K.S. Kvols, and S.B. Jacobsen, Source-independent call acceptance procedures in ATM networks, *IEEE JSAC* **9**, 1991, 351-358.

[51] K. M. Rege. Equivalent Bandwidth and Related Admission Criteria for ATM Systems-A Performance Study. *International Journal of Communications Systems*, 7:181–197, 1994.

[52] H. Saito and K. Shiomoto, Dynamic call admission control in ATM networks, *IEEE JSAC* **9**, 1991, 982-989.

[53] H. Saito, Call admission control in an ATM network using upper bound of cell loss probability, *IEEE Trans. Comm.* **40** , 1992, 1512-1521.

[54] K.-I. Sato, S. Ohta, and I. Tokizawa, Broadband ATM network architecture based on virtual paths, *IEEE Trans. Comm.* **38**, 1990, 1212-1222.

[55] Y. Sato and K. Sato, Evaluation of statistical cell multiplexing effects and path capacity design in ATM networks, *IECE Trans. Comm.* **E75-B**, 1992, 642-648.

[56] K. Sohraby, On the asymptotic behavior of heterogeneous statistical multiplexer with applications, *INFOCOM'92*, 839-847.

[57] K. Sohraby, On the Theory of General On-Off Sources With Applications in High-Speed Networks, *INFOCOM'93*, 401-410.

[58] T. E. Tedijanto and L. Gün, Effectiveness of dynamic bandwidth management mechanisms in ATM networks, *INFOCOM '93*, 358-367.

[59] D. P. Tranchier, P. E. Boyer, Y. M. Rouaud, and J.-Y. Mazeas, *Fast bandwidth allocation in ATM networks*, Tech. Rept., CNET-Lannion, 1992.

[60] J. S. Turner, *A proposed bandwidth management and congestion control scheme for multicast ATM networks*, Tech. Rept., Computer and Communications Research Center, Washington Univ., 1991.

[61] J. S. Turner, Bandwidth management in ATM networks using fast buffer reservation, *Proc. Australian Broadband Switching and Services Symposium*, Melbourne 15-17 July 1992.

[62] F. Vakil and R. P. Singh, Shutter: A flow control scheme for ATM networks, 7th ITC Specialists Seminar, Morristown, Oct. 1990.

[63] G. De Veciana, G. Kesidis, and J. Walrand. Resource Management in Wide-Area ATM Networks Using Effective Bandwidth. *IEEE Journal on Selected Areas in Communications*, 13:1081–1090, 1995.

[64] M. Yamamoto, T. Hirata, C. Ohta, and H. Tode, Traffic control scheme for interconnection of FDDI networks through ATM network, *INFOCOM '93*, 411-420.