4-1997

An Adaptive Scheme for Admission Control in ATM Networks

Saragur M. Srinidhi
NASA Lewis Research Center

William H. Thesling
Cleveland State University

Vijaya K. Konangi
Cleveland State University, V.KONANGI@csuohio.edu

Publisher's Statement

Original Citation

Srinidhi, S. M., Thesling, W. H., , & Konangi, V. K. (1997). An adaptive scheme for admission control in ATM networks. Computer Networks and ISDN Systems, 29(5), 569-582.

Repository Citation

Srinidhi, Saragur M.; Thesling, William H.; and Konangi, Vijaya K., "An Adaptive Scheme for Admission Control in ATM Networks" (1997). *Electrical Engineering & Computer Science Faculty Publications*. 71.
https://engagedscholarship.csuohio.edu/enece_facpub/71

An adaptive scheme for admission control in ATM networks

Saragur M. Srinidhi [a,*], William H. Thesling [b], Vijaya K. Konangi [b,1]

[a] *Sterling Software (US) Inc. [2], NASA Lewis Research Center, MS 142-1, 21000 Brookpark Road, Cleveland, OH 44135, USA*
[b] *Department of Electrical Engineering, Cleveland State University, SH-332, Euclid Avenue at E24th Street, Cleveland, OH 44115, USA*

1. Introduction

A common problem in designing telecommunication systems is determining an effective mechanism for allocating system resources, like buffers or channels, in order to attain and maintain the desired system performance. Due to the reusability of telecommunication resources, this problem is different from classic inventory problems in which perishable or repairable items are considered. An acceptable strategy should adapt to traffic fluctuations and should be implementable in a distributed processing environment.

ATM networks are characterized by their switch-based network architecture, and the communication between any pair of devices is established through a switch. This switched network is capable of supporting multiple connections simultaneously. The aggregate bandwidth of an ATM switch is expected to be several Gbps. The most important advantage of ATM is its flexibility in supporting a broad range of services, characterized by divergent traffic behaviors. In ATM networks, fixed length cells belonging to different calls are multiplexed together for transport. An ATM network service provider can optimize the efficiency by defining criteria for resource allocation and management to meet the quality-of-service requirements and to ensure optimal resource utilization.

During the call setup phase, the ATM network must decide whether to admit the new call or not.

This depends primarily on the availability of bandwidth to support the connection along the end-to-end path. The amount of bandwidth allocated for the new call is based on various statistical properties of the call, and on the number of calls already in session. The network uses three routing considerations to assign bandwidth to new calls: network routing of the Virtual Channel Identifier (VCI); path selection inside the switching fabric performed on a per call basis; and semi-permanent virtual connection and virtual path routing performed on a long term basis under the purview of network management. Once the network accepts a connection, it monitors the cell stream to ensure that the user does not exceed the values established in the call setup phase. If the user exceeds these values, the local ATM node's Policing Unit takes corrective policing action. This enforcement mechanism is applied at the point of traffic origination, prior to multiplexing.

Given the very bursty nature of ATM traffic, a methodology is required to efficiently allocate bandwidth on a broadband link according to the simple criterion that a call request be granted if the resource is available. The objective here is to maximize the number of connections on the link while maintaining a particular quality of service specified in terms of the cell loss probability. The scheme should be relatively easy to implement in real time. Call acceptance will be based on the estimated distribution of the calls arriving at the switching node and their associated bounds on cell loss probability.

The choice of traffic control algorithm directly impacts on a network's resource allocation strategy. For example, if only the peak bit rate of a connection were considered for admission, then this peak bit rate would have to be allocated for the connection. If this connection had a low average bit rate, then most of the time the network would exhibit a poor efficiency. The goal is to simultaneously: maximize the ATM network efficiency; and meet the users' Quality of Service (QOS) requirements.

2. Related work

Most of the previous work in the area of resource allocation has been limited to studying a single source. Extending this to multiple sources spanning virtual circuits over several nodes has not provided an implementable model. Much of the work in modeling packet switching systems has been extended to ATM connections as well [6,19,20]. Server allocation policies have been extensively studied in the queuing theory literature, but only a qualitative structure of the optimal policy has resulted. No bounds are available for the time required to perform a fair allocation of the bandwidth in a network of a given size, and very few authors have addressed this problem. An early distributed broadband network proposed by Prycker and Somer [25] was capable of supporting transport services of any rate based on the fast packet switching concept. A static load control mechanism was used and calls were admitted according to the level of the load. They did not explore the dynamic behavior of the network, and the ''burstiness'' of the offered traffic was not effectively modeled.

Aicardi et al. [1] proposed a two level control structure for the dynamic allocation of bandwidth in a multiservice network. Their approach was to provide a simple time invariant, randomized strategy to decide traffic scheduling on the basis of the local state information. No specific model was used, and the approach suggested that STM be used to describe a finite state Markov chain. Pattavina [24] also proposed a two step bandwidth allocation scheme. His scheme was based on the definition of the data link layer connections, whose capacity is not bounded by the capacity of a single broadband packet channel and was shown to be feasible in Batcher–Banyan switch. His scheme was well suited for an internal path allocation in a packet switch.

Gersht and Lee [11] proposed a bandwidth management framework for fast packet switched broadband networks. They proposed a threshold based strategy and derived an approximate solution technique to obtain the call blocking probabilities using first moment matching and an integer optimization algorithm using the reward function method to find optimal threshold parameters. Hui [16] provided a methodology for a multilayer bandwidth allocation scheme using the congestion measures of packet blocking, burst blocking, and call blocking. His work provided an insight into traffic engineering issues such as the appropriate link load, traffic integration, and bandwidth reservation criteria for bursty services. Jordan and Varaiya [18] posed a resource allocation problem for a multiple services and multiple resources model of a communications system that can process general types of requests, each of which requires several types of resources.

A traffic measurement method and its application for cell loss probability estimation was provided by Yamada and Sumita [33]. They presented a method that could be used for traffic measurement problems related to ATM networks and, in particular, to cell traffic measurements. They provided a method for the on-line evaluation of the first and the approximate second moment statistics during any interval. They estimated the cell loss probability through the use of a queuing model with an input process determined from the traffic flow description. Saito and Shiomoto [28] have another procedure to estimate the distribution of the number of arriving cells. Guerin, Ahmadi, and Nagshineh [13] developed a unified metric to represent the effective bandwidth load on the network links. They proposed approximate expressions for the equivalent capacity or bandwidth requirement of both individual and multiplexed connections. They also proposed an exact approach to the computation of the equivalent capacity, but the associated computation complexity made it infeasible for real time network traffic control applications. Saito, Kawarasaki, and Yamada [29] analyzed traffic characteristics of an ATM network and showed that by lowering the peak bit rate of a video source, the bandwidth utilization could be effectively improved. They concluded that statistical bandwidth allocation is ineffective without QOS control.

Connection admission strategies in ATM networks without any Poisson or renewal assumption were used by Rasmussen and Jacobson [27] to derive bounds on congestion for a simple admission control algorithm. Sriram and Whitt [31] proposed the use of the Index of Dispersion for Intervals (IDI) to characterize the superposition of the voice and data arrival processes, and in a later contribution Fendwick and Whitt [8] proposed the use of a scaled version of the variance-time curve as a measurement to describe the variability of the traffic offered to a queue. Heffes and Lucatoni [14] use a Markov Modulated Poisson Process (MMPP) to model the dependencies present in voice and data sources. They obtain the moments of the voice and data delay distributions as well as the queue length distributions. Ohba, Murata, and Miyahara [23] analyze the interdeparture processes for bursty traffic in the ATM environment. They model the ATM switch as a discrete time single server queue at which three kinds of arrival processes are allowed to join together. They offer an exact analysis to derive the waiting time distributions and interdeparture time distributions for arriving cells subject to admission control.

3. Gaussian approximation

The bandwidth needed to support the traffic generated by an on-demand connection must be available throughout the lifetime of the connection. Allocating the bandwidth for a call at its peak rate expends the bandwidth resource acutely, while reservation at its average rate cannot guarantee the QOS (Quality of Service) during periods when the source is transmitting at its peak rate.

In this model (Fig. 1) arriving cells from each call are buffered in a 2 cell FIFO. This enables the ATM switch to statistically multiplex the traffic, and the cells are output at the data rate of the link. Since the multiplexer cannot transmit a cell instantaneously, a 2 cell FIFO is required to allow the multiplexer the time to transmit one cell while a new cell is arriving. Larger FIFOs can be used, although this would

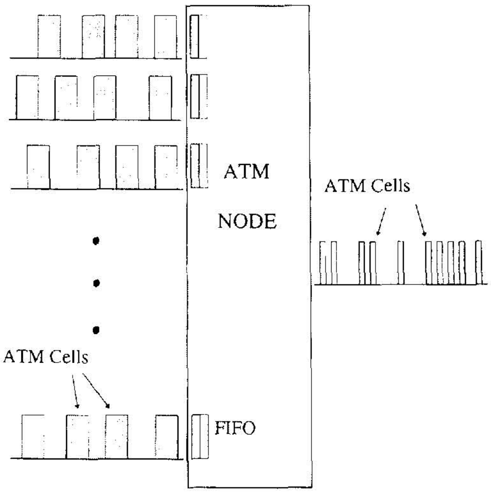

Fig. 1. ATM node model.

increase the average queuing delay and increase the cost of the hardware implementation. The link bandwidth utilization would also increase, however the benefit is unclear since this requires temporal information of the distribution of cells on the link. We have assumed a 2 cell FIFO which is the minimum requirement since a cell cannot be transmitted until it is completely received. The multiplexing switch therefore has to transmit this cell before the next one arrives completely so as to not "overwrite it" (or get dropped) and thus resulting in a cell loss.

A two cell FIFO (at each node or switch) will still yield a small (and bounded) *random* delay; however this delay is small and minimal. In essence, if the source transmits cells at a constant rate, the sink will receive cells at the same average rate, but with variable delay. This variable delay is bounded if the number of hops is bounded. It is recommended that a small delay buffer be used at the sink which is equal to the maximum network delay so that the sink can ensure a "smooth" constant cell flow as delivered by the source.

A FIFO larger than 2 cells does not change this. However it allows for a larger peak delay from the source to the sink. This raises the question: "How large is too large?" This of course depends on the nature of the application. The larger buffer will reduce the "effective" variance of the individual sources seen by the network. The variance is smaller when larger buffers are used but the analysis remains the same. For example, a 100 second call with a 25 Mbps burst for 4 consecutive seconds has the same mean and variance as a 100 second call made up of 100 40 ms bursts of 25 Mbps separated by 1 second. However a 1 MB buffer will not reduce the effective variance of the first situation much, but will essentially eliminate the variance of the second scenario. This is because buffering is a type of low pass filtering. If the call can be characterized a priori on an average time basis, we can then compute the effective variance throughout the duration of the call.

The state of the transmission link is completely described by the net mean and variance which increases linearly with the number of calls. It can then be reasonably assumed that the traffic on the link tends towards a Gaussian distribution as the number of users increases [5,13]. Fig. 2 illustrates the p.d.f. of the link traffic and shows the effect of the burstiness on the distribution of the statistically multiplexed traffic on the link. The distribution tends to shift and widen as traffic on the link grows. When the demand on the link exceeds the capacity of the link, cells are lost and we can compute the Cell Loss Probability (CLP) analytically by using the Gaussian approximation.

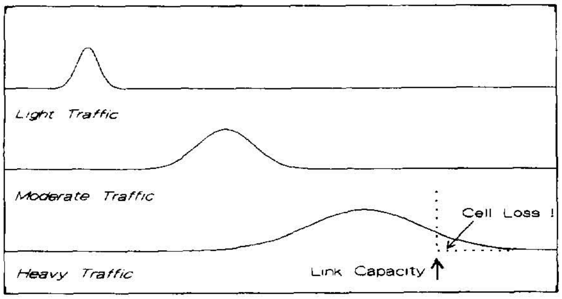

Fig. 2. Gaussian approximation.

Using the central limit theorem to estimate the distribution of the superimposed traffic, we apply the Gaussian approximation to compute the cell loss probability under heavy load conditions [13,17,23]. Whenever the effect of statistical multiplexing is of significance, the distribution of the stationary bit rate can be fairly accurately approximated by a Gaussian distribution. The assumption allows us to use the standard approximations to estimate the tail of the bit rate distribution. If the bit rate is not approximately Gaussian distributed, the difference between the assumed distribution and the Gaussian envelope may become so large that it is better to use an upper bound. The variance of a connection with bit rate x_t, mean bit rate μ_i, and peak bit rate $R^{(i)}_{peak}$ satisfies the following inequality:

$$\begin{aligned}\mathrm{Var}(x_t) &= E(x_t)^2 - (E(x_t))^2\\ &\leqslant E\left(R^{(i)}_{peak}\cdot x_t\right) - (E(x_t))^2\\ &= \mu_i\left(R^{(i)}_{peak} - \mu_i\right).\end{aligned} \tag{1}$$

The upper bound $\mu_i(R^{(i)}_{peak} - \mu_i)$ coincides with the bit variance if and only if the connection has on/off characteristics, i.e., has only two states: sending at peak bit rate or not sending at all. Using this upper bound, only two parameters – the mean bit rate μ_i and the peak bit rate $R^{(i)}_{peak}$ – are needed to characterize a connection.

4. Cell loss probability

It will now be shown that the instantaneous cell loss probability of a call i using a link with N active connections can be computed from the Complementary Error Function. The region of cell loss is as depicted in Fig. 2. Within the bounds of the Gaussian approximation, when no buffer provision is made to absorb burst scale congestion, cell loss probability depends on the instantaneous demand on the link bandwidth. If the instantaneous bandwidth demand (x_d) exceeds the maximum link capacity, the excess data $(x_d - B)$ is lost. The cell loss probability is then $\{E(x_d - B)/\mu\}$.

Mathematically, the cell loss rate is the expected value of the instantaneous demand in excess of the link capacity.

$$\text{Cell Loss Rate} = \int_B^\infty P(x_d)(x_d - B)\,\mathrm{d}x_d. \qquad (2)$$

By the assumption, $P(x_d)$ is the Gaussian distribution given by

$$P(x_d) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x_d-\mu)^2/2\sigma^2}. \qquad (3)$$

Substituting Eq. (3) into Eq. (2) yields

$$\text{Cell Loss Rate} = \frac{\sigma}{\sqrt{2\pi}} e^{-((B-\mu)/\sigma)^2/2} + (\mu - B)Q\left(\frac{B-\mu}{\sigma}\right), \qquad (4)$$

where $Q(x)$ is the complementary error function or the co-error function. It can also be expressed in the form

$$Q(x) = \frac{1}{2} erfc(x/\sqrt{2}). \qquad (5)$$

The cell loss probability (CLP) therefore is

$$\text{CLP} = \text{Cell Loss Rate}/\mu. \qquad (6)$$

Substituting Eq. (4) into Eq. (6), we have the CLP as

$$\text{CLP} = \frac{1}{\mu}\left[\frac{\sigma}{\sqrt{2\pi}} e^{-((B-\mu)/\sigma)^2/2} + (\mu - B)Q\left(\frac{B-\mu}{\sigma}\right)\right]. \qquad (7)$$

The CLP as computed in Eq. (7) is shown for each of three classes of calls in Figs. 3(a), 3(b) and 3(c) for link capacities of 155 Mbps, 1 Gbps, and 2 Gbps respectively. The Class A connection has a 10 Mbps peak rate at 10% duty cycle, the Class B has a 4 Mbps peak requirement at 25% duty cycle, while the Class C transmits at 25 Mbps at 4% duty cycle. The choice of the class characteristics reflects the different burstiness levels. Class C is the most bursty and Class B the least. Some care must be exercised to avoid situations where the Gaussian assumption used in the stationary approximation does not hold. This typically happens with small numbers of very bursty calls with high peak rates, low duty cycle and long bursts.

5. Statistical gain

In a circuit switched network, each user is allocated the required peak bandwidth for the duration of the call. This results in inefficient allocation of bandwidth resources. When a particular user is idle, the allocated bandwidth is nevertheless reserved throughout the duration of the call and no other user may use that bandwidth. Consider a number of users each demanding a peak bandwidth of R_{peak} and having an average bandwidth requirement of μ. The average bandwidth efficiency, $\eta_{BW,\,CircuitSwitched}$, of a circuit switched network is given by

$$\eta_{BW,CircuitSwitched} = \frac{\mu(B/R_{peak})}{B} = \frac{\mu}{R_{peak}}, \qquad (8)$$

where B is the total link capacity. Let the bandwidth efficiency of a statistically multiplexed link be denoted by $\eta_{BW,\,StatisticallyMux}$. The statistical gain of a statistically multiplexed link is a measure of the increased bandwidth efficiency with respect to a circuit switched network and is given by:

$$\text{Statistical Gain} = \frac{\eta_{BW,StatisticallyMux}}{\eta_{BW,CircuitSwitched}}. \qquad (9)$$

Consider the three classes of calls discussed earlier. The statistical gain for each of the three classes is shown in Figs. 4(a), 4(b), and 4(c). The statistical gain, whose upper bound is the reciprocal of the duty cycle for each class, is a strong function of the CLP. For example, in Fig. 4(b) with a CLP of, say, 10^{-11}, we achieve an approximate statistical gain of 10 for

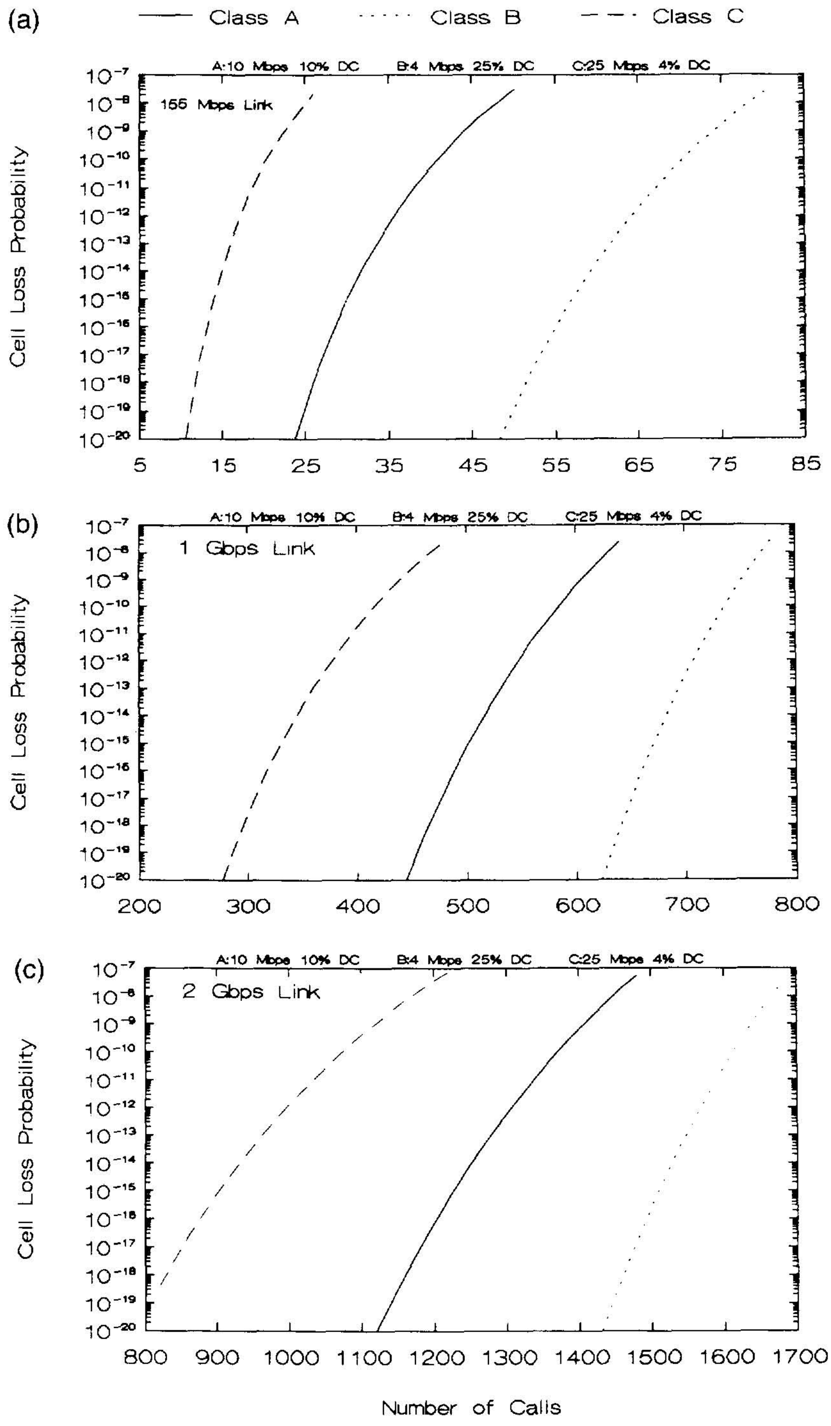

Fig. 3. (a) CLP for 155 Mbps link. (b) CLP for 1 Gbps link. (c) CLP for 2 Gbps link.

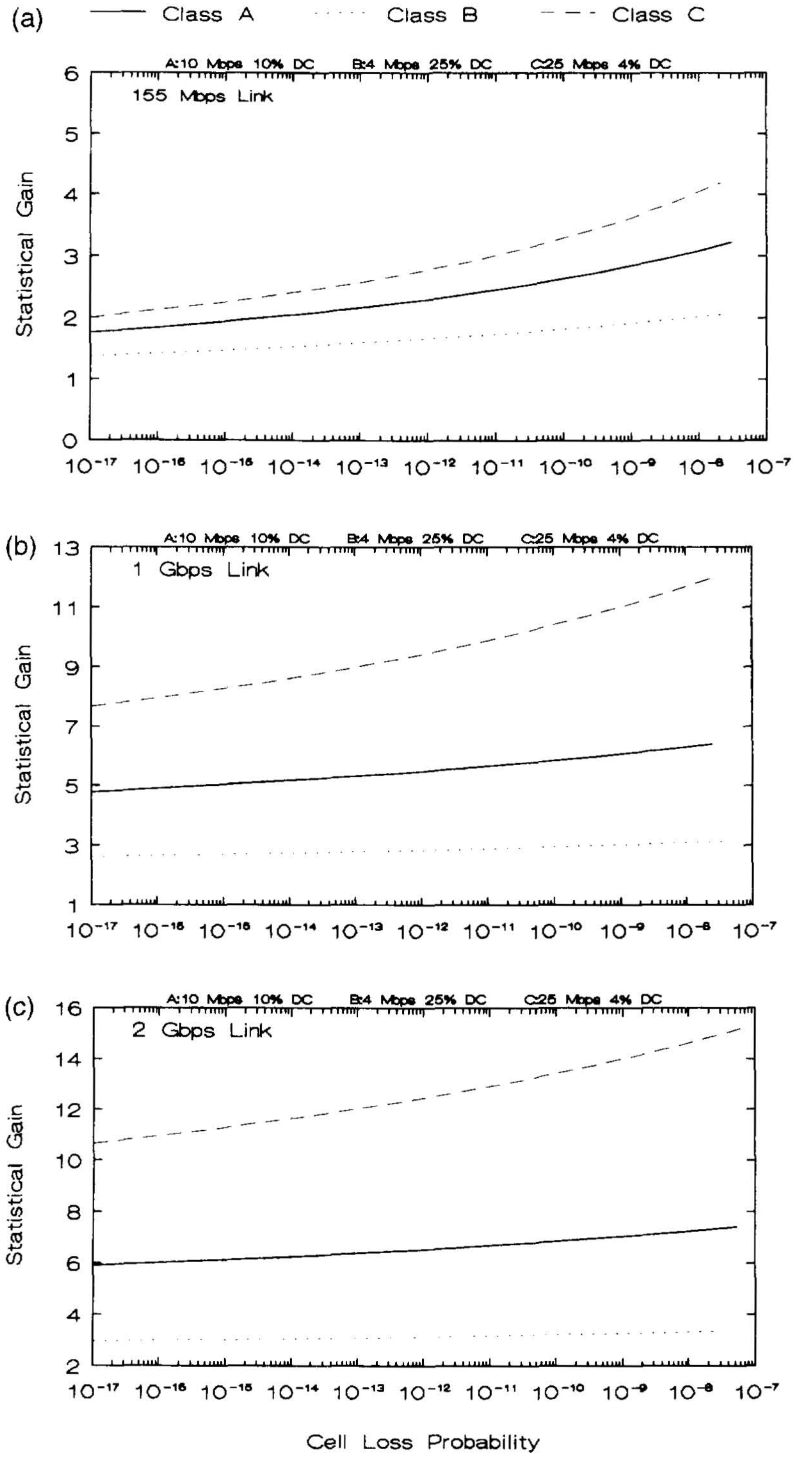

Fig. 4. (a) Statistical gain for 155 Mbps link. (b) Statistical gain for 1 Gbps link. (c) Statistical gain for 2 Gbps link.

a Class C connection. This implies that for a Class C connection, we can statistically multiplex 400 connections on a 1 Gbps link at a cell loss probability of 10^{-11}. A circuit switched network however would be capable of supporting only 40 calls at its peak rate. Fig. 4 indicates that a higher statistical gain can be achieved when burstiness increases.

6. Call admission algorithm

By modeling the call stream as a deterministic fluid we achieve simplicity and analytical tractability of the framework. A computationally simple approximation is used for the equivalent bandwidth requirement of a single or multiplexed connections on the basis of their statistical characteristics. Our intention is to employ the approximation approach applied by Guerin [5] in the development of plaNET [2] to arrive at an analytically simple model for a call admission algorithm as well as the concurrent desire for versatility. The cell descriptor used by Guerin is proposed for this model as it best characterizes the key parameters of the call.

When connections are statistically multiplexed, their aggregate statistical behavior differs from their individual representation. The two state fluid-flow model adopted here is based on the rationale that a source is either in an "idle state" transmitting at zero bit rate, or in a "burst state" and transmitting at its peak rate [13]. The arrival process for the calls and the distribution of their duration can be characterized by a Poisson process and by an exponential distribution, respectively. In the two-state fluid flow model, the correlated generation of cells within a call can be modeled as an Interrupted Poisson Process (IPP) [1]. In the IPP model, the transition from"burst" state to "idle" state and vice versa occurs with probabilities P_a and P_b, respectively. In the discrete time case, the burst and idle periods are geometrically distributed with means $1/P_a$ and $1/P_b$, respectively. The discrete time analog is a Bernoulli distribution with cells generated with rate λ in the burst period.

The aggregated cell arrivals when N connections are multiplexed depend on the number of sources in the burst state. The probability P_n that n out of N connections are in the burst state (for a discrete time system) is given by

$$P_n = \binom{N}{n}\left(\frac{P_a}{P_a+P_b}\right)^n\left(\frac{P_b}{P_a+P_b}\right)^{N-n}$$

$$\text{for } 0 \leqslant n \leqslant N. \qquad [10]$$

If the call arrival process is modeled as a renewal process, the consequence of multiplexing many independent call streams is not nearly a renewal process. The statistical behaviour of bursty traffic, after the mixing of many sources, is far from Poisson and makes the analysis very difficult. The choice therefore is to model this as a Markov Modulated Poisson Process (MMPP) [6]. The MMPP is a doubly stochastic Poisson process where the rate process is determined by the state of a continuous-time Markov chain. The IPP which is used to describe a single source is a special case of the two state MMPP.

The network provides services for κ different classes of calls each of which requires different amounts of the bandwidth resource for connection. During the call setup, each intermediate node on the route independently decides whether to accept the incoming call or not, based on the requested amount of bandwidth B_i. If the requested rate cannot be allocated due to lack of bandwidth, the call has to be blocked. Also, the call will be rejected if it is not setup in a predetermined amount of time.

An algorithm is needed to decide whether a new call is to be admitted to the system, or is queued for later consideration. The approach considered here exploits the mean and variance statistics associated with each call to maximize the link utilization. The process is as follows: We have a set of calls contending for admission into the system. When the link is lightly loaded, we admit any call which does not violate the Isoprobability threshold (explained later in this section). When the link is heavily loaded, we would like to admit as many calls as possible, as long as the QOS can be maintained. The overall link efficiency is enhanced, if calls which are *burst reducing* are admitted. These are calls with low variances and high means.

Since QOS is defined in terms of the cell loss probability, the algorithm uses the current mean and variance of the link along with the mean and variance of the new call to determine if admitting the

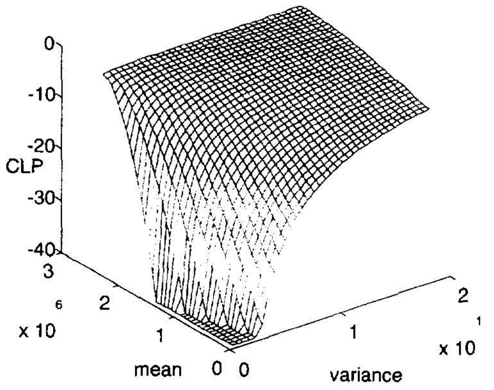

Fig. 5. Cell loss probability.

new call would violate the QOS constraint of the link. This is purely a function of the net mean and the net variance that describes the state of the link. Fig. 5 shows the surface plot of Eq. (7). The Z axis represents the log of CLP as a function of the net mean and the net variance on the link. We require that our CLP never exceed some threshold, say, 10^{-11}. Mathematically, if B_u and V_u represent the link bandwidth and link variance in use respectively, then the current state of the link can be defined as

$$X=(B_u, V_u). \tag{11}$$

If the ith call has a mean bandwidth requirement of μ_i and variance σ_i^2, then by our admission criteria, CLP which is now a function of $(B_u+\mu_i, V_u+\sigma_i^2)$ must satisfy

$$\text{CLP} = f(B_u + u_i, V_u + \sigma_i^2) \leqslant 10^{-11}. \tag{12}$$

For any state of the link $X=(B_u, V_u)$ there exists a specific CLP, and we can extract the constant loss probability curve corresponding to that CLP. The resulting Isoprobability curve is shown in Fig. 6. Fig. 6 defines the desired operating region and the call reject zone for all calls that would take the state of the link above the threshold set by the Isoprobability curve. Implementing the QOS constraint means that we accept no call that would drive the CLP of the link beyond this specified threshold. The operating region is where the aggregate traffic characteristics within the network are "well behaved". In this region, it is possible to achieve reasonable bandwidth utilization across a wide range of traffic types and mixes. Traffic which is generally not suitable for statistical operation, such as constant bit rate traffic (CBR), can be smoothly integrated into this admission scheme.

The first step of the proposed call admission procedure is to ensure that any call to be considered for admission meets the QOS constraint. If no call in the active set meets the QOS constraint, no call will be admitted to the network for this epoch. If the total demand on the link is small, then most, if not all the calls will meet the QOS constraint and be admitted. If the total demand on the link is large, then admitting any and all calls which meet the QOS constraint will tend to drive the operating point of the link towards the Isoprobability curve. However, when the demand is large (i.e., total resources requested by all incoming calls exceeds the statistically multiplexed bandwidth available for the means and variances of these calls) there will typically be many calls in the active set. The call admission criteria can become more selective as the size of the active set grows thereby enabling the call admission algorithm to achieve greater overall link bandwidth efficiency. The approach is to assign a reward metric for each call in the active set. The call that meets the QOS constraint and has the maximum reward metric is admitted, and the state of the link is updated.

The obvious question is: "How do we compute the reward metric?". The objective of the call admission scheme is to admit those calls which maximize the link efficiency. This is tantamount to admitting those calls which maximize the bandwidth in use, B_u, while maintaining the QOS requirement. The

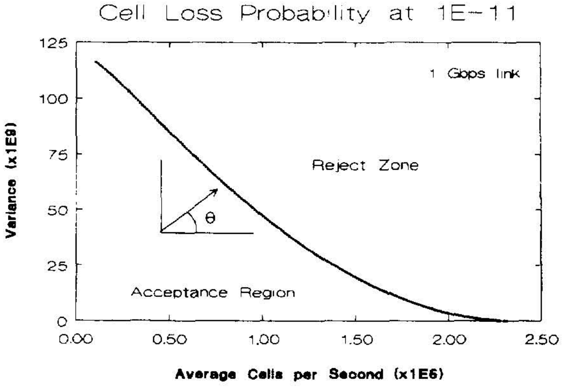

Fig. 6. Isoprobability (constant CLP) curve.

approach taken here is to let the reward metric be a linear combination of the mean and variance for each call. The linear combination essentially measures the correlation between each call and a vector χ on the mean–variance plane. Since all reward metrics will be compared to each other, we can consider χ to be a unit vector. The reward metric is therefore defined as

$$\Re = \mu_i \cos\theta + \sigma_i^2 \sin\theta, \qquad (13)$$

where θ is the angle of the vector χ still required to uniquely identify this vector. Simulation was used to determine the optimum angle.

7. Simulation and numerical results

A time based simulation was performed to determine the optimum correlation angle, and to evaluate the performance of the link under certain statistics. The base simulation used the following parameters:

(1) B = Link Capacity = 1 Gbps

(2) CLP = Cell Loss Probability = 10^{-11}.

(3) Duration of calls in the active set before being dropped = waiting time = 20 seconds.

(4) Call arrivals occur according to a Poisson process with an arrival rate of 1 call per second.

(5) Active set queue size = infinite. Due to the call arrival rate and the waiting time, this value tended towards 20, and seldom exceeded 40.

(6) Call mean bandwidth requirement was an exponential random variable with an average value of 1 Mbps.

(7) Call variance requirement was an exponential random variable with an average value of 2.5×10^{13}. (The mean bandwidth requirement and variance requirement corresponds to traffic at 25 Mbps with 4% duty cycle; this was defined earlier as the Class C type of call.)

(8) Call duration was an exponential random variable with an average value of 800 seconds.

(9) Offered traffic ≈ twice simple capacity. The call duration along with the arrival rate correspond to an equilibrium offered traffic load of 800 calls. As seen in Fig. 4, this link can only support approximately 400 Class C calls. The 400 calls are referred to as the simple capacity since there is no call admittance scheme; or the call admission scheme is to simply allow any and all calls on the link which meet the QOS constraint on a first come first served basis.

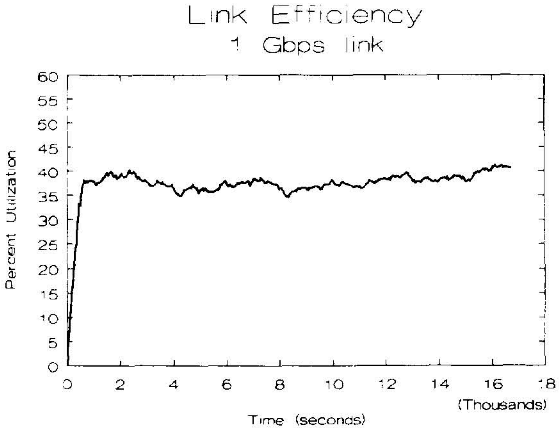

Fig. 7. Link utilization of calls meeting QoS.

To verify the model, we used the simple admittance algorithm of allowing any and all calls on the link which met the QOS constraint. The bandwidth in use B_u was sampled every 5 seconds of simulated time. A typical plot of B_u as a function of time is shown in Fig. 7. Notice that the bandwidth in use reached an equilibrium value near 38%. This corresponds to the value in Fig. 7 for a CLP of 10^{-11}. To determine the optimum correlation angle, 36 simulation runs were performed varying the correlation by 10° each time. Each run represented approximately 5 hours of operation time and the link bandwidth in use B_u was averaged over the 5 hours to get a measure of the average link utilization. It should be mentioned that the first 40 minutes were ignored since we were interested only in the equilibrium behavior of the link. The average link utilization (bandwidth in use) as a function of correlation angle for the given parameters is shown in Fig. 8 for two values of the waiting time $W_t = 20$ and $W_t = 50$ seconds. The set corresponding to the 50 second waiting time has the greater variation owing to the increased number of calls in the active set (on average). Both curves reveal a relatively flat optimum between $-20°$ (340°) and 10°. From this we can conclude that a correlation angle of 0° is near optimum. A zero degree correlation angle is of particular significance since any implementation would benefit greatly from the simplicity of a zero degree correla-

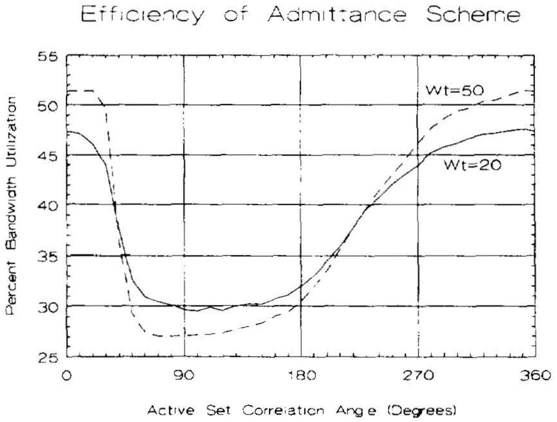

Fig. 8. Link utilization vs. correlation angle.

tion angle. A zero degree correlation angle corresponds to a reward metric which is simply equivalent to the value of the average bandwidth required by each call, and the variance parameter is ignored.

It should be clearly stated that these results are based on the statistics of the arriving calls as given above. The simulation results did not provide any empirical data suggesting any particular distribution on the mean and variance requirements of any call, or any relationship between a call's mean and variance. For this reason we used the somewhat simplistic approach of modeling the arriving calls as having exponential mean and exponential variance within a Poisson process.

8. Implementation complexity

The proposed algorithm is relatively simple in its approach, and very tractable for hardware implementation. The algorithm requires some measure or estimate of the links statistics (current state). One method might be to simply add the means of all the calls and the variances of all the calls to arrive at an estimate of the link. This method suffers from the estimation error of each individual call. A better approach is to measure this value directly, if possible. Since the mean and variance are expected values, they can only be estimated in any real system. Therefore we require a method to estimate these parameters. Instantaneous bandwidth demand can be measured by adding the state of all buffers at a given instant. If a buffer is full, its state is 1, if a buffer is empty, its state is 0. The addition of these state values can easily be accomplished in analog circuitry. An average value is accomplished by some averaging, or low pass filtering. The result is a measure or estimate of the bandwidth in use on the link $\hat{B}_u$. Similarly, we can also arrive at an estimate of the variance in use on the link $\hat{V}_u$. Thus, we can directly estimate the state of the link. The QOS constraint is obtained by evaluating

$$\mathrm{CLP} = f\left(\hat{B}_u + \mu_i, \hat{V}_u + \sigma_i^2\right), \tag{14}$$

where the function can be computed directly, or evaluated via a look up table. The reward metric is simply the bandwidth required μ_i. For all calls in the active set which meet the QOS constraint ($\mathrm{CLP} \leqslant 10^{-11}$), a search is performed to find the one with the maximum average bandwidth requirement μ_i. This is the call which is admitted.

9. Discussion

The ATM switch model uses a 2 cell FIFO which is a minimum requirement. Larger FIFOs can be used, although this would increase the average queuing delay and the cost and complexity of any hardware implementation. The effect that a multi cell FIFO has on bandwidth utilization is uncertain since this requires temporal information about the link traffic. Specifically, the autocorrelation function of the link demand is needed without which it is not possible to determine the tradeoff between the buffer size (and hence delay) and the effective increase in the average bandwidth utilization. With the 2 cell FIFO, this tradeoff is not considered. Bandwidth utilization is increased solely by shaping traffic by selectively admitting calls that have favorable statistics.

Call acceptance is based on the on-line evaluation of the upper bound on the Cell Loss Probability. The quality of service can be assured using this control when there is no estimation error. The control mechanism is effective when the number of calls is large. It tolerates loose bandwidth enforcement and loose policing control. Under high demand conditions, the call admission scheme increases the overall link efficiency by selecting only those calls that are well behaved. Simulation results of this algorithm demonstrate the effectiveness of this control.

The loss performance when modeled as a MMPP may be large. However, when considered with respect to the number of users on the link, the error is small, typically around 5%. Nonetheless, it should be stressed that our intent is to show that this admittance scheme can substantially increase the overall link efficiency as demonstrated by our results. The fidelity with which the traffic parameters accurately reflect true ATM traffic is unclear. The virtual circuit acceptance routine is straightforward and supports fully heterogeneous traffic. The control architecture schematic shown in Fig. 9 can be effectively implemented in hardware. A schematic of an ATM switching node with the CAC unit is also shown. Finally, we believe that our approach can achieve significant link efficiencies even under very bursty traffic.

A number of possible extensions and topics for future research can be identified. As indicated earlier, the size of the active set is a strong factor influencing our approach. Further studies could be conducted to find an optimal length of this queue and perhaps define an effective polling procedure that would be required as the queue becomes larger. Additionally, performance guarantees for multiple transport performance classes can be supported, and traffic with very stringent performance requirements could be given nonstatistical treatment. This premise warrants a detailed evaluation. The results of this study serve as a starting point in further characterization of an optimal policy, in understanding and designing effective heuristic rules, and in developing

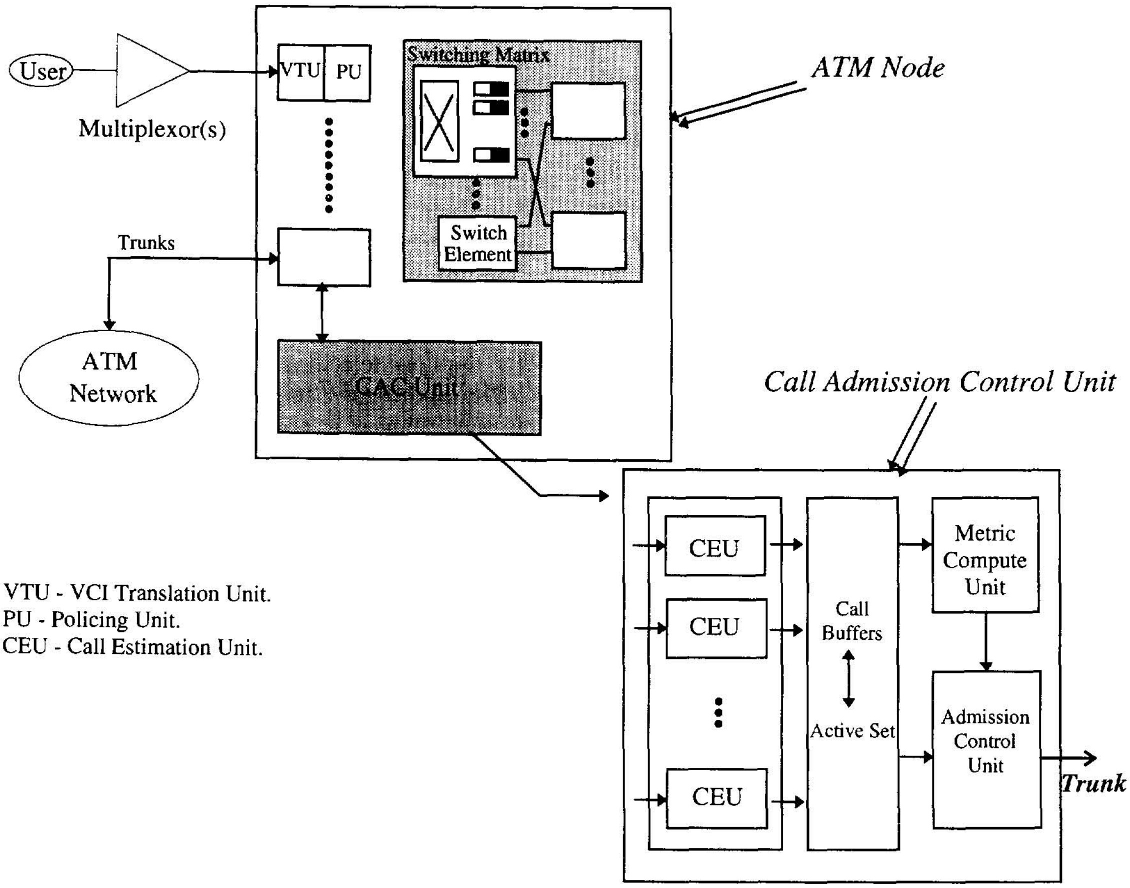

Fig. 9. ATM local node reference architecture with CAC unit.

normative-descriptive ATM Connection Admission Control models.

10. Conclusions

We have presented an analytically simple yet versatile call admission scheme that dynamically shapes the traffic on the link and uses the burstiness associated with traffic sources to effect dynamic assignment of bandwidth. The negotiated quality of service is guaranteed using this control scheme, provided that there is no estimation error. The control mechanism is effective when the number of calls is large, and tolerates loose bandwidth enforcement and loose policing control. The proposed approach is very effective in the connection oriented transport of ATM networks where the decision to admit new traffic is based on the a priori knowledge of the state of the route taken by the traffic.

References

[1] M. Aicardi, F. Davoli and R. Minciardi, A model for combined user premises local exchange dynamic bandwidth assignment in broad-band communication networks, in: *Conf. Record IEEE Globecom'89*, 1989.

[2] S.L. Albin, Approximating queues with superposition arrival processes, Doctoral Dissertation, Columbia University, 1981.

[3] J.J. Bae and T. Suda, Survey of traffic control schemes and protocols in ATM networks, *Proc. IEEE* 79 (2) (1991).

[4] J. Burgin, Management of capacity and control in Broadband ISDN, *Internat. J. Digital and Analog Cabled Systems* 2 (1989) 155–165.

[5] I. Cidon, I. Gopal and R. Guerin, Bandwidth management and congestion control in plaNET, *IEEE Commun. Mag.* (October 1991).

[6] J.N. Daigle and J.D. Langford, Models for analysis of packet voice communications systems, *IEEE JSAC* 4 (6) (1986).

[7] A. Descloux, Stochastic models for ATM switching networks, *IEEE JSAC* 9 (3) (1991).

[8] K. Fendwick and W. Whitt, Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue, *Proc. IEEE* 77 (1) (1989) 171–194.

[9] G. Gallassi, G. Rigolio and L. Fratta, ATM, bandwidth assignment and bandwidth enforcement policies, in: *Conf. Record IEEE Globecom'89* (1989) 1788–1793.

[10] G. Gallassi, G. Rigolio and L. Verri, Resource management and dimensioning in ATM networks, *IEEE Networks Mag.* (May 1990).

[11] A. Gersht and K.J. Lee, Virtual-circuit-load control in fast packet switched broadband networks, in: *Conf. Record IEEE Globecom'88* (1988) 214–219.

[12] H. Gilbert, O. Aboul-Magd and V. Phung, Developing a cohesive traffic management strategy for ATM networks, *IEEE Commun. Mag.* (October 1991), p. 36.

[13] R. Guerin, H. Ahmadi and M. Naghshineh, Equivalent capacity and its application to bandwidth allocation in high-speed networks, *IEEE JSAC* 9 (7) (1991).

[14] H. Heffes and D.M. Lucantoni, A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance, *IEEE JSAC* 4 (6) (1986).

[15] B.E. Helvik, P. Hokstad and N. Stol, Correlation in ATM traffic streams – some results, in: *Proc. 13th Internat. Teletraffic Congress (ITC-13) Workshop*, Copenhagen, Denmark, 1991.

[16] J.Y. Hui, Resource allocation for broadband networks, *IEEE JSAC* 6 (9) (1988).

[17] J.Y. Hui, *Switching and Traffic Theory for Integrated Broadband Networks* (Kluwer Academic Publishers, 1990).

[18] S. Jordan and P.P. Varaiya, Throughput in multiple service, multiple resource communication networks, *IEEE Trans. Commun.* 39 (8) (1991).

[19] B. Kraimeche and M. Schwartz, Analysis of traffic access control strategies in integrated service networks, *IEEE Trans. Commun.* 33 (10) (1985).

[20] B. Kraimeche and M. Schwartz, Bandwidth allocation strategies in wide-band integrated networks, *IEEE JSAC* 4 (6) (1986).

[21] S. Low and P.P. Varaiya, A simple theory of traffic and resource allocation in ATM, in: *Conf. Proc. IEEE Globecom*, 1991.

[22] J.A.S. Monteiro, M. Gerla and L. Fratta, Input rate control for ATM network, in: *Proc. 13th Internat. Teletraffic Congress (ITC-13) Workshop*, Copenhagen, Denmark, 1991.

[23] Y. Ohba, M. Murata and H. Miyahara, Analysis of interdeparture processes for bursty traffic in ATM networks, *IEEE JSAC* 9 (3) (1991).

[24] A. Pattavina, Multichannel bandwidth allocation in a broadband packet switch, *IEEE JSAC* 6 (9) (1988).

[25] M.De. Prycker and M.De. Somer, Performance of a service independent switching network with distributed control, *IEEE JSAC* 5 (8) (1987).

[26] V. Ramaswami and G. Latouche, Modeling packet arrivals from asynchronous input lines, *12th Int. Teletraffic Congress*, Turin, June 1988.

[27] C. Rasmussen et al., Source independent call acceptance procedures in ATM networks, *IEEE JSAC* 9 (3) (1991).

[28] H. Saito and K. Shiomoto, Dynamic call admission control in ATM networks, *IEEE JSAC* 9 (7) (1991).

[29] H. Saito, M. Kawarasaki and H. Yamada, An analysis of statistical multiplexing in an ATM transport network, *IEEE JSAC* 9 (3) (1991).

[30] S.M. Srinidhi and V.K. Konangi, A resource allocation model for BISDN/ATM, in: *Proc. Internat. Phoenix Conf. on Computers and Communication*, Tempe, AZ (1993) 282–288.

[31] K. Sriram and W. Whitt, Characterizing superposition arrival

processes in packet multiplexers for voice and data, *IEEE JSAC* 4 (7) (1986).

[32] E. Wallmeier and C.M. Hauber, Blocking probabilities in ATM pipes controlled by a connection acceptance algorithm based on mean and peak bit rates, in: *Proc. 13th Internat. Teletraffic Congress (ITC-13) Workshop*, Copenhagen, Denmark, 1991.

[33] H. Yamada and S. Sumita, A traffic measurement method and its application for cell loss probability estimation in ATM networks, *IEEE JSAC* 9 (3) (1991).

[34] U. Yechiali, Optimal dynamic control of polling systems, in: *Proc. 13th Internat. Teletraffic Congress (ITC-13) Workshop*, Copenhagen, Denmark, 1991.