A Measurement-Based CAC Strategy for ATM Networks

Kunyan Liu, David W. Petr, Cameron Braun,*

Telecommunications & Information Sciences Laboratory
Department of Electrical Engineering and Computer Science
The University of Kansas
Lawrence, KS 66045-2228

*Sprint Corporation, Overland Park, KS 66212

Abstract

In ATM networks, Connection Admission Control (CAC) has been recognized as one of the most important means to provide satisfactory quality of service (QoS) and protect the network from congestion, especially for supporting real-time services such as voice and video. An efficient CAC strategy, which can both guarantee the QoS of admitted connections and achieve good resource utilization, then becomes a crucial issue for ATM network providers. The concept of a user-network traffic contract has been introduced by ATM Forum. Starting from this point, we propose a measurement-based CAC strategy. We first discuss how to obtain an accurate description (traffic parameters) of user traffic by using on-line measurement in conjunction with dynamic renegotiation. The results show that the proposed strategy is reliable and simple to implement. We then move on to examine the possibility and methodology for exploiting the effect of statistical multiplexing in resource allocation to achieve higher network resource utilization.

1: Introduction

A significant portion of traffic in future ATM-based B-ISDN will consist of real-time services, such as voice, multimedia and especially video applications. Currently, the service classes defined by ATM Forum for these applications are Constant Bit Rate (CBR) and Variable Bit Rate (VBR). However, since most traffic sources are intrinsically variable bit-rate, it is generally accepted that using the VBR class will result in better information quality. Meanwhile, from the network provider's point of view, supporting VBR service might mean higher resource utilization due to the prospect of exploiting the statistical multiplexing effect. Hence we focus on VBR services in this paper.

Another characteristic of most applications of this type is that they are either non-controllable or will suffer unacceptable qualify degradation from forced rate-control. Therefore, Connection Admission Control (CAC) becomes the primary method to allocate limited network resources in such a way that the user requirement for QoS can be satisfied. This presents a new challenge not found in traditional telecommunication networks. First of all, available resources (bandwidth and buffer) are fixed in amount and limited compared with user demand. On the other hand, the user traffic is ever-changing and QoS requirements can be stringent. This problem becomes even more difficult since so far there is no generic theoretical analysis method available to model the behavior of traffic sources (e.g., the long-range dependent video source [15]) and to predict performance accordingly.

So far, numerous efforts have been made to find an efficient CAC strategy which includes determining resource requirements for VBR traffic [3] [4] [5] [6] [7]. However, the proposed schemes are often limited to a particular aspect of the problem and lack a simple, generic strategy. In this paper, we try to address this issue by combining a variety of techniques, including on-line traffic measurement, dynamic renegotiation, and utilizing the statistical multiplexing effect.

The rest of this paper is organized as follows: Section 2 presents an overview of our CAC strategy as well as the underlying network architecture. In section 3, we discuss the issue of determining the user traffic descriptors. Section 4 addresses the problem of statistical multiplexing in CAC. Finally, section 5 draws a conclusion for the paper.

2: An Overview of CAC Strategy

As defined by ATM Forum [9], CAC is the set of actions taken by the network at virtual connection establishment in order to determine whether a connection can be accepted or should be rejected. Generally, CAC has to make the decision based on whether or not all connections (including both the existing ones and the new connection) will be able to achieve their QoS, given limited network resources.

Pricing policy is also important in any CAC strategy. Usually, price can be based on either or both of the following:

- Resource allocation, which may be measured in terms of declared traffic parameters.
- Actual usage (cell counts).

We take the position that pricing based on both factors is necessary to satisfy both users and network providers. Such policies provide incentives for both users and network providers to maintain consistency between resource allocation and actual usage.

A successful CAC strategy should achieve a good balance between the users' desire for QoS guarantees (conservative resource allocation) and the network provider's desire for maximum revenue (aggressive resource allocation). Furthermore, it should be relatively simple to implement, suitable to a wide range of traffic types, and able to deal with time-varying traffic.

As part of a comprehensive traffic management solution, CAC needs support from the following two aspects:

- A traffic description method accepted by both user and network. Currently ATM forum has chosen the Generic Cell Rate Algorithm (GCRA) as the basis of the user-network contract, as well as Usage Parameter Control (UPC) parameters at the user-network interface (UNI). The basic traffic parameters for VBR services are Sustainable Cell Rate

The authors can be contacted via e-mail at dwp@tisl.ukans.edu. This research is supported by Sprint Corporation

(SCR) and Burst Tolerance (BT). [1]

- An efficient underlying resource management scheme. Such a scheme should allocate resources to each connection to guarantee its QoS, yet allow dynamic resource sharing. In this paper, we use a bandwidth management scheme [1] based on an edge-core network architecture and Weighted Round-Robin (WRR) [10] [11] [12] cell scheduling. The basic idea can be expressed as follows:

    1. Partition the ATM network into VC-based edge network and VP-based core network.
    2. Ensure at least one CBR/VBR VP and one ABR/UBR VP between each edge-node-pair. The VP bandwidth is semi-permanently allocated based on long-term management considerations. In terms of CAC, this is the total amount of available bandwidth for that VP.
    3. Achieve both bandwidth allocation and dynamic sharing by using both VC-based WRR at the edge and VP-based WRR in the core.
    4. Throttle the traffic entering a VP at the edge, so that the loss and delay in the core will be minimal. Consequently, per VC bandwidth allocation at the edge would be enough to guarantee end-to-end QoS, allowing the CAC decision to be made at the ingress edge without requiring information from other nodes.

In this context, we propose a two-part CAC strategy:

1. Choose proper traffic descriptors (SCR and BT values) for each incoming connection and maintain accurate values using on-line measurement and dynamic renegotiation. Note that as shown later in this paper, if we allocate WRR bandwidth according to SCR, delay and loss bounds (QoS) can then be derived from SCR and BT. The baseline CAC strategy can then be expressed as: *allocate bandwidth and buffer according to SCR and BT. If sufficient spare resources are available, the call can be accepted. Otherwise it should be rejected.*
2. Since the baseline strategy is likely to be quite conservative, the second part of the strategy is to enhance the CAC performance (number of admissible connections) by taking the statistical bandwidth multiplexing (enabled by WRR) into consideration. Note that proper traffic descriptors will still be essential for the success of this enhanced strategy.

3: Problem 1: Traffic Descriptor Selection

The (SCR, BT) traffic description that will yield zero violation for a given type of traffic is not unique. For example, given an SCR value, there exists a BT_{min} such that for any $BT \geq BT_{min}$, the policer based on (SCR, BT) will give zero cell-tagging. More important, BT_{min} itself will vary with SCR. Clearly, the total number of admissible (SCR, BT) pairs is infinite. The choice of (SCR, BT) is important since it is directly related to resource allocation (bandwidth and buffer) and QoS (delay and loss ratio).

The problem then becomes: Given a certain delay bound and CLR requirement, determine the corresponding (SCR, BT) value that will satisfy the requirements and yet minimize the amount of resource allocation. In this section we propose addressing this problem by employing two key techniques: *virtual buffer measurement* and *UPC parameter renegotiation*.

3.1: Virtual Buffer Measurement

In virtual buffer(VB) measurement [2], the virtual buffer is actually a cell counter, which increases by one on cell arrival and decreases at a preset drain rate. The counter value is sampled (perhaps on every cell arrival) for measurement processing. Depending on the processing techniques, various kind of results can be obtained, such as the maximum VB counter value and probability of VB overflow. In practice, the data processing function should maintain currency, for example by working in a sliding-window fashion.

The importance of the above measurement is two-fold:

1. Simulates a FIFO with fixed serving rate equal to allocated bandwidth, providing worst-case delay and loss estimates.
    - The VB counter value is equal to the worst-case FIFO queue length. The maximum observed VB counter value corresponds to the buffer size that will yield zero cell loss.
    - The queue length in the VB observed at cell arrival is directly proportional to the worst-case delay experienced by the arrived cell.
2. Simulates a GCRA (leaky bucket) policer, allowing UPC adjustments.
    - SCR corresponds to the VB drain rate, and BT to the size of VB.
    - The VB overflow event is equivalent to cell-tagging in the corresponding UPC function.

One great advantage of VB-based measurement is that it is very simple and low-cost. The network provider can easily put it at the UNI or any other place to monitor traffic. Also it is very flexible and can be applied for many different purposes, which we will discuss in the rest of this paper.

3.2: Off-Line Traffic Characterization

Given a certain kind of incoming traffic, if we use a number of VBs with different drain rates (SCR) in parallel and record the maximum delay for each SCR, the result will be a delay-bound vs. SCR curve, which is an important characteristic of this traffic. If the curve is already known, the network provider can then easily determine GCRA parameters and resource allocation by picking an SCR/BT pair that satisfies the user's delay requirement.

Unfortunately, the exact curve generally can not obtained until a connection is admitted to the network. However, since the curve itself is an important characteristic of the particular traffic type, the result measured from pre-sampled trace files can serve as a guideline for user-network traffic contract during the initial CAC.

Since it is commonly believed that digital video traffic will be a significant portion of VBR traffic in B-ISDN, we have examined a number of video samples, including both MPEG-I coded and JPEG coded traces, for the above purpose. The results are shown in figures 1 - 5. In all these figures, we have assumed a frame-level “bursty” source, i.e., the source segments a whole video frame and transmits the resulting cells at very high link rate (OC-3, or 155 Mb/s).

The result as shown in figure 1 is acquired by measuring a 2-hour MPEG-I video trace (the movie “Star Wars”, from BELLCORE). There are four curves, representing CLR requirement of 0 (no loss/tagging happens), 10^{-5}, 10^{-4} and 10^{-3} respectively. The most striking feature is that a sharp knee is present in all four curves. This seems to be a universal characteristic, since we observed the same feature in all our experiments. The implication is that there exists a bandwidth threshold operation point

[1] Peak Cell Rate (PCR) and Cell Delay Variation Tolerance (CDVT) are also defined but the values for these are usually determined by equipment configurations.

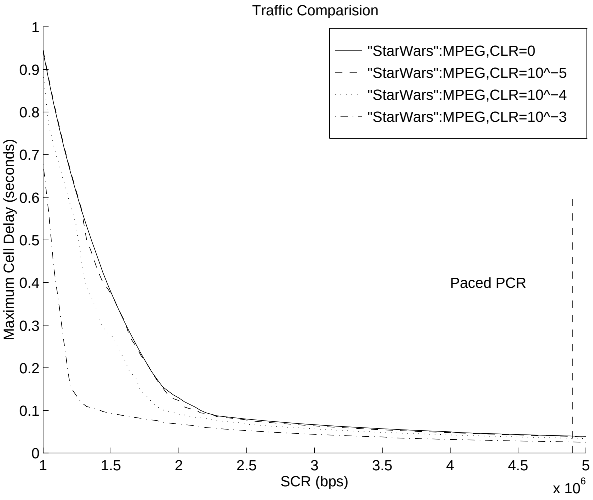

Figure 1. SCR-Delay relationship: different CLR

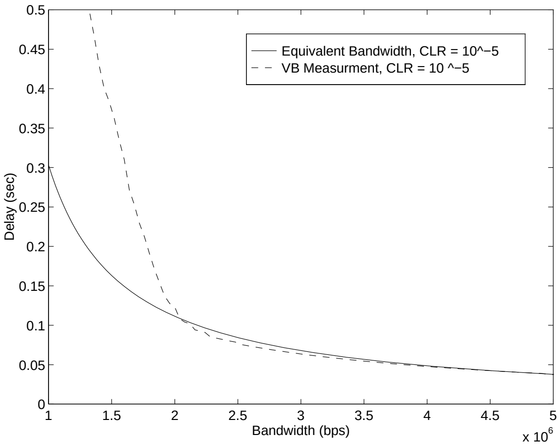

Figure 2. Comparison with Equivalent Bandwidth: MPEG Star Wars

below which the delay and buffer requirement are very sensitive to bandwidth, but above which the delay and buffer requirement changes little. The natural choice of operational bandwidth is then somewhere above, yet "safely" close to the threshold. Though the exact threshold value will depend on the type of traffic, note that the threshold bandwidth for this video trace is much lower than the often-suggested paced peak rate (maximum frame size divided by frame interval) of $4.9 Mb/s$ despite the fact that we have used a very bursty source.

We can also see that the bandwidth requirement varies, sometimes considerably, with different CLR requirements. For example, the threshold bandwidth drops almost 30% when the CLR requirement changes from 10^{-4} to 10^{-3}.

To further investigate the bandwidth requirement for the video sources, we compare the results obtained from VB measurement with those obtained by the well-known Equivalent Bandwidth (EBW) [13] [14] method. Based on an on-off fluid flow model, the EBW method estimates the bandwidth requirement from mean burst length, mean bit rate, peak rate, buffer size, and CLR requirement. In our experiments, burst lengths and rates are measured from the trace files, and the delay bound is considered as the buffer size divided by the resulting bandwidth.

As shown in figure 2, for the same "Star Wars" MPEG trace, the bandwidth estimation by the EBW method is close to that obtained by VB measurements when the delay bound is less than 100 ms. For larger delay bound, the EBW method tends to underestimate the bandwidth requirement considerably. Figure 3 shows a similar comparison using a JPEG encoded TV broadcast program trace. Here the EBW method always *overestimates* the bandwidth requirement. However, in other cases not shown due to space limitations, we have also observed that EBW can underestimate bandwidth compared to VB measurement. These results are further evidence that the bandwidth requirement for VBR traffic generally can not be obtained from currently existing traffic models.

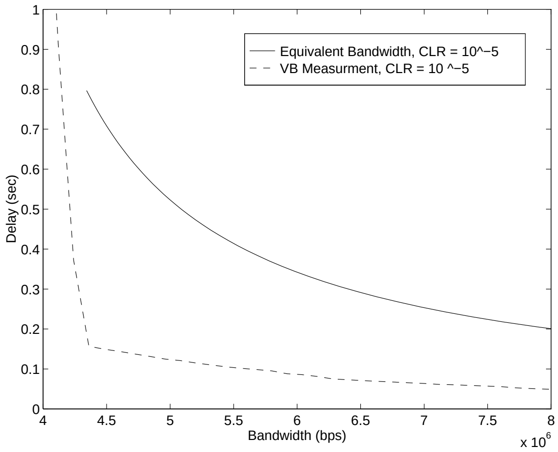

Figure 3. Comparison with Equivalent Bandwidth: JPEG TV program

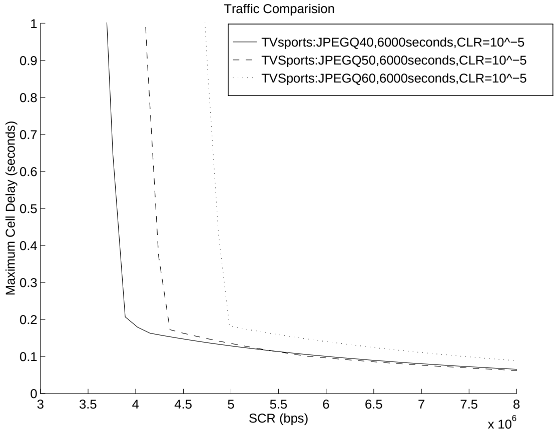

Figure 4. SCR-Delay relationship: different quality

Figure 3-5 are all based on sequential JPEG encoded video. All video samples are obtained by recording 100 minutes of broadcast TV programs. The sampling and encoding is done using the SunVideo video system on SUN SPARC workstations.

In figure 4, we examine the sensitivity to different encoding quality (Q40 is lower quality than Q50). The result is just as expected: the bandwidth requirement for the same QoS rises as the encoding quality gets better. However, the shape of curve remains the same, which means the effect of varying encoding quality is generally predictable.

Finally, in figure 5, we illustrate the sensitivity to the program content. Generally, there can be considerable variation caused by program content. As a result, the network operator and user may have to choose the worst-case curve (the rightmost one) for initial CAC decisions, since neither of them are likely to have an accurate estimation of the precise content of traffic. However, it is possible that more in-depth and systematic study will reveal some general principle regarding this case.

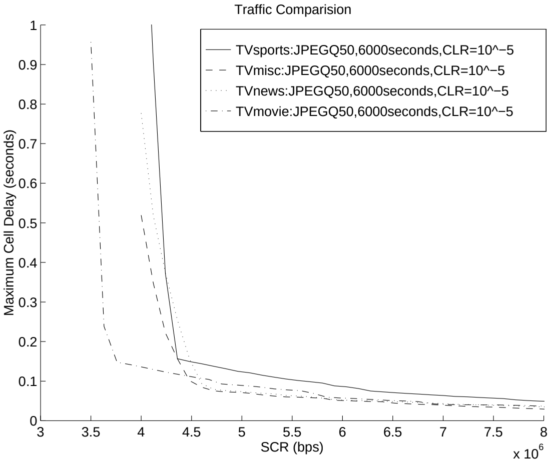

Figure 5. SCR-Delay relationship: different material

3.3: UPC Parameter Renegotiation

Although results acquired from representive trace files can provide a starting point for the user-network traffic contract, more accurate knowledge can only be obtained *during* the connection by using VB-based measurement on-line. Once the better traffic estimation is obtained, a *renegotiation* is necessary to change the UPC parameters accordingly. For example, if current UPC parameters (and hence the allocated resource) is more than necessary to guarantee QoS, it is beneficial to both user (lower resource allocation cost) and network (higher potential usage revenue from other connections) to change it to a lower setting.

Generally, there are two kinds of renegotiation scenarios:

User-initiated renegotiation Based on the knowledge of future traffic or user-side measurements, a user may initiate a renegotiation procedure. If the new set of UPC parameter requires additional resource, the network will decide whether to accept the request according to currently available resource, using a similar procedure as the one for initial CAC. If less resource is demanded, the request can always be accepted.

Network-initiated renegotiation In this case, the network keeps track of actual resource usage of user traffic and initiates a renegotiation at an appropriate time.

In this paper we mainly focus on the latter case. However, it should be kept in mind that, in either case, it is up to the user to make the final decision whether to adapt to new UPC parameters or not, especially when the new UPC parameters mean lower SCR. Once the user agrees to lower his SCR (e.g., in exchange for lower price), he takes the risk that the network may allocate the corresponding resource to other users, and he may not be able to get it back in case he needs it in the future. Given all those considerations, we propose the following network-initiated UPC renegotiation procedure:

1. Initial stage: there are two options in this stage:

   Option 1: user specifies a general traffic type (e.g., a MPEG encoded video) as well as delay and CLR requirements. The network then looks into the trace-file data base and suggests a safe set of UPC parameters for the user, who must give final approval.

   Option 2: the user specifies UPC parameters directly without consulting the network.

2. The network continuously monitors the traffic submitted by the user (by doing VB-based measurement) and generates new reference UPC parameters periodically. The length of period depends on the nature of the particular traffic.

Traffic Type	MPEG-1	MPEG-1	JPEG video	JPEG video
Number of Sources	6	6	6	6
Delay Bound (ms)	100	50	100	50
$\sum SCR$ (Mbps)	10.2	19.9	69.5	73
Multiplexed ($Mbps$)	5.72	11.5	44.4	46.3
Multiplexing Gain	44%	42%	36%	37%

Table 1. Statistical Multiplexing Gain on Video Sources

3. The network then checks the currently available resource. If the new UPC parameters can be guaranteed, the network then initiates renegotiation by sending the suggested UPC parameters to the user. Otherwise no renegotiation should be started. Other useful information, such as current cell-tagging ratio, could also be provided.

4. The user then evaluates the suggested UPC parameters against foreseeable future usage and cost implications, and decides whether to accept the new traffic contract.

4: Problem 2: Dealing with Statistical Multiplexing Effect

While the CAC approach that relates the traffic descriptors directly to resource requirements does provide QoS guarantees for admitted connections, it ignores a great advantage of ATM networks, *statistical multiplexing*. Actually, supported by a bandwidth-sharing scheme such as WRR, it is possible to reduce the total amount of required bandwidth considerably.

Table 1 shows some results of video source multiplexing. The sources used here are six 20 minute segments taken from the movie “Star Wars”, either MPEG-1 or JPEG encoded. The fourth row is the sum of SCRs corresponding to the required delay bound. The fifth row is the total bandwidth required to achieve the same delay bound after the traffic from the sources is multiplexed using WRR. The multiplexing gain is defined as $(\sum SCR - MultiplexedBW)/\sum SCR$. Clearly, there are significant multiplexing gains regardless of delay requirements, even if the total number of sources is relatively small. Furthermore, many studies [6] [7] show that, for MPEG video sources, when the number of multiplexed sources increases, the aggregated bitrate distribution becomes more Gaussian and narrow. As a result, the aggregated peak rate tends to get closer to aggregated mean rate, and the effect of statistical multiplexing becomes even more significant.

However, in real life a VP will probably carry many kinds of traffic with greatly-varying characteristics, and the statistics such as mean rate, peak rate and burst size, which are required in many previous studies, can only be obtained *during* the lifetime of the connection. Meanwhile, when a network operator decides to take advantage of statistical multiplexing to increase total admission, he also takes the risk of possible over-admission and the resulting QoS degradation. For example, there is no guarantee that all users will not transmit at SCR at the same time. To avoid this situation as much as possible, it is necessary to constantly monitor current resource usage. Therefore, a practical CAC strategy considering the multiplexing effect should generally employ some kind of on-line measurement.

To deal with the above problem, we now propose a CAC strategy based on estimation of actual usage of bandwidth. The strategy can be expressed as follows:

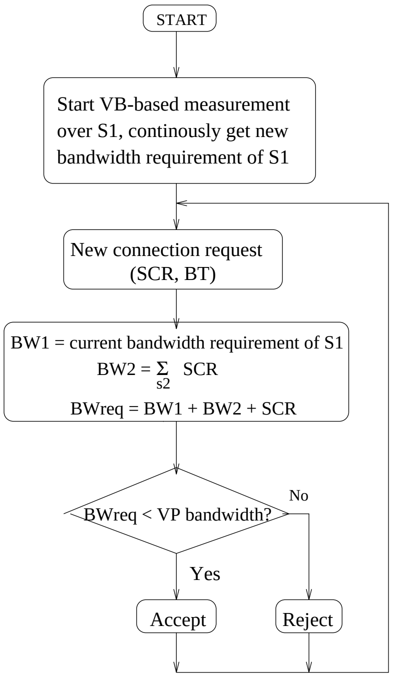
START
Start VB-based measurement over S1, continously get new bandwidth requirement of S1
New connection request (SCR, BT)
BW1 = current bandwidth requirement of S1
BW2 = Σ s2 SCR
BWreq = BW1 + BW2 + SCR
BWreq < VP bandwidth?
No
Yes
Accept
Reject

Figure 6. CAC algorithm based on actual usage

Let T_w be a pre-defined measurement window width, and for each VC, let T_s be the time elapsed since admission. Define two sets $S_1 = \{VCs : T_s > T_w\}$, $S_2 = \{VCs : T_s < T_w\}$, so that S_2 contains VCs for which measurements are not available.

As shown in figure 6, the network constantly monitors the actual aggregate bandwidth usage of all VCs belonging to S1 by VB-based measurement as described in section 3.2:. Again, the result is a delay-bound vs. SCR curve. The actual VP bandwidth usage is then estimated as the SCR value that satisfies the most stringent delay and CLR requirement of all the VCs in the VP under investigation.

For those VCs belonging to S_2, the bandwidth is always estimated as the claimed SCR. Similarly, the bandwidth requirement of the incoming connection request is also estimated as its SCR. The admission criteria then becomes:

if the sum of estimated bandwidth of both VCs already admitted and the incoming request is greater than VP bandwidth, reject the new call, otherwise the call can be accepted.

As we stated before, this kind of CAC approach can be risky and should be applied with caution. For example, in practice it is probably desirable for the network operator to set a high watermark of bandwidth usage (e.g., 90% of physical VP bandwidth), and use that as VP bandwidth in the above CAC procedure.

Another open issue is the choice of measurement window T_w. Generally, a larger T_w means a more conservative strategy, and a smaller T_w means more aggressive. Furthermore, it may be desirable to combine results from several measurement windows. We hope a good rule can be found through further experiments and analysis.

5: Conclusion

In this paper, we propose a new CAC strategy for real-time VBR services in ATM networks. First, we introduce the idea of Virtual Buffer measurement for resource usage and UPC parameters. Then, we discuss and illustrate how to obtain accurate UPC parameters for user traffic, by employing Virtual Buffer measurement and dynamic renegotiation. The baseline (conservative) CAC strategy is tightly couples resource allocation and UPC parameters. From this basis, we move further to examine the possible resource gain from statistical multiplexing effects, and propose a more aggressive CAC strategy to exploit these effects. Further work will focus on performance characterization and implementation details.

References

[1] K. Liu, H. Zhu, D. W. Petr, V. S. Frost, C. Braun, and W. Edwards, “Design and Analysis of a Bandwidth Management Framework for ATM-Based Broadband ISDN,” *Proc. IEEE ICC'96*, pp. 1712-1716, 1996.

[2] H. Zhu and V. S. Frost, “In-Service Monitoring and Estimation of Cell Loss Ratio QoS in ATM Networks,” *IEEE/ACM Transactions on Networking*, Vol. 4, No. 2, pp. 240-248, April, 1996.

[3] F. Guillemin , C. Rosenberg and J. Mignault, “On Characterizing an ATM Source via the Sustainable Cell Rate Traffic Descriptor,” *Proc. IEEE INFOCOM'95*, pp. 1129-1136, 1995.

[4] A. R. Reibman and A. W. Berger, “Traffic Descriptors for VBR Video Teleconferencing Over ATM networks,” *IEEE/ACM Transactions on Networking*, Vol. 3, No. 3, pp.329-339, June 1995.

[5] D. J. Reininger, D. Raychaudhuri and J. Y. Hui, “Bandwidth Renegotiation for VBR Video over ATM Networks,” *IEEE JSAC*, Vol. 14, No. 6, pp. 1076-1085, August 1996.

[6] J. Mata, G. Pagan and S. Sallent, “Multiplexing and Resource Allocation of VBR MPEG Video Traffic on ATM Networks,” *Proc. IEEE ICC'96*, pp. 1401-1405, 1996.

[7] M. Krunz, R. Sass and H. Hughes, “Statistical Characteristics and Multiplexing of MPEG Streams,” *Proc. IEEE INFOCOM'95*, pp. 455-462, 1995.

[8] The ATM Forum Technical Committee, *User-Network Interface (UNI) Specification Version 3.1*, 1994.

[9] The ATM Forum Technical Committee, *Traffic Management Specification Version 4.0*, AF-TM 0056.000, April, 1996.

[10] K. Sriram, “Dynamic Bandwidth Allocation and Congestion Control Schemes for Voice and Data Multiplexing in Wideband Packet Technology,” *Proc. IEEE ICC'90*, Atlanta, Georgia, pp. 1003-1009, April 1990.

[11] M. Katevenis, S. Sidiropoulos and C. Courcoubetis, “Weighted Round-Robin Cell Multiplexing in a General-Purpose ATM Switch Chip,” *IEEE JSAC*, Vol. 9, No. 8, pp. 1265-1279, October 1991.

[12] Y. Wang, T. Lin and K. Gan, “An Improved Scheduling Algorithm for Weighted Round-Robin Cell Multiplexing in an ATM Switch,” *Proc. IEEE ICC'94*, pp. 1032-1037, 1994

[13] Roch Guérin, Hamid Ahmadi, and Mahmoud Naghshineh, “Equivalent Capacity and Its Application to Bandwidth Allocation in High-Speed Networks” *IEEE JSAC* Vol. 9, No. 7, pp. 969-981, September 1991.

[14] C. Braun, D. W. Petr and T. G. Keley, “Performance Evaluation of Equivalent Capacity for Admission Control,” *Proc. IEEE Wichita Conference on Communications, Networking and Signal Processing*, April 1994

[15] J. Beran, R. Sherman, M. S. Taqqu and W. Willinger, “Long-Range Dependence in Variable-Bit-Rate Video Traffic,” *IEEE Transactions on Communications*, Vol. 43, No. 2/3/4, pp. 1566-1579, Feb./Mar./Apr. 1995.