Paper: TSCS-113

Title: Design and simulation of ATM-ABR end system congestion control

Authors:
Sonia Fahmy
Department of Computer Sciences, Purdue University
Raj Jain
Department of CIS, The Ohio State University
Rohit Goyal
Axiowave Networks
Bobby Vandalore
Nokia

Raj Jain is now at
Washington University in Saint Louis
Jain@cse.wustl.edu
http://www.cse.wustl.edu/~jain/

Corresponding author:
Sonia Fahmy
1398 Computer Science
West Lafayette, IN 47907-1398
USA
Tel: (765) 494-6183, Fax: (765) 494-0739,
E-mail: fahmy@cs.purdue.edu

Design and simulation of ATM-ABR end system congestion control

Sonia Fahmy[1]
Department of Computer Sciences, Purdue University
Raj Jain
Department of CIS, The Ohio State University
Rohit Goyal
Axiowave Networks
Bobby Vandalore
Nokia

Abstract—**We develop a simulation model for the ATM ABR service, and use it to engineer ABR congestion control behavior. Although significant work has been performed on ABR rate allocation algorithms at network switches, little work has focused on the end system behavior, which we examine in this paper. The effect of the speed of links on the path from the source to the destination, and the connection round trip time on the selection of ABR parameter values is studied. Simulation results illustrate the impact of the key parameters that control rate reduction in the absence of network feedback on performance, in terms of connection throughputs, queue lengths at the switches and link utilizations. These results have been incorporated into the ABR standards, and can be generalized to cooperative congestion control with explicit congestion notification (ECN) in the Internet.**

Keywords— **traffic management, congestion control, simulation models, ATM networks, available bit rate (ABR) service, ABR end system, ABR parameter tuning**

I. Introduction

ATM networks offer six service categories: constant bit rate (CBR), real-time variable bit rate (rt-VBR), non-real time variable bit rate (nrt-VBR), available bit rate (ABR), unspecified bit rate (UBR), and guaranteed frame rate (GFR). The ABR, UBR, and GFR service categories are specifically designed for data traffic. The ABR service provides better service for data traffic than UBR and GFR by frequently indicating to the sources the rate at which they should transmit. ABR can thus provide minimum rate guarantees and low cell loss to ABR sources.

The ABR source end system is allowed to send data at a given rate called Allowed Cell Rate (ACR), which ranges between a negotiated Peak Cell Rate (PCR) and Minimum Cell Rate (MCR). Immediately after establishing a connection, ACR is set to an Initial Cell Rate (ICR), which is also negotiated with the network. The source sends a Resource Management (RM) cell every Nrm$-$1 data cells (default Nrm value is 32), and the destination end system turns the RM cells around. As seen in figure 1, RM cells traveling from the source to the destination are called forward RM cells (FRMs), while the RM cells traveling from the destination back to the source are called backward RM cells (BRMs). The RM cells collect network feedback and return to the source, which adjusts its allowed cell rate according to this feedback [2], [9]. Most RM cells generated by the sources are considered part of the source load in the sense that the total rate of data and RM cells should not exceed the source ACR. Such RM cells are called "in-rate" RM cells. Under exceptional circumstances, switches, destinations, or sources can generate extra RM cells. These "out-of-rate" RM cells are not counted in the ACR of the source and are distinguished by having their cell loss priority (CLP) bit set, meaning that the network will transport them only if there is enough bandwidth, and discard them if congested.

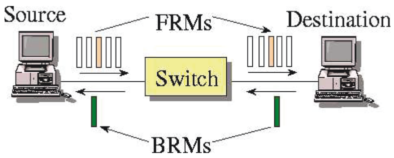


Fig. 1. Forward and backward RM cells

At the time of connection setup, ABR sources negotiate several parameters with the network. A complete list of parameters used in the ABR mechanism is presented in table I. There are three ways for the network switches to indicate their feedback to the sources. First, each cell header contains a bit called Explicit Forward Congestion Indication (EFCI), which can be set by a congested switch. Second, RM cells have two bits in their payload, called the Congestion Indication (CI) bit and the No Increase (NI) bit, also set by congested switches (such switches are called relative rate marking (RRM) switches). Third, the RM cells have a field called explicit rate (ER) that can be reduced by congested switches to any desired value. When sources receive the returning RM cells, they adjust their ACR accordingly. A complete explanation of the source and destination rules is presented in [5].

This paper develops an ATM-ABR simulation model and uses it to identify key factors which influence ABR performance. We focus on end system behavior, and not rate allocation algorithms in network switches, which are studied in most of the earlier ABR work [7], [8]. Our results have been incorporated into the ABR standards, and similar techniques have recently appeared in the Internet active queue management literature [10]. In the remainder of this paper, we examine each set of related parameters in consecutive sections, identifying the effect of each on performance, and giving guidelines on setting each.

II. Rate Upper and Lower Bounds

Role. The peak cell rate (PCR) and the minimum cell rate (MCR) are used in **source rule 1**. The rule states that source should always transmit at a rate equal to or below its computed ACR, which cannot exceed PCR and need not go below MCR, i.e., MCR $\leq$ ACR $\leq$ PCR and Source Rate $\leq$ ACR. PCR is the maximum value at which a source can transmit. It must be negotiated down and has no default value. MCR is the minimum value that a source need not reduce its rate beyond. It is negotiated down to the minimum acceptable MCR, MCRmin, only if MCRmin is signaled.

Values. ABR sources may initially set PCR to the maximum possible value. For example, PCR can be set according to the capacity of the application or host, or the bandwidth of the link from the host to the next node. Of course, pricing considerations play a key role in parameter selection: sources may select a lower PCR value if they are unwilling to incur the costs. MCR can be set according to user requirements (e.g., video applications require some minimum rate guarantee), and the pricing policy. Unless the traffic is high priority, MCR is usually set to zero, making the service a best effort one. Most applications, especially TCP/IP applications, however, work better with an

[1]– The authors can be reached at address: Sonia Fahmy, 1398 Computer Science, West Lafayette, IN 47907-1398, USA, Tel: (765) 494-6183, Fax: (765) 494-0739, E-mails: fahmy@cs.purdue.edu, jain@cis.ohio-state.edu, rgoyal@axiowave.com, bobby.vandalore@nokia.com

TABLE I

ABR PARAMETERS

Label	Expansion	Units and Range	Default	Signaled?
PCR	Peak Cell Rate	cells/second from 0 to 16M	–	down
MCR	Minimum Cell Rate	cells/second from 0 to 16M	0	down to MCRmin
ACR	Allowed Cell Rate	cells/second from 0 to 16M	–	no
ICR	Initial Cell Rate	cells/second from 0 to 16M	PCR	down
TCR	Tagged Cell Rate	constant	10 cells/s	no
Nrm	Number of cells between FRM cells	power of 2 from 2 to 256	32	optional
Mrm	Controls bandwidth allocation between FRM, BRM and data cells	constant	2	no
Trm	Upper Bound on Inter-FRM Time	milliseconds, 100 × power of 2 from −7 to 0	100 ms	optional
RIF	Rate Increase Factor	power of 2 from 1/32768 to 1	1	down
RDF	Rate Decrease Factor	power of 2 from 1/32768 to 1	1/32768	up, or down by ≤ RIF decrease factor
ADTF	ACR Decrease Time Factor	seconds, from 0.01 to 10.23 seconds in steps of 10 ms	0.5 s	optionally down
TBE	Transient Buffer Exposure	cells from 0 to 16,777,215	16,777,215	down
CRM	Missing RM-cell Count	integer of unspecified size	$\lceil\frac{TBE}{Nrm}\rceil$	computed
CDF	Cutoff Decrease Factor	zero or a power of 2 from 1/64 to 1	1/16	optionally up
FRTT	Fixed Round-Trip Time	microseconds from 0 to 16.7 seconds	–	accumulated

MCR greater than zero, to prevent timeouts. Observe that charging considerations may limit PCR to a multiple of MCR, i.e., $PCR = k \times MCR$, where: $2 \leq k \leq 10$. This simplifies traffic shaping.

Switches can reduce the PCR and MCR according the connection admission control (CAC) algorithm. One possible simple policy is to ensure that, after admitting the new connection: $\Sigma PCR_{CBR} + \Sigma SCR_{VBR} + \Sigma MCR_{ABR} \leq$ link bandwidth. Hence, the MCR of the new connection can be computed as: $MCR_i \leq min($User-requested:MCR_i, link bandwidth $- \Sigma PCR_{CBR} - \Sigma SCR_{VBR} - \Sigma_{j \neq i} MCR_{ABRj})$. If the signaled MCR is less than the minimum acceptable MCR, i.e., $MCR_i < MCRmin_i$, the connection is rejected.

The PCR of the ABR connection is only limited by the bandwidth of the links on the path from the source to the destination: $PCR_i = min(PCR_i, \forall j, j \in$ links on path,minimum (link bandwidth)$_j)$. Therefore, PCR and MCR are dependent on the bottleneck link bandwidth, but not on the round trip time (RTT) of the connection.

III. RM CELL FREQUENCY CONTROL

Role. The three parameters Nrm, Mrm and Trm control the frequency of generation of resource management cells at the source. They are used in **source rule 3**. At any instant, sources have three kinds of cells to send: data cells, forward RM cells, and backward RM cells (corresponding to the reverse flow). The relative priority of these three kinds of cells is different at different transmission opportunities.

The sources are required to send an FRM after every Nrm cells. If the source rate is low, however, the time between RM cells will be large and network feedback will be delayed. To overcome this problem, a source should send an FRM cell if more than Trm milliseconds have elapsed since the last FRM was sent. This introduces another problem for low rate sources. In some cases, at every transmission opportunity, the source may finds that it has exceeded Trm and needs to send an FRM cell. In this case, no data cells will be transmitted. To overcome this problem, an additional condition must be satisfied there must be at least two (Mrm) non-FRM cells between FRMs. A waiting BRM has priority over waiting data, given that no BRM has been sent since the last FRM. Of course, if there are no data cells to send, waiting BRMs may be sent. The second and third part of source rule 3 ensure that BRMs are not unnecessarily delayed and that all available bandwidth is not used up by the RM cells. Figure 2 illustrates the scheduling of FRMs, BRMs and data cells.

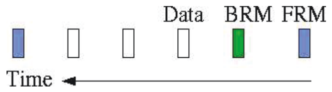


Fig. 2. Scheduling of forward RM, backward RM, and data cells

Values. Mrm is constant at 2, and is not negotiated at connection setup. We next discuss the setting of Nrm and Trm.

Nrm. The specification [2] select a default value of 32 for Nrm to ensure that the control overhead does not exceed approximately 6% (the value with window-based flow control). During normal operation, $\frac{1}{32}^{nd}$ or 3% of all cells are FRM cells. Another 3% of cells are BRM cells resulting in a total overhead of 6% [3]. Nrm is independent of link speed and round trip time, since it is simply a ratio.

In practice, the choice of Nrm affects the responsiveness of control system, and the computational overhead at the end systems and switches. For a connection running at 155 Mbps, the inter-RM cell time is 86.4 μs, while it is 8.60 ms for the same connection running at 1.55 Mbps. The inter-RM interval determines the responsiveness of the system. Sources, destinations, and switches may wish to increase Nrm if their processing power is limited, or if they wish to minimize rate variations of the ABR connection, or increase the data cell frequency. They may wish to decrease Nrm if fast rate changes are desirable, and responsiveness to network feedback is advantageous. At high data rates, a small RM cell interval can result in high frequency rate variations caused by the ABR feedback. If real-time video traffic is transported over ABR (which we have shown is feasible with the correct parameter choices [11]), rate variations must be minimized to reduce variations in the quality of service. One way of reducing the ABR rate changes is to send RM cells less frequently, i.e., set Nrm to a large value, instead of 32. Sending RM cells at end of each video frame is one possible option.

We build a complete ATM simulation model which includes all six service categories, and various application traffic models, and switch rate allocation algorithms. We use this model to vary Nrm and examine the allowed cell rates at the sources, the queue lengths at bottleneck switches, the link utilizations, and the throughput at the destinations. Since the Nrm value must be a power of two that is allowed to range between 2 and 256 [2]), we conduct experiments with all the allowed Nrm values (2, 4, 8, 16, 32, 64, 128 and 256). However, we only show the simulation results for Nrm = 8 and 256 here. This is because values smaller than 8 incur a very high control cell overhead and are not very realistic; and 256 is the maximum allowed value. In our simulations, all links are 155.52 Mbps. The initial cell rate (ICR) of all sources is set to 150 Mbps, while the remaining ABR parameters are set to their default values as given in the specifications In particular, note that the value of the rate increase factor (RIF) parameter is set to 1/16. The ERICA [7] scheme is used in this study, with switch averaging interval set to a fixed time of 5 ms, and target utilization set to 90% of the link capacity. The configuration simulated consists of two ABR sources: source 1 sends data at its ACR throughout the simulation, while source 2 is a transient source that comes on at 100 ms and sends data for about 100 ms. All link lengths are 1000 km. The main aim of this simple dumbbell configuration is to test the effect of Nrm on responsiveness of the system.

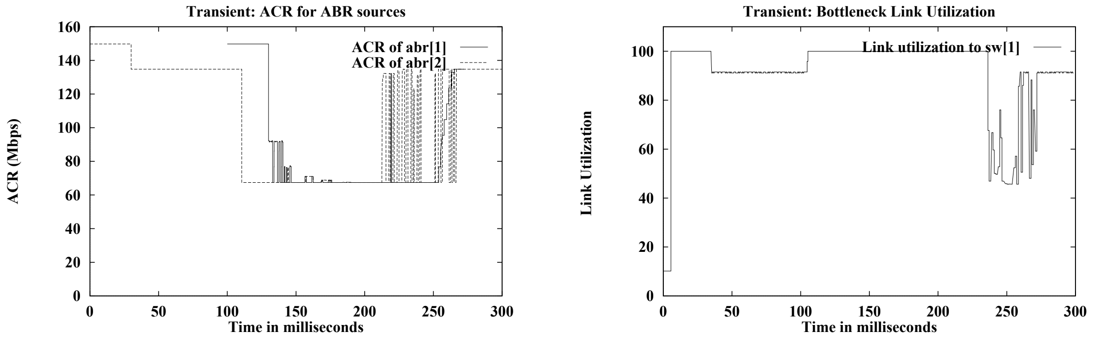
ABR performance for this transient configuration is shown in figure 3 and 4 for Nrm = 8 and 256. We show the ACRs of the two sources, and the link utilization at the bottleneck link. In all cases, source 1 ACR is rapidly reduced to its target value of about 140 Mbps. When source 2 starts sending data, the ACRs of both sources are reduced to 70 Mbps. When source 2 stops sending data, the ACR for source 1 returns to 140 Mbps. There is a difference in the rate of increase of ACR for the Nrm values. Since RIF is set to 1/16, the ACR increases in steps on the receipt of every BRM cell. Since the source receives BRMs more frequently with Nrm = 8 versus 256, the ACR for source 1 reaches 140 Mbps fastest in this case. The overhead with small Nrm values is quite high, however. This can be clearly observed by measuring the throughput at the application layer at the destinations (these plots are not shown here). Another interesting observation is that for smaller Nrm values, source 1 does not start rising as rapidly as with larger Nrm values, because the high RM cell overhead causes the data of the second source to take a longer time to be transmitted, and hence the two sources must share the bottleneck link for a longer time. Table II shows the variation of inter-RM cell time with link speed and with Nrm value. The source is assumed to be sending at link rate for the values shown in the table. A general heuristic is to use Nrm of 32 at speeds below OC-3 and to use Nrm of 256 for OC-3 and higher speeds.

TABLE II
INTER-RM CELL TIME FOR DIFFERENT SPEEDS AND NRM

Total ABR Capacity	DS0 64 kbps	T1 1.5 Mbps	OC-3 155 Mbps	OC-24 1.2 Gbps
Nrm = 8	0.5 s	24 ms	24 μs	3 μs
Nrm = 32	2.3 s	96 ms	96 μs	12 μs
Nrm = 256	18.4 s	768 ms	768 μs	96 μs

Trm. The Trm parameter is used with low rate sources: Trm is compared to the time elapsed since the last in-rate FRM cell was sent. Sources may be limited to a low ACR due to high amplitude VBR traffic sharing the same resources as the ABR connection, a large number of ABR sources, or low bottleneck link speeds (T1 links). Smaller Trm values result in shorter time between RM cells, leading to faster transient response (rise from low rate to high rate). Small Trm values, however, increase overhead with low rate sources. The choice of Trm depends on the link speed. For example, at a rate of 155 Mbps, the inter-cell time is 2.7 μs, while at a rate of 1.5 Mbps, the inter-cell time is 270 μs, and at a rate of 2.4 Gbps, the inter-cell time is 0.42 ns. Thus, a Trm value of 100 ms seems more appropriate for $1.5-155$ Mbps than with higher (2.4 Gbps$+$) speeds, where a Trm of 100 ms is too long to wait before sending an FRM cell to sense the state of the network. Trm should be reduced in such cases. The switches or destination can compare Trm to the inter cell time calculated as the reciprocal of the negotiated PCR (which may indicate the bottleneck link bandwidth). A good heuristic value for Trm is: $Trm = \frac{1}{PCR} \times c$. One choice of c can be $\frac{1,000,000}{27}$. This is based upon the intuition that 100 ms was observed to be suitable for OC-3 links (2.7 microsecond = 0.0027 millisecond inter-cell time). Trm is independent of the round trip time, i.e., whether the connection is local to a LAN, crosses a WAN, or traverses a satellite link of hundreds of milliseconds delay. This is because Trm is compared to the time since the last in-rate FRM cell was sent, so it is independent of the time the RM cell reached the destination, or the time the RM cells returns back to the source.

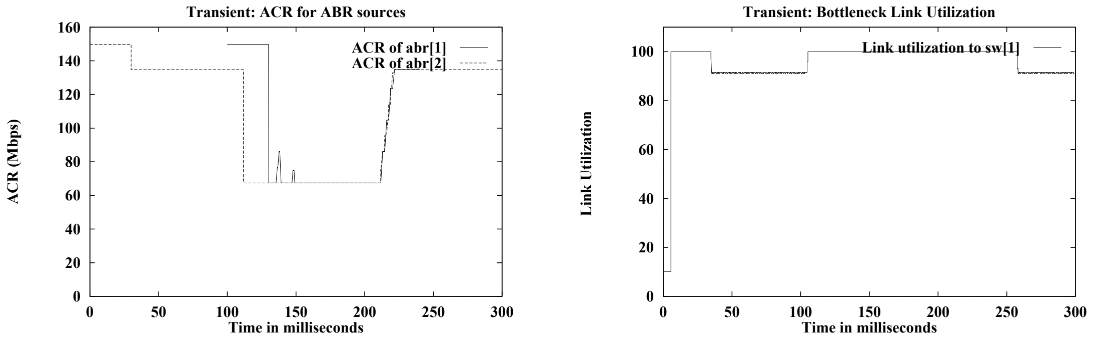
We experiment with Trm values of 1 and 100 ms with low rate sources. Our multi-class scheduler at ATM switches gives VBR traffic higher priority than ABR. We model VBR as a simple on/off source with 20 ms periods, and 138 Mbps peak rate. Simulation results (see figure 5) illustrate that in this case, capacity remains unused for a long time for large Trm values (100 ms), after VBR stops and capacity for ABR becomes available. Lower Trm values increase RM cell frequency and reduce response time. This is especially important for small or zero minimum cell rate connections.



(a) Allowed Cell Rate

(b) Link Utilization

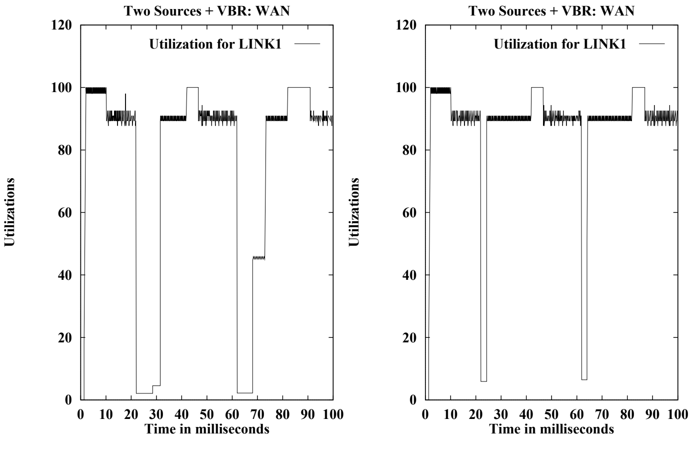
Fig. 3. Results for a WAN transient configuration Nrm = 8



(a) Allowed Cell Rate

(b) Link Utilization

Fig. 4. Results for a WAN transient configuration Nrm = 256



(a) Trm = 100 ms

(b) Trm = 1 ms

Fig. 5. Link utilization results for two sources and VBR in a WAN

IV. Rate Increase and Decrease Factors

Role. The rate increase factor (RIF) and rate decrease factor (RDF) are used in **source rules 8 and 9**. Source rules 8 and 9 describe how the source reacts to network feedback. The feedback consists of the explicit rate (ER), congestion indication bit (CI), and no increase bit (NI). A source does not simply change its ACR to the new ER due to the following reasons: (1) If the new ER is very high compared to current ACR, switching to the new ER may cause sudden overload in the network. Therefore, the rate increase factor (RIF) parameter determines the maximum allowed increase in any one step: RIF × PCR; (2) If there are any EFCI or relative rate marking (RRM) switches in the path, they do not change the ER field but set EFCI bits in the cell headers, or CI and NI bits in RM cells. The destination monitors EFCI bits in data cells, and returns the last seen EFCI bit in the CI field of a BRM. A CI of 1 means that the network is congested and that the source should reduce its rate by the rate decrease factor (RDF) parameter. Unlike the increase, which is additive, the decrease is multiplicative; and (3) The no-increase (NI) bit handles mild congestion by allowing a switch to specify an ER and instruct the source not to increase its rate if ACR is already below the specified ER. The actions corresponding to the possible values of CI and NI are:

NI	CI	Action
0	0	ACR ← min (ER, ACR + RIF × PCR, PCR)
0	1	ACR ← min (ER, ACR − ACR × RDF)
1	0	ACR ← min (ER, ACR)
1	1	ACR ← min (ER, ACR − ACR × RDF)

Once the ACR is updated, subsequent cells sent from the source conform to the new ACR value. However, if the earlier ACR was very low, it is possible that the very next cell had been scheduled to be sent a long time ahead. In such a situation, it is advantageous to "reschedule" the next cell, so that the source can take advantage of the high ACR allocation immediately [6].

Values. RIF and RDF play an important role when a connection traverses EFCI or RRM switches. In addition, some ER schemes work better with conservative RIF values, while others, such as ERICA [7] are insensitive to the RIF value, and work well with an RIF of 1.

RIF. The rate increase factor determines the maximum increase when a BRM cell indicating underload is received. If the RIF is set to a fraction less than one, the maximum increase at each step is limited to RIF × the peak cell rate for the connection. Setting RIF to small values is a more conservative strategy that controls queue growth and oscillations, especially during transient periods. It, however, may slow down the response of the system when capacity suddenly becomes available, leading to network underutilization.

If there are no EFCI switches in a network, setting RIF to 1 allows ACRs to increase as fast as the network directs (through the ER field) increasing utilization. For EFCI networks, or a combination of ER and EFCI networks, RIF should be set conservatively to avoid unnecessary oscillations [8]. Thus, sources can initially set RIF to large values, according to application requirements. During connection setup, any switch which does not implement an explicit rate scheme or implements a scheme which requires a conservative RIF (such as EPRCA) must reduce RIF to a conservative value such as 1/16 or less. RIF should also be set to more conservative values for high speeds (as indicated by PCR) and long round trip times (as indicated by FRTT) to avoid high congestion-related losses. The fixed part of the round-trip time (FRTT) is accumulated during connection setup. This is the minimum delay along the path and does not include any queueing delays.

Figures 4 and 6 compare the performance of the transient configuration (which was used in the Nrm experiments) with RIF set to 1/16 (the default value) and RIF set to 1. The basic ERICA scheme [7] is used in these simulations. Nrm is set to 256 to slow down the feedback rate, in order to emphasize the effect of RIF. All other parameters are the same as with the Nrm experiments. It is clear from figure 4 that an RIF value of 1/16 results in a step-wise increase of the rate of the non-transient source when the transient source stops transmission. With RIF set to 1 (figure 6), the rate of the non-transient source increases to the full rate as soon as the inactivity of the transient source is detected.

RDF. When the network is congested (the CI bit is set), the source multiples its current rates by (1-RDF). Thus the RDF parameter determines how fast the rate is reduced in case of congestion. This multiplicative decrease only occurs if the CI bit is set, either by the switches, or by the destination when the EFCI bits of data cells are set. The source should initially set RDF to a moderate value. Switches should reduce RDF dependent on the schemes they use for setting EFCI bits or CI bits. Explicit rate switches need not modify RDF. The RDF parameter should be set more conservatively (to smaller values) for higher speeds and longer round trip times to avoid a large amount of cell loss during congestion. Switches should examine the round trip time (in the FRTT field) and bottleneck link speed (as indicated by the PCR), and reduce RDF accordingly. If the switch or destination detects a large FRTT or large PCR (indicating a high bottleneck link speed), then RDF should be reduced.

V. Abnormal Conditions and Idle Periods

Since CRM and CDF are used with source rule 6, they are both discussed together. Both CRM and ICR are computed using the TBE parameter, so TBE and ICR are also discussed here, as well as ADTF that is used in conjunction with ICR.

Roles. TBE, CRM and CDF. The three parameters transient buffer exposure (TBE), missing RM cell count (CRM), and cutoff decrease factor (CDF) are used in **source rule 6**. This rule deals with the following scenario: if a network link fails, or becomes highly congested, RM cells are blocked and the source does not receive feedback. To protect the network from continuous in-flow of traffic under such circumstances, sources are required to reduce their rate if the network feedback is not received in a timely manner. This improves network fault tolerance.

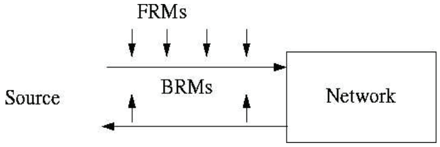
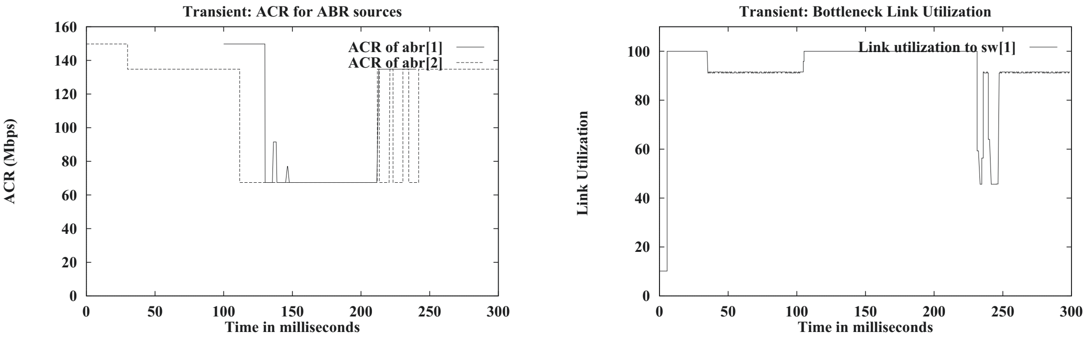


Fig. 7. Source rule 6 does not trigger if BRM flow is maintained

In steady state, a source should receive one BRM for every FRM sent (figure 7). The sources keep a count of the RM cells sent, and if no backward RM cells are received for a long time, the sources reduce their rate by a factor of “Cutoff Decrease Factor (CDF).” The “long time” is defined as the time to send CRM forward RM cells at the current rate. When rule 6 triggers, the condition is satisfied for all successive FRM cells until a BRM is received. Thus, this rule results in an exponential decrease of ACR. CRM is a function of a parameter called transient buffer exposure (TBE) negotiated at connection setup. TBE determines the maximum number of incoming cells at a switch during the first round trip, before the closed-loop phase of the control takes effect. During this time, the source can send TBE/Nrm RM cells. Hence, CRM $= \lceil \frac{\text{TBE}}{\text{Nrm}} \rceil$.

ICR and ADTF. Sources begin transmission at the initial cell rate (ICR) as specified by **source rule 2**. During the first round trip, a source may send as many as ICR × FRTT cells into the network. Since this number is negotiated separately as TBE, the following relationship exists between ICR and TBE: ICR × FRTT ≤ TBE, or: $ICR \leq \frac{TBE}{FRTT}$. The sources are required to use the ICR value computed above if it is less than the ICR negotiated with the network, i.e., ICR used by the source = min(ICR negotiated with the network, $\frac{TBE}{FRTT}$).

According to **source rule 5**, a source ACR is valid only for approximately ADTF seconds. If a source does not transmit any RM cells for this duration, it cannot use its previously allocated ACR, particularly if the ACR is high. The source should resense the network state by sending an RM cell and decreasing its rate to the initial rate (ICR) negotiated at connection setup. If the source ACR is already below ICR, it should not increase to ICR. The timeout interval is set to the ACR Decrease Time Factor (ADTF) parameter, whose default value is 500 ms. Rule 5 is intended to resolve the problem of *ACR retention*, when a source retains a rate allocated to it under light loads, and uses that rate to transmit when the network is highly loaded, caus-



(a) Allowed Cell Rate

(b) Link Utilization

Fig. 6. Results for a WAN transient configuration (Nrm = 256) RIF = 1

ing congestion. Several solutions to this problem (called *use it or lose it* (UILI) problem) were proposed, but the ATM Forum standardized a policy that simply reduces ACR to ICR when the timeout (ADTF) expires. Vendors are free to implement additional proprietary restraints at the source or at the switch.

Values. We first discuss the value of TBE and the two parameters which depend on it (CRM and ICR). Then, we discuss CDF, and finally ADTF.

TBE, CRM, and ICR. TBE determines the total number of cells that a switch may be suddenly "exposed" to during transients. TBE is specified in cells while CRM is specified in RM cells. Since there is one RM cell per Nrm cells, the relationship between CRM and TBE is: CRM $\leftarrow \lceil \text{TBE/Nrm} \rceil$.

In negotiating TBE, the switches have to consider their buffer availability, since a switch may receive TBE cells during the first round trip, and after long idle periods. For small buffers, TBE should be small and vice versa. On the other hand, TBE should also be large enough to prevent unnecessary triggering of rule 6 on long delay paths or with very high speeds. Thus, TBE is highly affected by the bandwidth-delay product of the connection, and should be set to: $PCR \times FRTT + \Sigma_i, i \in \{\text{switches on path}\}$, buffer sizes to account for the speed, link delays, and buffer sizes.

Effect of link speed and round trip time on TBE and CRM. For long-delay links, such as satellite links, our simulation results revealed that source rule 6 can unnecessarily trigger and cause oscillations during start up and after idle periods, unless TBE is large enough. This can degrade the throughput considerably. Figure 8 shows the configuration used to illustrate the problem. All the links are OC-3 links operating at a rate of 155.52 Mbps. The link connecting the two switches is a satellite link, while the links connecting the switches to the end systems are each 1 km long. The one-way propagation delay of the satellite link is 275 ms, while the propagation delay of each LAN link is 5 microseconds. The traffic is bidirectional, and the sources are persistent. The ERICA [7] algorithm is used with target utilization 90%. The ABR source parameter values are: PCR = 155.52 Mbps, MCR = 0 Mbps, ICR = 0.9×PCR = 140 Mbps, Nrm = 32, RIF = 1, CDF = 1/16, and CRM = 32, 256, 1024, 4096, 6144, 8192.

Figure 9 illustrates the performance of the system with CRM set to 32 (the default value before we proposed a change in August 1995). Figure 9(a) shows the allowed cell rate of the source over 1200 ms, and figure 9(b) shows the number of cells received at the destination during the same period of time [1]. As seen in figure 9(a), the initial rate is 140 Mbps (90% of 155 Mbps). After sending 32 RM cells (or CRM×Nrm = 32×32 = 1024 cells), rule 6 triggers and the rate rapidly drops. The first feedback is received from the network after around 550 ms (275 ms×2), because the one-way delay of the satellite link is 275 ms. The network asks the source to increase to 140 Mbps. The source increases its rate but rule 6 triggers again. This is because the period between returning RM cells is long (they were sent at a low rate). This phenomenon of increase and decrease repeats resulting in high-frequency oscillations between very low rates and very high rates. The rapid rate drops occur due to the triggering of source rule 6, while the rate increases occur because the network feedback is consistently at 140 Mbps (90% of 155 Mbps).

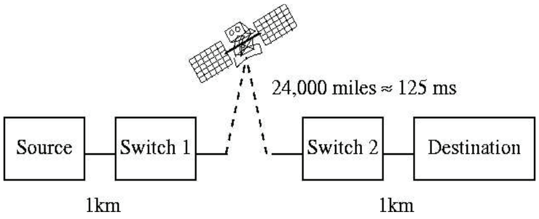
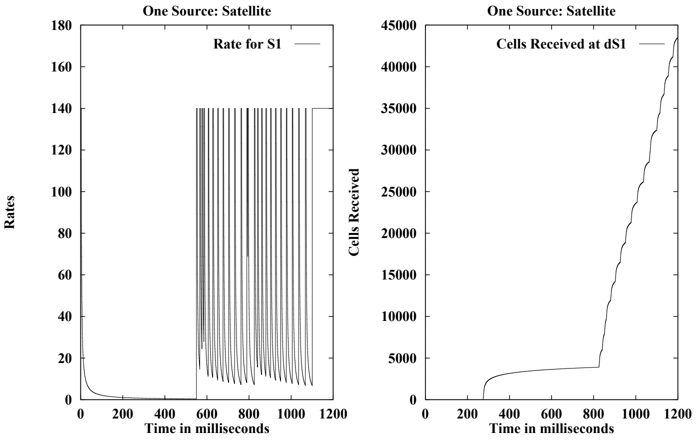


Fig. 8. Satellite configuration

Figure 9(b) shows the number of cells received at the destination. From this figure it is possible to compute instantaneous throughput denoted by the slope of the curve. It is also possible to compute average throughput over any interval by dividing the cells received (increase in the y-value) during that interval by the period of time (x-value) of the interval. The average throughput during the interval from 275 ms to 825 ms is 32 Mbps and that during the interval from 825 ms to 1200 ms is 45 Mbps. During the first 550 ms, the source is mostly sending at a very low rate until the first feedback is received after about 550 ms. The effect of the receipt of feedback can be observed at the destination after 550+275=825 ms. After the first feedback is received, the rate oscillations result in reduced throughput. The results do not significantly vary for different values of CDF. The low throughput values in figure 9(b) are a result of the unnecessary triggering of source rule 6 for small CRM values.



(a) Allowed Cell Rate

(b) Cells Received at Destination

Fig. 9. Simulation results for a one source configuration CRM = 32.

For full throughput, TBE must be set so that the number of cells in flight is large enough to fill the path both ways. This number is equal to the round trip time (FRTT) × PCR. The number of RM cells in flight (CRM) should be (1/Nrm) of this value: CRM $\geq \frac{\text{FRTT}\times\text{PCR}}{\text{Nrm}}$. For 155 Mbps links, CRM should be greater than or equal to 6144 (550 ms×365 cells per ms/32 cells). For 622 Mbps links, CRM $\geq$ 24576 (6144×4). For n 622 Mbps satellite hops, CRM $\geq$ 24576×n. Since the size of the TBE parameter is 24 bits and Nrm is normally 32, a 24-bit TBE allows a 19-bit CRM, which is sufficient in most situations.

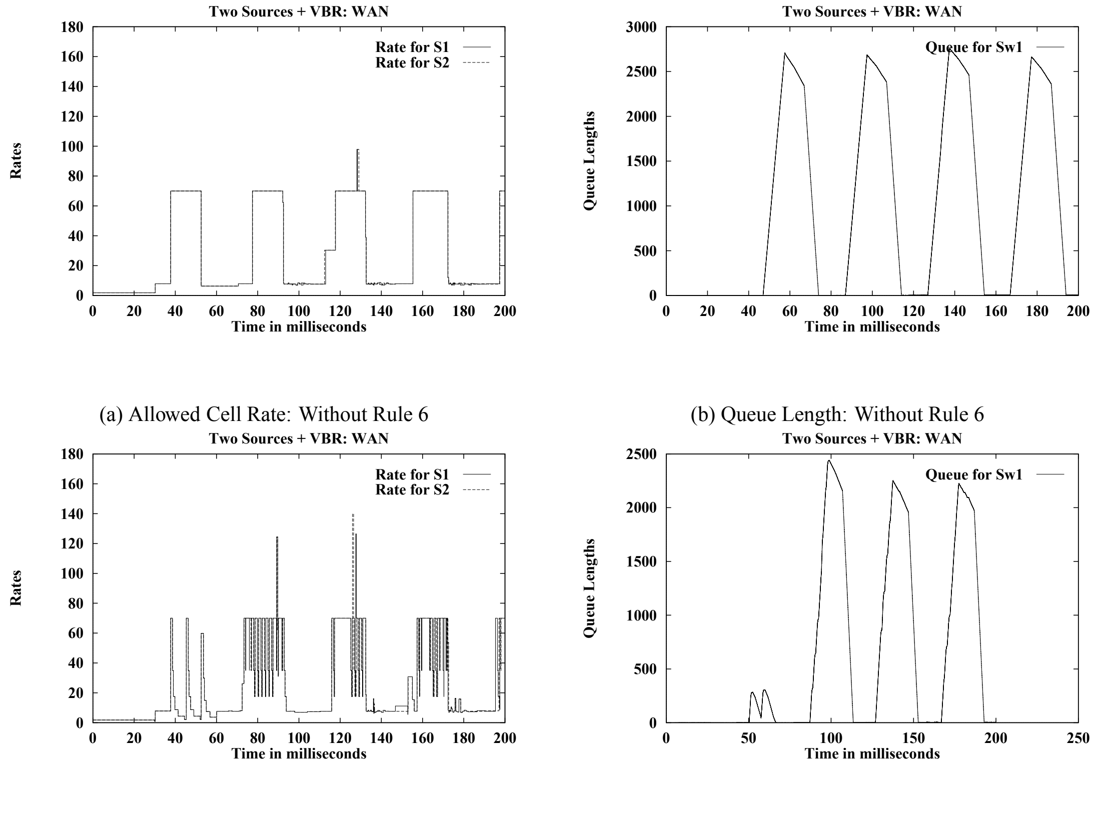
Effect of TBE on queue sizes. It had been incorrectly believed that cell loss could be avoided by simply negotiating a TBE value below the number of available buffers in the switches. We show in [4] that it is possible to construct workloads where queue sizes could be unreasonably high even when TBE is very small. TBE limits the queue length only during initial startup and after idle periods when there are no previous cells in the network from the same connection. In this case, the queue length can be given by the following equation: Queue length $=$ (number of sources $-$ 1) $\times$ $min(TBE,$ burst size$)$.

TBE cannot be relied upon during the closed-loop operation phase of a connection. During this latter phase, the contribution of a connection to the queue at a switch can be more than its TBE. The buffer usage at a switch can be more than the sum of TBEs allocated to active connections. In steady state, rule 6 rarely triggers and is overridden by subsequent explicit feedbacks. Since the reverse flow is not stopped completely, the forward flow continues and keeps filling the queues. TBE does not significantly affect the maximum queue length. Figure 10 depicts ACR and queue lengths for a network consisting of two ABR and one VBR sources going through two switches to corresponding destination. All simulation results use the ERICA switch algorithm [7] with 90% target utilization. All links are 155 Mbps and 1000 km long. All connections are bidirectional. The following parameter values are used: PCR = 155.52 Mbps, MCR = 0 Mbps, ICR = min{155.52, TBE/FRTT}, RIF = 1, Nrm = 32, RDF = 1/512, CRM = TBE/Nrm, Trm = 100 ms, FRTT = 30 ms, TBE = {128, 512, 1024}, CDF = {0, 0.5} = {Without rule 6, With Rule 6}. The VBR source generates a square waveform of 20 ms on and 20 ms off. During the on period, its amplitude is 80% of the link rate. The first VBR pulse starts at t=2 ms. to 42 ms and so on. The scheduler gives preference to VBR. Figure 10 shows the ABR ACRs and queue sizes for TBE of 128 cells. With just two sources, the queue length (without rule 6) is of the order of 2500 cells. The situation does not change significantly with rule 6. Rule six does trigger during initial start up, but is not triggered once the flow is set up (see ACRs). With TBE of 512 or 1024 cells (not shown), with or without rule 6, once again the queue length is around 3000 cells. This queue length is more than that with TBE of 128 but there is no simple relationship between TBE and queue length.

The reason for the inadequacy of rule 6 in limiting the queue growth can be explained as follows (figure 7). Assume that a certain source S is sending forward RM cells at an average rate of R cells per second (cps). The RM cells are turned around by the destination and the backward RM cells are received by S at a different rate r cps. In this case, the inter-forward-RM cell time at the source is $1/R$ while the inter-backward-RM cell time at the source is $1/r$. Source end system Rule 6 will trigger at S if the inter-backward-RM time is much larger (more than CRM times larger) than the inter-forward-RM time. That is, if: $1/r \geq$ CRM $\times$ $(1/R)$ or: $R \geq$ CRM $\times$ r. In the case of initial startup, r is zero and so after TBE cells, rule 6 triggers and protects the sources. Similarly, in the case of a bursty source, r is zero and rule 6 triggers after TBE cells. However, if the BRM flow is not totally stopped and $R < CRM \times r$, then the cells can accumulate in the network at the rate of $(R - r) \times Nrm$ and not trigger rule 6. In such cases, the queues can grow substantially. The maximum queue length is a function of PCR, the target utilization, and the VBR amplitude, multiplied by the feedback delay [4].

ICR. ICR should be set by the source as desired according to pricing and the application type. For TCP/IP applications and lower link speeds, ICR should be close to the peak cell rate (PCR). Switches should reduce their ICR to reflect their availability of buffers, as well as the bandwidth available for the connection. ICR is related to the availability of resources as computed during connection setup, and should correspond to the anticipated ACR for the connection at that time. Finally, the source takes the minimum of that ICR and $\frac{TBE}{FRTT}$ to correspond to the rate at which the source should initially send for the first round trip or after idle periods, before feedback is received. ICR thus depends on the bottleneck link speed and the round trip time.

CDF. When source rule 6 is triggered, the source reduces its rate by a factor of CDF, but not below the minimum cell rate. That is, ACR $\leftarrow$ max (MCR,ACR$-$ACR $\times$ CDF), where the value of CDF can be zero (for no rate decrease), or it can be a power of two that ranges from 1/64 to 1. This means that after CRM RM cells are sent (or CRM×Nrm total cells are sent), and no backward RM cell is received: ACR $=$ ACR$_{initial}$ $\times$ $(1 -$ CDF$)$ Note that if rule 6 is triggered once, it usually triggers on sending successive forward RM cells (as long as no backward RM cells are being received). Thus, after CRM$+k$ RM cells (or (CRM$+k$)×Nrm cells) are sent: ACR $=$ ACR$_{initial}$ $\times$ $(1 -$ CDF$)^{k+1}$ Such repeated rate reductions result in an exponential rate drop when source rule 6 triggers, as long as no feedback is being received. The smaller the CDF value, the more rapid the rate decrease. It may be desirable to disable source rule 6 (by



(a) Allowed Cell Rate: Without Rule 6

(b) Queue Length: Without Rule 6

(c) Allowed Cell Rate: With Rule 6

(d) Queue Length: With Rule 6

Fig. 10. Two Sources and VBR on a WAN, TBE = 128 cells

setting CDF to zero) if TBE cannot be set to a reasonable value, or if TBE must be set to a small value to decrease ICR. Disabling rule 6 in this manner, however, risks high cell loss in case of link failures or congestion collapse. CDF can be set to smaller values for high speeds and long RTTs to avoid big losses, according to the application type, confidenc in TBE value, confidenc in links, and availability of resources.

ADTF. As previously mentioned, the purpose of the ADTF timeout is to avoid the ACR retention problem that may cause congestion. ACR retention can cause sudden queue growth of: $(ACR -$ source rate) × feedback delay × (number of sources $-$ $1)$. Connections that disable rule 5 (e.g., by setting ICR=PCR) can be vulnerable to sudden arrivals. The default value of 500 ms was selected to correspond to the timer granularity used with most TCP/IP implementations using slow start. ADTF especially affects bursty traffic ADTF is independent of the bottleneck link speed of the connection since traffi is smoothed in the ATM network. $ADTF$ must be greater than RTT to prevent unnecessary rate reductions for long round trip times. Sources can set ADTF according to the application traffi characteristics (the expected burstiness of the traffic) Switches can reduce ADTF if they have limited resources.

VI. Out of Rate RM Cells

Although the tagged cell rate (TCR) is not signaled, we include a brief discussion on its role and settings.

Role. As stated in **source rule 11**, the out-of-rate FRM cells generated by sources are limited to to a rate below the tagged cell rate (TCR) parameter, which has a default value of 10 cps.

Values. Although higher TCR values improve transient response with zero or very low ACRs, since feedback is more frequent, increased TCR does increase the RM cell overhead in such cases. *Rescheduling* becomes important in cases where ACR is very low and the new ACR will allow cells to be scheduled earlier than their previously scheduled time [6]. There are no guidelines on how to space out-of-rate RM cells. TCR should depend on the bottleneck link speed, and perhaps a ratio, such as Nrm, should be used. 10 cps may be too low for very high speeds, e.g., 2.4 Gbps$+$. It is better to state that no more than$x$%, say 2.7×10^{-5}%, of the link bandwidth should be used for out-of-rate RM cells. The value 2.7×10^{-5}% is based on the intuition that 10 cps is a good value for OC-3 links (10 cps out of 365 cells per millisecond).

VII. Conclusions

Table III summarizes the discussion in this paper. For each of the parameters, the table indicates what the value the source end system sets for the parameter, how switches and destinations negotiate the parameter, how the parameter is affected by link speeds, and how it is affected by the round trip time of the connection.

References

[1] S. Fahmy, R. Jain, S. Kalyanaraman, R. Goyal, and F. Lu. On source rules for ABR service on ATM networks with satellite links. In *Proceedings of the First International Workshop on Satellite-based Information Systems*, pages 108–115, November 1996.

TABLE III

SUMMARY OF PARAMETER DEPENDENCIES AND RECOMMENDATIONS

Parameter	Speed?	RTT?	Source initializes according to	Switch/Dest. modifies according to
PCR	increases	no effect	link bandwidth or host/application capacity and pricing	bottleneck link bandwidth
MCR	increases	no effect	application requirements (e.g, video) and pricing	connection admission control (resources)
ICR	increases	source takes minimum of signaled ICR and $\frac{TBE}{FRTT}$	pricing, host capacity and application	buffering and resources, PCR and FRTT
Nrm	maybe should increase with speed	no effect	processing speed and application type (real-time should increase it)	switch scheme and switch speed
Trm	decreases	no effect	processing speed and application type	switches can reduce Trm for a high PCR, or increase it for low switch speed
RIF	no, but may be decreased	no, but may be decreased	application requirements	EFCI and RRM switches and ER switches sensitive to RIF should reduce it depending on FRTT, PCR and scheme
RDF	no, but may be decreased	no, but may be decreased	application requirements	EFCI and RRM switches should reduce dependent on FRTT, PCR and scheme
ADTF	no effect	no effect (may increase)	application traffic characteristics	if limited resources, reduce
TBE	increases	increases	application type, pricing and host capacity	buffering and resources, PCR and FRTT
CDF	may be smaller for high speeds	may be smaller for long RTTs	application type, confidence in TBE value	confidence in TBE and links, availability of resources

[2] The ATM Forum. The ATM forum traffic management specification version 4.0. ftp://ftp.atmforum.com/pub/approved-specs/af-tm-0056.000.ps, April 1996.

[3] R. Jain. Congestion control and traffic management in ATM networks: Recent advances and a survey. *Computer Networks and ISDN Systems*, 28(13):1723–1738, November 1996.

[4] R. Jain, S. Fahmy, S. Kalyanaraman, R. Goyal, F. Lu, and S. Srinidhi. More strawvote comments: TBE vs queue sizes. ATM Forum/95-1661, December 1995.

[5] R. Jain, S. Kalyanaraman, S. Fahmy, R. Goyal, and S. Kim. Source behavior for ATM ABR traffic management: An explanation. *IEEE Communications Magazine*, 34(11):50–57, November 1996.

[6] R. Jain, S. Kalyanaraman, S. Fahmy, and F. Lu. Out-of-rate RM cell issues and effect of Trm, TOF, and TCR. ATM Forum/95-0973R1, August 1995.

[7] S. Kalyanaraman, R. Jain, S. Fahmy, R. Goyal, and B. Vandalore. The ERICA Switch Algorithm for ABR Traffic Management in ATM Networks. *IEEE/ACM Transactions on Networking*, 8(1):87–98, February 2000.

[8] A. Koike, H. Kitazume, H. Saito, and M. Ishizuka. On end system behavior for explicit forward congestion indication of ABR service and its performance. *IEICE transactions on communications*, E79-B(4):605–610, April 1996.

[9] D. Lee, K. K. Ramakrishnan, W. M. Moh, and A. U. Shankar. Performance and correctness of the ATM ABR rate control scheme. In *Proceedings of the IEEE INFOCOM*, volume 2, pages 785–794, March 1997.

[10] K. Ramakrishnan and S. Floyd. A proposal to add explicit congestion notification (ECN) to IP. RFC 2481, 1999.

[11] B. Vandalore, S. Fahmy, R. Jain, R. Goyal, and M. Goyal. QoS and multipoint support for multimedia applications over the ATM ABR service. *IEEE Communications Magazine*, pages 53–57, January 1999.

AUTHOR BIOGRAPHIES

Sonia Fahmy is an assistant professor of Computer Sciences at Purdue University. She completed her PhD degree at the Ohio State University in 1999. Her work is published in over 30 papers and 40 ATM Forum contributions. She has served on the program committees of IEEE INFOCOM, ICNP and ICC, and chaired an SPIE conference.

Raj Jain is a co-founder and Chief Technology Officer of Nayna Networks– an optical systems company. He is currently on a leave of absence from Ohio State University, where he is a Professor of Computer & Info. Science. He is a fellow of IEEE and ACM, and is on the editorial boards of several journals. He was an ACM lecturer and an IEEE distinguished visitor. He is currently an IEEE distinguished lecturer. He received a Ph.D. degree in Computer Science from Harvard in 1978 and is author of “Art of Computer Systems Performance Analysis” and “FDDI Handbook.” He has 14 patents and over 90 publications. URL: http://www.cis.ohio-state.edu/∼jain/

Rohit Goyal received his M.S. and PhD. degrees in computer & info. science from the Ohio State University. He is currently with Axiowave Networks, Marlborough, MA. He has several journal, conference and standards publications, and is an active member of the ATM Forum and the IETF.

Bobby Vandalore received his MS and PhD in Computer & Info. Science from the Ohio State University in 1995 and 2000 respectively. He is currently a software engineer at Nokia, Mountain View, CA. He is the author of several papers and ATM Forum contributions.