Traffic Management: A Review of Call Admission Control Schemed for ATM Networks

Khaled M. Fuad Elsayed
Department of Communications and Electronics
Faculty of Engineering, Cairo University
Giza, Egypt 12613
kelsayed@alpha1-eng.cairo.eun.eg

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

Abstract

Connection Admission Control (CAC) is one of the primary mechanisms for traffic management in ATM networks. A substantial number of CAC schemes have been proposed. In this paper, we review the salient features of some of these algorithms.

1 Introduction

In recent years, there has been a tremendous growth in the development and deployment of ATM networks. One area which is of significant importance to ATM networks is traffic management. Congestion control is one of the primary mechanisms for traffic management. 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 schemes 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: *connection admission control* (CAC) and *bandwidth enforcement*. 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 *connection 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/routers. On each of these ATM switches, resources have to be allocated to the new connection.

The connection 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 into a) non-statistical allocation, or peak bandwidth allocation, and b) statistical allocation. Below, we examine these two cases. As will be seen, it is challenging to design good connection 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 1, each output port is provided with a finite (non-shared) 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 ϵ for the existing traffic. Assuming that the new connection is accepted, would the cell loss probability be also of the order of ϵ for the total traffic carried by the port?

1.1 Non-statistical allocation (Peak bandwidth allocation)

Suppose a source has an average bandwidth of 2 Mb/s and a peak bandwidth of 5 Mb/s. Peak bandwidth allocation, otherwise known as non-statistical allocation, requires that 5 Mb/s be reserved at the output port for the specific source, independent of whether the source transmits continuously at 5 Mb/s. 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. 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

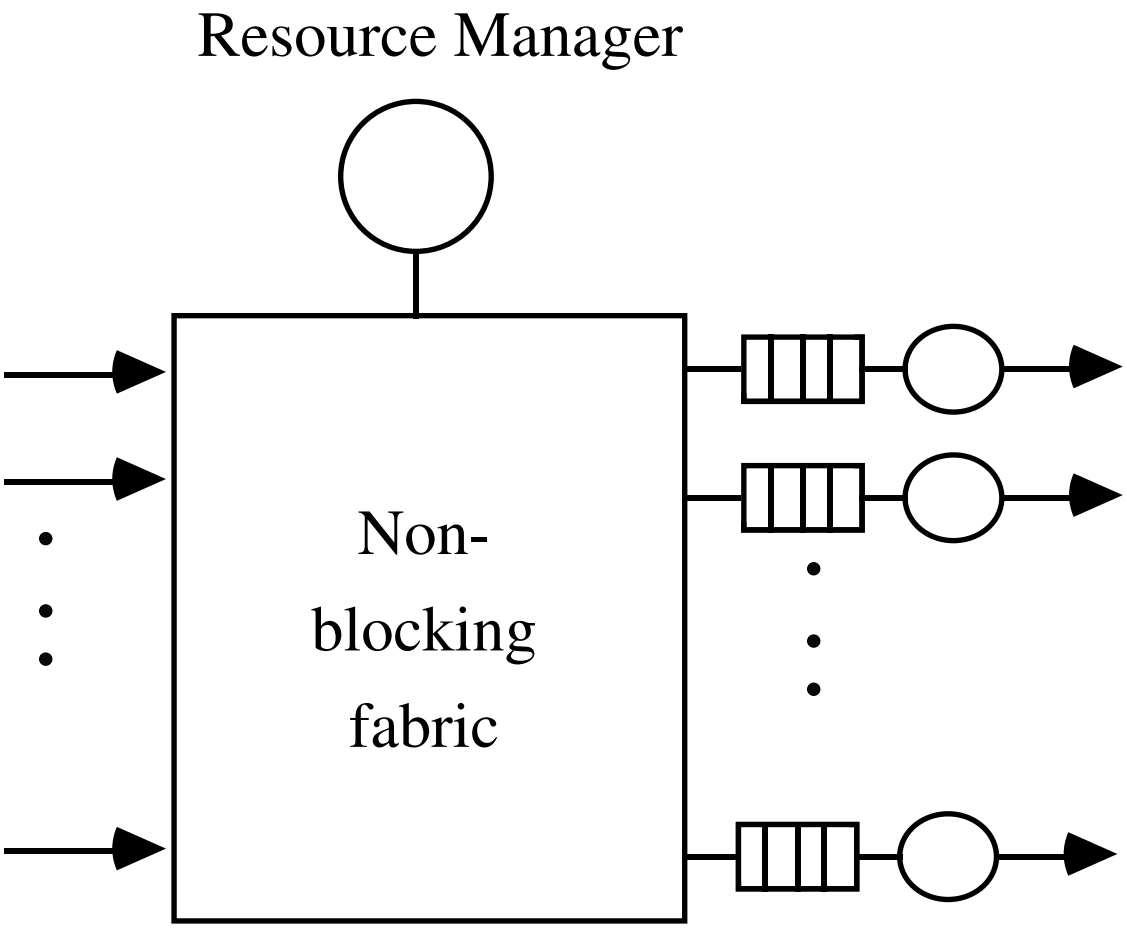

Figure 1: An ATM switch with output buffering

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 may 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 the arrival process of ATM cells and lack of understanding as to how this arrival process is shaped deep in the ATM network.

Another difficulty in designing a connection 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 connection or not may be formulated as a queueing problem. For instance, let us consider the non-blocking switch shown in figure 1. The connection 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 [1] and Li [2]). It gets even more complicated, if we assume complex arrival processes. 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.

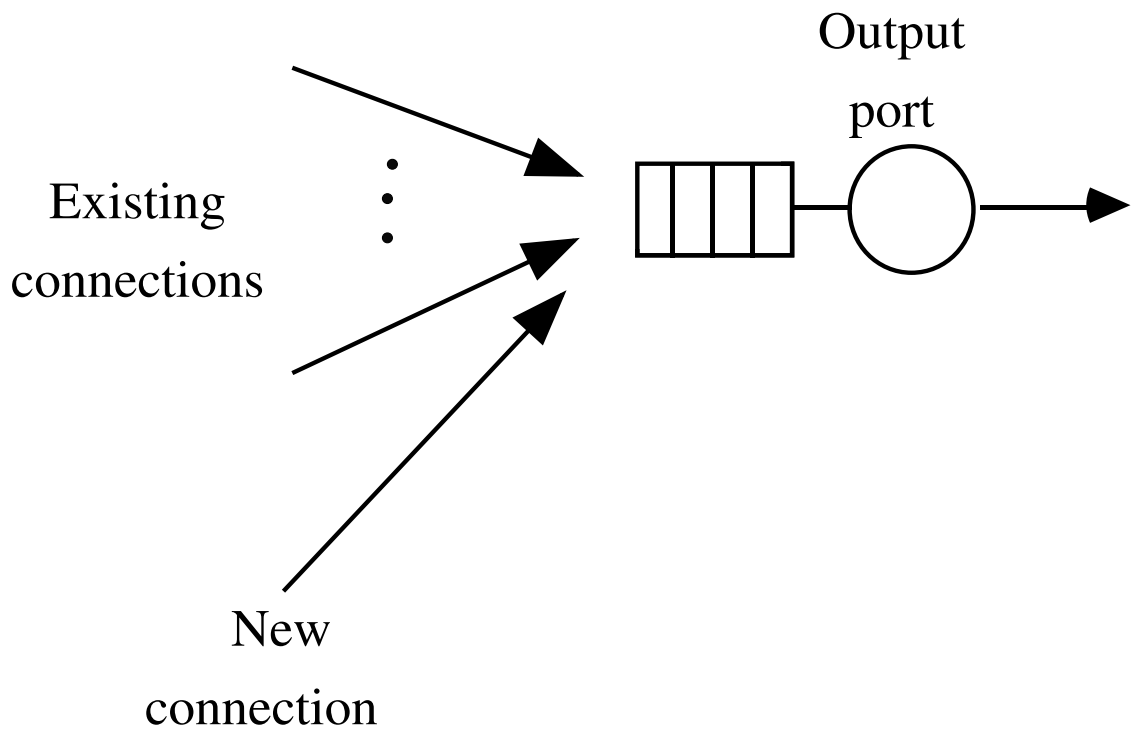

Figure 2: An ATM multiplexer

In this paper, we will review some of the connection 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.

Over the last few years, we have gone through several paradigm shifts regarding our understanding as to 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 were in the form of the Markov Modulated Poisson Process (MMPP), its discrete-time counterpart Markov Modulated Bernoulli Process (MMBP), and the 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 [3]). 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 behavior of an ATM multiplexer under on/off periods with arbitrarily distributed on and off periods can be quite difficult (see Elsayed [4] and Guibert [5].

Finally, we should mention that several auto-regressive types of models have been proposed to characterize the traffic due to video (see for example Magalaris et al. [6], Heyman, Tabatabai and Lakshman [7], and Grünenfelder et al. [8]). Also, 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. [9], Erramilli, Gordon and Willinger [10] and Duffield, Lewis and O'Connel [11] 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, sustainable 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 sustainable rate. These parameters are fairly inadequate when it comes to bandwidth allocation. For, it can be verified 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, the 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 times are affected as the traffic from 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.

3 Classification of connection admission schemes

A variety of different connection 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 as follows:

1. Effective bandwidth
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 a scheme. Below, we discuss the salient features of each class of CAC schemes and review some of the proposed

schemes within each class.

3.1 Effective bandwidth

Let us consider a single source feeding a finite capacity queue. Then, the effective bandwidth of the source is the service rate of the queue that corresponds to a cell loss of ϵ. The effective bandwidth for a single source can be derived as follows, see Guérin, Ahmadi, and Naghshineh [12]. 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. The effective bandwidth e is given by:

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

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 effective bandwidth is again the service rate e which ensures that the cell loss for all the sources is ϵ. Guérin, Ahmadi, and Naghshineh [12] proposed the following approximation for multiple sources:

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

where

- e_i is the equivalent capacity of the ith source calculated using expression, and $\sum_{i=1}^{N} e_i$ 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

- $a' = \sqrt{-2ln(\epsilon) - ln2\pi}$.

Elwalid and Mitra [13] showed that the effective bandwidth 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. Some studies (see Choudhury, Lucantoni, and Whitt [14] and Elsayed and Perros [15]) have clearly indicated the inaccuracy of effective bandwidth methods in some situations. In particular, the effective bandwidth method fails when a bufferless system subject to the same input traffic has a small probability that the traffic load exceeds the link capacity. In the effective bandwidth approach, this probability is assumed to be close to one (and is taken as one in the calculations). Rege [16] compares various approaches for effective bandwidth and proposes some modifications to enhance the accuracy of the scheme. Elwalid et al. [17] proposed a method in which they combined Chernoff bounds and effective bandwidth approximation to overcome the shortcomings of the effective bandwidth. This method provides better accuracy than effective bandwidth for the case mentioned above of a bufferless multiplexer that can achieve substantial statistical gain. However, in some other cases, the method does not solve all the problems with the inaccuracy of effective bandwidth.

Kulkarni, Gün, and Chimento [18] considered the effective bandwidth vector for two-priority on/off source. Chang and Thomas [19] introduced a *calculus* for evaluating source effective bandwidth at output of multiplexers and upon demultiplexing or routing. On-line evaluation of effective bandwidth have been proposed by De Veciana, Kesidis and Walrand [20]. Duffield et al. [21] proposed maximum entropy as a method for characterizing traffic sources and their effective bandwidth. Further relevant references are Gibbens and Hunt [22], Kelly [23], Kesidis, Walrand and Chang [24], Guérin and Gün [25] and Dziong, Juda, and Mason [26].

3.2 Heavy traffic approximation

Sohraby [27] proposed an approximation for bandwidth allocation based on the asymptotic behavior of the tail of the queue-length distribution It is known that the steady-state queue-length distribution exhibits a geometrically distributed tail. For sufficiently large i, Sohraby suggests

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

where

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

where $r = \sum_{i=1}^{N} r_i R_i$. 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

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

This is a good approximation when there is a large number of sources with each source peak rate very small compared to link capacity. Only in this case, the system may be operated (efficiently) at the heavy traffic region (e.g. utilization above 85for obtaining the results above. If this condition is violated, the method provides inaccurate results.

3.3 Upper bounds of the cell loss probability

Several other connection admission schemes have been proposed which are based on an upper bound for the cell loss probability. Saito [28] 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

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

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

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

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 $D/2$.

The main problem of this method is the absence of burst size in the calculation and thus a worst case behavior is assumed for the source. This method works well in the case when the actual source behavior is close to the worst case behavior assumed in the above calculation.

For other upper bounds on the cell loss probability see Rasmussen et al. [29], Castelli, Cavallero, and Tonietti [30], Doshi [31] and the closely related work by Elwalid, Mitra, and Wentworth [32].

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 [33] 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 [34, 35] 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 connection admission procedure is prescribed.

A related work is by Doshi and Heffes [36, 37]. They 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 [38] 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[39].

Vakil and Singh [40] 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 connection admission control schemes

Dynamic bandwidth allocation was investigated by Tedijanto and Gün [41], Saito and Shiomoto [42], Bolla, Danovaro, Davoli, and Marchese [43] and Jamin et al. [44].

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 [45]. These schemes, though they were developed specifically for cell-level congestion control, lend themselves to an approach for connection admission control. See Gersht and Lee [46], Makrucki [47, 48], and Jagannath and Viniotis [49]. Recently, the ATM Forum adopted a feedback-base congestion control scheme referred to as Available Bit Rate (ABR).

Déjean, Dittman, and Lorenzen [50] and Lorenzen and Dittman [51] 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 Bovopoulos [52], and Evans [53]. Also, neural nets have been used for connection admission control, see Hiramatsu [54], Faragó [55], Nordström [56], Gällmo, Nordström, Gustafsson, and Asplund [57], Uehara and Hirota [58], and Youssef, Habib, and Saadawi[59].

A different approach for connection admission control has been proposed by Gibbens, Kelly, and Key [60]. They propose using Bayesian decision theory to provide a simple and robust connection 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.

Connection admission control schemes for virtual paths have been examined in Sato and Sato [61], and Sato, Ohta, and Tokizawa [62]. See also Yamamoto, Hirata, Ohta, and Tode [63]. Finally, additional references can be found in [64].

References

[1] 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.

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

[3] 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, 313–323, North Holland 1994.

[4] K. Elsayed, On the Superposition of Discrete-Time Markov Renewal Processes and Applications to Statistical Multiplexing of Bursty Traffic Sources, *GLOBECOM'94*i, 1994.

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

[6] 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.

[7] 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.

[8] 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 JSAC*, 9:284–293, 1991.

[9] 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.

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

[11] 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)*, 473–477, 1995.

[12] R. Guérin, H. Ahmadi, M. Naghshineh, Equivalent Capacity and its Application to Bandwidth Allocation in High-Speed Networks, *IEEE JSAC*, 9:968–981, 1991.

[13] A. Elwalid and D. Mitra, Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High-Speed Networks, *IEEE Transactions on Networking*, 1:329–343, 1993.

[14] 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)*, 411–420, 1994.

[15] K. Elsayed and H. G. Perros. Analysis of an ATM Statistical Multiplexer with Heterogeneous Markovian On/Off Sources and Applications to Call Admission Control. To appear, *Journal of High Speed Networks*, 1997.

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

[17] 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 JSAC*, 13:1004–1016, 1995.

[18] V. Kulkarani, L. Gün, and P. Chimento, Effective Bandwidth Vector for Two-Priority ATM Traffic, *INFOCOM'94*, 1056–1064, 1994.

[19] C.-S. Chang and J. A. Thomas. Effective Bandwidth in High Speed Networks. *IEEE JSAC*, 13:1091–1100, 1995.

[20] G. De Veciana, G. Kesidis, and J. Walrand. Resource Management in Wide-Area ATM Networks Using Effective Bandwidth. *IEEE JSAC*, 13:1081–1090, 1995.

[21] 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 JSAC*, 13:981–990, 1995.

[22] R. J. Gibbens and P. J. Hunt, Effective Bandwidths for the Multi-Type UAS Channel, *Queueing Systems*, 9:17–26, 1991.

[23] F. P. Kelly, Effective bandwidths at Multi-Class Queues, *Queueing Systems*, 9:5–16, 1991.

[24] 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, 1993.

[25] R. Guérin and L. Gün, A Unified Approach to Bandwidth Allocation and Access control in Fast Packet-Switched Networks, *INFOCOM'92,*, 1–12, 1994.

[26] Z. Dziong, M. Juda, and L. G. Mason, A Framework for Bandwidth Management in ATM Networks–Aggregate Equivalent Bandwidth Estimation Approach, *IEEE Transactions on Networking*, 5(1):134–147, 1997.

[27] K. Sohraby, On the Asymptotic Behavior of Heterogeneous Statistical Multiplexer with Applications, *INFOCOM'92*, 839–847, 1992.

[28] H. Saito, Call Admission Control in an ATM Network Using Upper-Bound of Cell Loss Probability, *IEEE Transactions on Communications*, 40:1512–1521, 1992

[29] C. Rasmussen, J.H. Sørensen, K.S. Kvols, and S.B. Jacobsen, Source-Independent Call Acceptance Procedures in ATM networks, *IEEE JSAC*, 9:351–358, 1991.

[30] 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, 847–852, 1991.

[31] 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)*, 591–600, 1994.

[32] 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 JSAC*, 13:1115–1127, 1995.

[33] D. P. Tranchier, P. E. Boyer, Y. M. Rouaud, and J.-Y. Mazeas, *Fast Bandwidth Allocation in ATM Networks*, Tech. Rept., CNET-Lannion, 1992.

[34] 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.

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

[36] B. T. Doshi and H. Heffes, Performance of an in-call Buffer-Window Reservation/Allocation Scheme for Long File Transfers, *IEEE JSAC*, 9:1013–1023, 1991.

[37] 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, 441–446, 1991.

[38] S. J. Golestani, Congestion-Free Communication in Broadband Packet Networks, *IEEE Transactions on Communications*, 39:1802-1812, 1991.

[39] T. Faber and L. Landweber, Dynamic Time Windows: Packet Admission Control with Feedback, *SIGCOMM'92*, 124–135, 1992.

[40] F. Vakil and R. P. Singh, Shutter: A Flow Control Scheme for ATM Networks, 7th ITC Specialists Seminar, Morristown, Oct. 1990.

[41] T. E. Tedijanto and L. Gün, Effectiveness of Dynamic Bandwidth Management Mechanisms in ATM Networks, *INFOCOM'93*, 358–367, 1993.

[42] H. Saito and K. Shiomoto, Dynamic Call Admission Control in ATM Networks, *IEEE JSAC*, 9:982–989, 1991.

[43] R. Bolla, F. Danovaro, F. Davoli, and M. Marchese, An Integrated Dynamic Resource Allocation Scheme for ATM Networks, *INFOCOM'93*, 1288–1297, 1993.

[44] S. Jamin, P. B. Danzig, S. J. Shenker, and L. Zhang, A Measurement-Based Admission Control Algorithm for Integrated Service Packet Networks, *IEEE Transactions on Networking*, 5(1):56–70, 1997.

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

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

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

[48] B. Makrucki, Explicit forward congestion notification in ATM networks, in: H. G. Perros, (Ed.), *High-Speed Communication Networks,* Plenum Press, 73–96, 1992.

[49] 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, 97–108, 1992.

[50] J. H. Déjean, L. Dittman, and C. N. Lorenzen, String Mode - a New Concept for Performance Improvement of ATM Networks, *IEEE JSAC*, 9:1452–1460, 1991.

[51] C. N. Lorenzen and L. Dittman, Evaluation of the String Mode Protocol in an ATM Network, in: H. G. Perros, G. Pujolle, and Y. Takahashi (Eds.), *Modelling and Performance Evaluation of the ATM Technology*, North-Holland, 211–227, 1993.

[52] A. D. Bovopoulos, *Optimal Burst Level Admission Control in a Broadband Network*, Tech. Rept., Comp. Sci. Dept., Washington Univ., 1992.

[53] 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, 159–173, 1990.

[54] A. Hiramatsu, Integration of ATM Call Admission Control and Link Capacity Control by Distributed Neural Networks, *IEEE JSAC*, 9:131–1138, 1991.

[55] A. Faragó, A Neural Structure as a Tool for Optimizing Routing and Resource Management in ATM Networks, *IWANNT'93,* Princeton, 1993.

[56] E. Nordström, A Hybrid Admission Control Scheme for Broadband ATM Traffic *IWANNT'93,* Princeton, 1993.

[57] 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.

[58] K. Uehara and K. Hirota, Fuzzy Connection Admission Control for ATM Netowrks Based on Possibility Distribution of Cell Loss Ratio, *IEEE JSAC*, 15:179–190, 1997.

[59] S. A. Youssef,,, I.W. Habib, and T. N. Saadawi, A Neurocomputing Controller for Bandwidth Allocation in ATM Networks, *IEEE JSAC*, 15:191–199, 1997.

[60] R. J. Gibbens, F. P. Kelly, and P. B. Key. A Decision-Theoretic Approach to Call Admission Control in ATM Networks. *IEEE JSAC*, 13:1101–1113, 1995.

[61] Y. Sato and K. Sato, Evaluation of Statistical Cell Multiplexing Effects and Path Capacity Design in ATM Networks, *IECE Trans. Comm.*, E75-B:642–648, 1992.

[62] K.-I. Sato, S. Ohta, and I. Tokizawa, Broadband ATM Network Architecture Based on Virtual Paths, *IEEE Transactions on Communications*, 1212–1222, 1990.

[63] M. Yamamoto, T. Hirata, C. Ohta, and H. Tode, Traffic Control Scheme for Interconnection of FDDI Networks through ATM Network, *INFOCOM'93*, 411–420, 1993.

[64] IEEE Communications Magazine, special issue on Bandwidth Allocation in ATM Networks. Vol. 35 No. 5, May 1997.