Rate-Based Congestion Control for ATM Networks

Hiroyuki Ohsaki, Masayuki Murata, †Hiroshi Suzuki, †Chinatsu Ikeda and Hideo Miyahara

Department of Information and Computer Sciences
Faculty of Engineering Science, Osaka University
1-3 Machikaneyama, Toyonaka
Osaka 560, Japan

†C&C System Research Laboratories, NEC Corporation
4-1-1, Miyazaki, Miyamae
Kawasaki, Kanagawa 216, Japan

E-mail: oosaki@ics.es.osaka-u.ac.jp

Abstract

Congestion control plays an important role in the effective and stable operation of ATM networks. This paper first gives a historical overview of rate-based congestion control algorithms developed in the ATM Forum, showing how the current ATM Forum standard regarding the traffic management control methods is exploited, by these algorithms. Then, analytical approach is used to quantitatively evaluate their performance and show the effectiveness of the rate-based approach. In presenting the numerical examples, we emphasize that appropriate control parameter settings are essential for proper traffic management in an ATM network environment.

1 Introduction

The capability of ATM (Asynchronous Transfer Mode) networks to provide large bandwidth and to handle multiple quality of service (QOS) guarantees can only be realized by preparing effective traffic management mechanisms [1]. Traffic management includes congestion control, call admission control, and virtual path (VP)/virtual channel (VC) routing. Especially, essential for stable and efficient operation of ATM networks is congestion control, which is performed between ATM end systems. The ATM end system is a point where an ATM connection is terminated, and the connection goes up to the ATM adaptation layer (AAL). It is also defined as a point where VC connections are multiplexed, or demultiplexed, or both. Therefore, each ATM segment — a private local area network (LAN) or a public wide area network (WAN) — may, depending on its characteristics, adopt a different congestion control scheme.

Two congestion control strategies for ATM networks have been discussed: open-loop control and closed-loop control. With open-loop control, the limit on each connection's usable bandwidth is based on the notion of a *traffic contract* [2, 3]. To assign bandwidth to the connection, each end system has to declare its traffic parameters to the network before the connection is established. These traffic parameters are described, for example, in terms of such measures as peak cell rate, cell delay variation, sustainable cell rate, and burst length tolerance. Then, once the connection request is admitted, its QOS is guaranteed throughout the session. Because a lack of network resources may cause a newly requested connection to be rejected, open-loop control is sometimes referred to as preventive congestion control. This sort of congestion control scheme can be applied to real-time communications called constant-bit-rate (CBR) and variable-bit-rate (VBR) traffic, by which audio and motion video can be accommodated.

Open-loop control, however, is insufficient for data communications because each connection can never emit cells exceeding its negotiated rate, not even when there is unused bandwidth in the network [1]. Furthermore, the bandwidth requirements for data traffic are not likely to be known at connection setup time. Instead, the cell transmission rate can be adjusted if up-to-date information about the congestion status of the network is used appropriately. These are the reasons that closed-loop rate control is promising for data communications and is being applied to ABR (Available Bit Rate) service in the ATM Forum. Closed-loop control is sometimes called reactive congestion control, and it dynamically regulates the cell emission process of each connection by using feedback information from the network. It is therefore especially suitable for data transfer service.

For implementation of closed-loop control, two kinds of schemes have been proposed in the ATM Forum: rate-based and credit-based. The credit-based scheme is based on a link-by-link window flow control mechanism [4]. Independent flow controls are performed on each link for different connections, and each connection must obtain buffer reservations for its cell transmission on each link. This reservation is given in the form of a *credit balance*. A connection is allowed to continue cell transmission as long as it gains credit from the next node. When the connection is starved of credit, it should wait for credit. Owing to this link-by-link fast feedback mechanism, transient congestion can be relieved effectively. In addition, no cell loss occurs because no connection can send cells unless it has credit.

The rate-based scheme, on the other hand, controls the cell emission rate of each connection between end systems. It is simpler than credit-based flow control schemes in which each switch requires complicated queue management for every connection. Typical examples of the rate-based approach are forward explicit congestion notification (FECN) and backward explicit congestion notification (BECN) [5], which are well-known congestion control strategies in conventional packet-switched networks. One of the main purposes of this paper is to discuss the rate-based congestion control schemes proposed and developed so far in the ATM Forum. Implementation aspects of FECN-like and BECN-like methods in these control schemes are also described. We will also introduce more intelligent schemes in which the switch controls the rate of connections explicitly. We then present results of quantitative evaluation (provided through an analytical approach) showing the effectiveness of the rate-based congestion control method. The numerical examples here emphasize that the appropriate control parameter settings are essential for the rate-based congestion control to provide proper traffic management in an ATM network environment. Excellent discussions for traffic management methodologies in recent high speed network environments can be found in [6], and interested readers should refer to it.

The rest of this paper is organized as follows. In Section 2 we introduce several rate-based congestion control schemes proposed in the ATM Forum and qualitatively evaluate each of these schemes. In Section 3 we analyze the Enhanced Proportional Rate Control Algorithm (EPRCA), which is a basis of the rate-based control schemes in the ATM Forum, and use numerical examples to illustrate its performance. In Section 4 we offer some concluding remarks and mention some open issues. The control parameters used in the rate-based congestion control methods are summarized in Appendix A.

2 Rate-based Control Schemes

Several proposals had been contributed in the rate-based congestion control framework to the ATM Forum by the end of 1993: the methods based on FECN [7, 8] and the methods based on BECN [5, 9]. The Rate-Based Traffic Management Ad-Hoc working group was then established to discuss various aspects of rate-based congestion control methods. The result, which will be precisely described in the next subsection, was published as an ATM Forum Contribution [10]. The ATM Forum standard regarding traffic management specifies only the source and destination end systems behaviors; the methods for implementing the switches are left to the manufacturers. We will describe here how the behavior of end systems is standardized and how the various switches proposed in the ATM Forum can cooperate with the *standardized* end systems.

2.1 Interval-based Approach

In this subsection we explain the original rate-based scheme, which was proposed in [10, 11]. Figure 1 illustrates a basic configuration of the rate-based congestion control scheme in which the ATM connection is terminated at the source and destination end systems. A permitted cell transmission rate ACR (Allowed Cell Rate) of the source end system is changed according to the congestion status of the network. An initial rate ICR, a maximum allowable rate PCR, and a minimum cell rate MCR are specified by the network at connection setup time, and the source is then allowed to emit cells at a rate that ranges from 0 to ACR. When this scheme is compared with later proposals described in the following subsections, a distinctive point of the original scheme is that the operation of both end systems is based on interval timers. And the polarity of the feedback information from the network is *negative*; that is, the source end system receives feedback information only when the network falls into congestion.

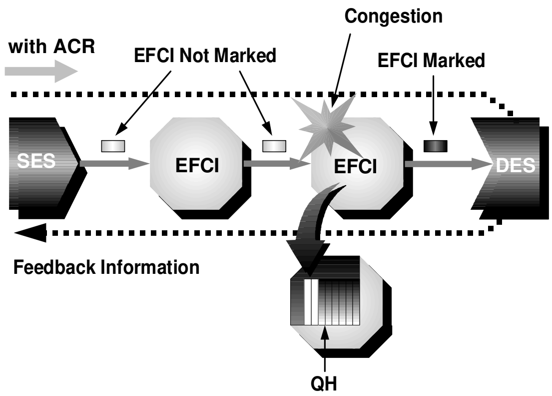

Figure 1: Basic Configuration of the Rate-Based Congestion Control.

An occurrence of congestion is detected at each intermediate switch by monitoring the queue length of the cell buffer. When the queue length exceeds a threshold value (Q_H), congestion is signaled to the source by a special cell called an RM (resource management) cell, whose Payload Type Identifier (PTI) is "110". The FECN-like signaling mechanism is defined in this scheme. That is, each switch signals its congestion information to its downstream switches by setting an EFCI (Explicit Forward Congestion Indication) bit in the header of passing data cells. When the destination end system receives a data cell in which the EFCI bit is marked, it sends an RM cell back to the source along the backward path. Then the source end system must decrease its ACR, multiplicatively according to this feedback information, as

$$ACR \quad \leftarrow \quad max(ACR \times MDF, MCR), \quad (1)$$

where MDF is the multiplicative decrease factor and MCR is the minimum cell rate for the ACR. A time interval RMI (RM Interval) is defined at the destination end system, and only one RM cell is allowed to be sent in an RMI. The source end system is also provided with an interval timer UI (Update Interval). When the timer expires without an RM cell having been received, the source recognizes no congestion in the network. Then it increases ACR additively as

$$ACR \quad \leftarrow \quad min(ACR + AIR, PCR),$$

where AIR is the additive increase rate and PCR is the peak cell rate of the connection.

As an implementation option, the network can be divided into two or more segments by introducing *intermediate* networks that should act as a *virtual* destination end system for the source and as a virtual source end system for the destination (Fig. 2). As a destination end system, an intermediate network has to send RM cells back to the source according to the EFCI status of incoming cells. As a source end system, it is also required to regulate the flow of cells destined for the destination end system.

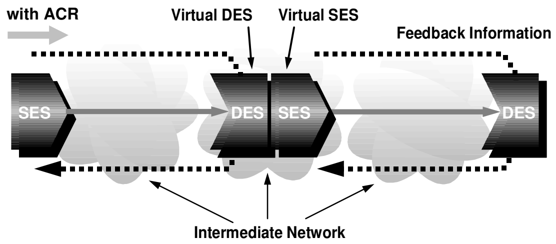

Figure 2: Network Segmentation by an Intermediate Network.

Since this approach requires interval timers at both end systems, it increases the complexity of implementation and could become expensive. And as pointed out in [12], the negative feedback mechanism could cause a collapse of the network in certain conditions. If the network is heavily congested, RM cells can be delayed or lost because of buffer overflow, with the result that timeliness of the congestion information is lost or — in a more serious situation — the source increases its cell emission rate because of the absence of RM cells.

2.2 Counter-based Approach

The timer-based approach described in the previous subsection was revised because of its drawbacks, and [13] proposed a proportional rate control algorithm (PRCA) with two major modifications: (1) the polarity of the feedback information is *positive*, and (2) the need for interval timers is eliminated. The origin of the name PRCA is in that opportunities for rate increases are given in proportion to the current sending rate ACR. In PRCA the source end system marks the EFCI bit in all data cells except for the first of every N_{RM} cells. The destination end system instantly sends an RM cell back to the source when it receives a cell with the EFCI bit cleared. If the EFCI bit is set by an intermediate switch because of its congestion, the destination takes no action. By this mechanism, a positive congestion signaling is established: receiving the RM cell implies that there is no congestion in the network, and therefore the source end system is given an opportunity to increase its rate.

The source end system sends the cells in the following way. Unless receiving an RM cell, the source determines the next cell transmission time at $1/ACR$ after the current time. This implies that the source continuously decreases its ACR (until receiving an RM cell) as

$$ACR \leftarrow \max(ACR - ADR, MCR). \qquad (2)$$

When the source receives an RM cell, the rate is increased as

$$ACR \leftarrow min(ACR + N_{RM}\, AIR + N_{RM}\, ADR, PCR),$$

which compensates the reduced rate since the source received the previous RM cell ($N_{RM}\, ADR$) and increases the rate by $N_{RM}\, AIR$. In an ideal situation with no propagation delay, this should give a linear increase of the cell transmission rate. The network can thus be restored even if heavy congestion results in all RM cells being discarded. This is because the rate is always decreased whenever the source does not receive an RM cell. The above operation provides FECN-like congestion management, but if switches have the ability to discard RM cells in the backward direction, BECN-like operation can also be achieved. This is one of the notable features of PRCA (Fig. 3).

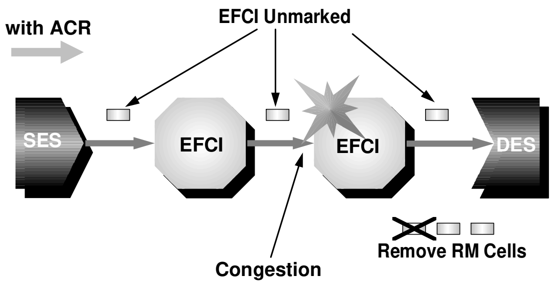

Figure 3: BECN-like Congestion Notification in PRCA.

Certain problems, however, remain even in PRCA and have been pointed out in [14]. One of them is referred to as an "ACR beat down" problem, and is explained as follows. Each source of active connections that experiences congestion in several switches has less opportunity to receive positive feedback than do sources of other connections with fewer switches. Once one of those relatively feedback-starved sources decreases its transmission rate to the minimum rate MCR, it is in some circumstances likely to remain at that rate indefinitely. Thus, fairness among connections cannot be achieved. Another problem is that PRCA requires a considerable amount of buffers when there is a large number of active connections. It is now widely recognized that when the propagation delays are large (as in the WAN environment), the queue length temporarily grows because of the control information delays.

This is an intrinsic and unavoidable feature of recent high-speed networks. The problem is, however, that such a long queue length occurs even in a LAN environment. During congestion, the rate is decreased by ADR (Eq. (2)). When the number of connections is very large, however, decreasing the aggregated input rate at the switch is too slow. For example, when the control parameters suggested in [13] are used, 241,000 cell buffers are required for assuring no cell loss even in the LAN environment for 1000 active connections [14]. Such a large buffer size is unacceptable given the current memory technology.

2.3 Enhanced PRCA Method

An improved version of PRCA — called EPRCA (Enhanced Proportional Rate Control Algorithm) — was then proposed in [15, 16]. New functionalities are added as implementation options in two ways. One, to achieve better fairness among connections, is a capability to send a congestion indication to particular sources rather than all sources. The fairness could be achieved if each connection is maintained separately at the switch, which is called *per-VC accounting* (see Subsection 2.4.2 for more detail). However, since it requires an additional control complexity, EPRCA adopts another method "intelligent marking", which is originated from the work in [17]. The other is the means for reducing the rate of each connection explicitly; that is, the switch can have a responsibility for determining the cell transmission rate of selected connections. While some modifications were required in order to incorporate these new features, EPRCA preserves a backward compatibility with PRCA. A switch supporting only PRCA can thus also be used in an EPRCA-based network. For distinguishing this switch from other new switches, it is called an EFCI bit setting switch.

EPRCA requires forward RM cells as well as backward RM cells. RM cells contain a CI (Congestion Indication) bit that is used to carry congestion information to the source. Instead of unmarking an EFCI bit of data cells as PRCA does, the source end system periodically sends a forward RM cell every N_{RM} data cells. When the destination end system receives the forward RM cell, it returns the RM cell to the source as a backward RM cell. When doing this, the destination end system sets the CI bit of the backward RM cell according to the EFCI status of the last incoming data cell. The source end system can thus be notified of the congestion detected at the intermediate switches by marking the EFCI bit of data cells in the forward path. This is a FECN-like implementation of congestion notification. Furthermore, the switch can be allowed to set a CI bit of backward RM cells as a BECN-like implementation.

The two major enhancements of EPRCA — *intelligent marking* and *explicit rate setting* — require additional information fields in each RM cell: CCR (Current Cell Rate) and ER (Explicit Rate) fields. An ER element is used to decrease the source rate explicitly, and is initially set to PCR by the source. One or more congested intermediate switches can change it to a lower value so that the rate of the source end system is rapidly decreased for quick congestion relief. The CCR element is set to the current ACR of the source in effect, and a fair distribution of the bandwidth can be achieved with these two values. The intermediate switch selectively signals congestion indication to the sources with larger ACR values. A more impartial sharing of the available bandwidth can be achieved by using this intelligent marking mechanism in conjunction with the explicit rate setting mechanism.

Three types of switch architectures with different functions are suggested in the form of pseudo-code in [16]; EFCI bit setting switches (EFCI), Binary Enhanced Switches (BES), and Explicit Down Switches (EDS). EFCI switches, which are already on the market, are same as switches supporting PRCA and are expected to be the least expensive.

In BES switches, two threshold values for indicating congestion are defined: QT and DQT (on behalf of Q_H in the EFCI switch). When a BES switch is congested (i.e., when the queue length in the cell buffer of a BES switch exceeds QT), the switch performs intelligent marking. It selectively reduces the rate of sources with larger ACR (Fig. 4), by which the ACR beat down problem can be avoided. For implementing this mechanism, the switch maintains a control parameter $MACR$ (Mean ACR) that should ideally be the mean of the ACR's of all active connections. When the rate of all connections is equal to $MACR$, the bandwidth is shared equally and the switch can be fully used without falling into congestion. The key to this is obtaining an accurate $MACR$. The BES switch updates its $MACR$ according to the CCR field of forward RM cells. For example, $MACR$ is calculated as

$$MACR \quad \leftarrow \quad MACR(1 - AV) + CCR \times AV,$$

where AV is used as an averaging factor [16]. When the switch becomes congested, it indicates its congestion to the sources having higher rates. More specifically, the switch marks the CI bit of the backward RM cells if its CCR value exceeds $MACR \times DPF$ (Down Pressure Factor), where a typical value of DPR is 7/8 for safe operation. The switch may remain congested, however, if only intelligent marking is used. Therefore when a BES switch becomes *very* congested (such that the queue length exceeds DQT), all backward RM cells are marked irrespective of their CCR values. Note that it is evident from the above description that a BECN-like quick congestion notification can be accomplished in BES switches.

The EDS switches are provided with an *explicit rate setting* capability in addition to intelligent marking (Fig. 5). These maintain $MACR$ as BES switches do, and they control the transmission rate of sources by setting the ER field of backward RM cells according to a degree of congestion. When a backward RM cell

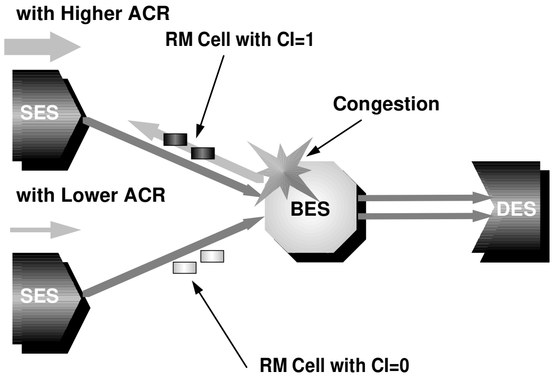

Figure 4: Intelligent Marking in a BES Switch.

with CCR larger than $MACR$ passes through the congested EDS switch, the value of its ER element is set to $MACR \times ERF$ (Explicit Reduction Factor). If the switch becomes very congested, $MACR \times MRF$ (Major Reduction Factor) is set in all backward RM cells to achieve quick congestion relief, which is called *major reduction*. Typical values of ERF and MRF shown in [16] are 7/8 and 1/4. A quantitative evaluation of these three types of switches will be presented in the next section.

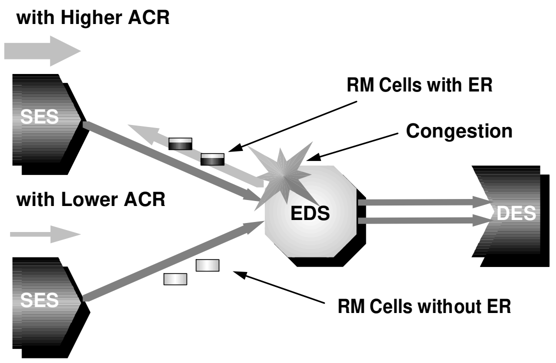

Figure 5: Intelligent Marking and Explicit Rate Setting in an EDS Switch.

Not only can three types of switches (EFCI, BES and EDS switches) coexist in EPRCA, but the operation of EFCI switches can also be enhanced by locating BES or EDS switches downstream. BES and EDS switches interpret the EFCI bit of forward data cells. When a BES or EDS switch is congested, entries in the VC table are marked according to the EFCI status of forward RM cells, and the EFCI bit is cleared. Then the CI bit of backward RM cells is set to notify the source of the switch's congestion if its associated entry in the VC table is marked. This mechanism enables a BECN-like quick congestion notification even if there are EFCI switches in the network: the enhanced switches can behave as virtual destination end systems for the EFCI switch.

The behavior of the source end system is simplified in [18], which will be included as an example source code in the standard. In that proposal, the source end system decreases its ACR by ACR/RDF every N_{RM} cells sent until reaching MCR. While this modification on EPRCA would degrade its performance to some extent, the complexity at the source end system can be decreased considerably.

2.4 Recent Proposals for Enhancement of EPRCA

This subsection introduces several proposals that enhance the switch capabilities. As mentioned in the introduction to Section 2, since the standard does not specify the switch's behavior, the methods shown in this subsection will not be reflected in the standard. We think, however, that these methods will help us understand how the current EPRCA can be improved for more effective congestion control.

2.4.1 Adaptive Proportional Rate Control

Although EPRCA shares resources more fairly than PRCA does and uses link bandwidth more efficiently, it still has a fault in that fairness is not assured in some configurations [19]. When a BES or EDS switch is very congested, it forces all connections to decrease their rates equally but not selectively. Thus, when the switch is very congested for a long time, *intelligent marking* does not work well. While it has been suggested that this problem can be avoided by eliminating this operation in *very congested* states, this results in excessive queue length rather than fairness [19].

Adaptive Proportional Rate Control (APRC), which is an originator of *intelligent marking* [17], is modified to solve this problem in [20]. Congestion in the switch is detected by evaluating the change of queue length in a fixed time interval rather than by comparing queue length with a threshold value. If the queue length increases in N cell times, the switch is expected to fall into congestion. Then each connection that has a higher rate than $MACR$ is adjusted to a lower rate by setting $MACR$ to the ER field of the backward RM cells. When the number of cells in the buffer exceeds DQT, the switch selectively sets $MACR \times DPF$ to ER only for connections with higher rates. This modification improves the responsiveness to congestion and therefore can reduce the maximum queue length. It also improves the fairness among connections.

This *intelligent marking* capability of APRC was incorporated into EPRCA and a new scheme called APRC2 was introduced in [21]. In EPRCA, the CCR value in the forward RM cell is used to compute $MACR$. At one switch, however, the effective rate of some connections that experience congestion at another switch may be quite different from the CCR values contained in the RM cells. This leads to misbehavior of the explicit rate control. This problem can be avoided by introducing UCR value at the switch in order to establish stable operation [21], where UCR is defined as a mean of CCR's only for connections with larger CCR than

$MACR$. The value of UCR is updated as

$$UCR \leftarrow UCR + a(CCR - UCR)$$

only when CCR is greater than $MACR$. Thus UCR is used to determine the explicit rate ER effectively.

The operation of the source end system and switches can also be improved by shortening *ramp-up* time, which is the time between when a connection begins/resumes its transmission and when the network settles into steady state. The ramp-up time is of importance because (1) since each connection starts its transmission with rate ICR regardless of the network status, it might cause a large queue buildup or underutilization of the link unless ICR is set properly, and (2) most networks operate in a transient state since many connections are established but idle because of the bursty nature of the ABR traffic. Simulation experiments in [22] show that APRC2 results in better ramp-up time, link utilization, and maximum queue length than do EPRCA and APRC. Excellent arguments on limitations of all existing rate control schemes can be found in [21]. Refer to it for understanding various trade offs: "intelligent marking vs. nonselective binary marking" and "counter-based vs. interval-based".

2.4.2 EPRCA+ and EPRCA++ Methods

Implementing *explicit rate setting* requires the number of active connections to be known by the switch. One way of assuring this is *per-VC accounting*, which can be implemented in several ways with additional hardware complexity. For instance, each switch can have a VC table to record the number of active connections. Each VC entry is marked or unmarked according to the status of its corresponding VC (active or inactive), and the number of marked entries represents the number of active connections. In this way, the rates of all sources are adjusted through RM cells in one round-trip time when there is one congested switch in the network. EPRCA+, proposed in [23], also uses this kind of scheme, and its simplified version can be found in [24].

In EPRCA+, congestion is detected by estimating the traffic load at the switch rather than by using a threshold value in the cell buffer. For this, the switch is provided with an interval timer and counts the number of cells received during a fixed time interval. The source end system is also equipped more expensively with an interval timer instead of a counter for sending RM cells. The rate of the source is kept unchanged until it receives a backward RM cell in which the explicit rate ER determined by the switch is contained.

One attractive feature of EPRCA+ is its small number of control parameters, which can be set easily by a network manager. Many control parameters required in EPRCA (see Appendix A) are eliminated in EPRCA+. Furthermore, in EPRCA+ the target utilization band (TUB) around which the switch is utilized, can be set freely. One may set the TUB of the switch under 95% link utilization, and then the queue size at the switch is smaller and cell delays are shorter. Although there is an additional expense for timers and the VC table, EPRCA+ can provide better fairness and responsiveness than EPRCA can [23].

Redundant complexities in the latest EPRCA are pointed out in [25]. For example, in the current proposal an active source end system should decrease its rate by ACR/RDF every N_{RM} cell sent until reaching MCR. Its necessity is, however, not well justified, and it might be unnecessary in stable environments. A new scheme called EPRCA++ proposed in [26] uses a counter at the source end system for forward RM cells instead of a timer as in EPRCA+. Furthermore, the source end system decreases its ACR only if no backward RM cell is received in $k \times N_{RM}$ cell times (where k is set to a rather large value). These modifications enable EPRCA++ to perform better than EPRCA+, especially in transient state.

2.5 Fairness Definitions

Rate-based congestion control schemes have been improved to obtain better fairness among connections as well as to increase link utilization and to reduce buffer size at the switch, but a clear definition for the fairness measure is needed for evaluating the degree of fairness. Since there is no definition of the fairness measure in the literature, the ATM forum is also discussing fairness criteria. The fairness for the ABR service is defined by a max-min criterion in [27]. In the max-min criterion, all active connections are served *fairly* if the following two conditions are met [28]: (1) each connection must pass through at least one bottlenecked switch, and (2) the available bandwidth should be shared fairly when it is assigned to connections that do not pass through non-bottlenecked switch. However, the max-min criterion becomes inadequate for ABR traffic where MCR, the minimum usable rate for each connection, is incorporated. In what follows, we introduce the four definitions of fairness measures proposed in [29, 30].

[1] MCR plus equal share

In this definition, each active source is first allocated its MCR and then the rest of the available bandwidth is assigned fairly according to the max-min criterion. That is, the nth active connection's rate B_n is given by

$$B_n = MCR_n + \frac{C - \sum_{i=1}^{N_{VC}} MCR_i}{N_{VC}}, \quad 1 \leq n \leq N_{VC}, \tag{3}$$

where C is the available bandwidth and N_{VC} is the number of active connections at the switch, and MCR_n is the MCR of the nth connection. It is known that this definition can be applied not only to enhanced switches like BES or EDS but also to EFCI switches [29].

[2] Maximum of MCR or Max-Min share

In this definition, each connection acquires larger bandwidth of MCR and the bandwidth equally divided by all connections:

$$B_n = max(\frac{C}{N_{VC}}, MACR_n), \quad (4)$$
$$1 \le n \le N_{VC}.$$

This assignment needs an iteration for the sum of B_n's to be settled at the available bandwidth C, and the required number of iterations cannot be estimated. For connections with larger MCR, however, more bandwidth can be allocated than in the previous case.

[3] ALLOCATION PROPORTIONAL TO MCR
The third definition assigns the available bandwidth to unconstrained connections in a weighted manner as

$$B_n = C\frac{MCR_n}{\sum_{i=1}^{N_{VC}} MCR_i}, \quad (5)$$
$$1 \le n \le N_{VC}.$$

This definition cannot be applied to connections with $MCR = 0$. The bandwidth allocated to each connection is proportional to its MCR.

[4] WEIGHTED ALLOCATION
The fourth definition is a hybrid of the first and third:

$$B_n = MCR_n + F_n(C - \sum_{i=1}^{N_{VC}} MCR_i), \quad (6)$$
$$1 \le n \le N_{VC},$$

where F_n is a weight for the nth connection and is defined as

$$F_n = \frac{b}{N_{VC}} + (1-b)\frac{MCR_n}{\sum_{i=1}^{N_{VC}} MCR_i},$$
$$0 \le b \le 1.$$

Note that this becomes identical to the first and third definitions with $b = 0$ and $b = 1$, respectively.

An appropriate definition of fairness depends on the environment in which it is used. Requirements for charges, for example, may become an important factor in deciding which one is appropriate. In the public WAN environment, the third definition is suitable if the tariff is determined in proportion to MCR. In private networks, on the other hand, the first definition might be more suitable because its implementation is less expensive. In the above definitions, however, the different values of PCR_n for connections are not taken into account. Since various speeds of interfaces are currently defined, PCR may also have to be incorporated.

3 Analysis of Enhanced PRCA

In this section, we present the analytical results of EPRCA with comparison among three types of switches suggested — EFCI, BES and EDS switches described in Subsection 2.3. Due to lack of space, the analysis for only EFCI switches is presented in this paper, and refer to [31, 32, 33] for more precise analysis and additional numerical examples.

3.1 Analytic Model

Our model is rather simple and consists of homogeneous traffic sources and a single bottleneck link (see Fig. 6). The number of active connections that share the bottleneck switch is denoted by N_{VC}. We assume that these connections behave identically; that is, they all have the same parameters ICR, PCR, AIR, and MDF (for definitions of these parameters, see Appendix A). The bandwidth of the bottleneck link (in cells/msec) is denoted by BW, and the propagation delays from the source to the switch and from the switch to the destination are denoted τ_{sx} and τ_{xd}, respectively. These parameters τ_{sx} and τ_{xd} are given according to the network configuration (i.e., LAN or WAN). The round-trip propagation delay from the source to the destination is denoted τ, and the following relation holds: $\tau = 2(\tau_{sx} + \tau_{xd})$. We further introduce $\tau_{xds}(= 2\tau_{xd} + \tau_{sx})$, which is the propagation delay of congestion indication from the switch to the source end system via the destination end system.

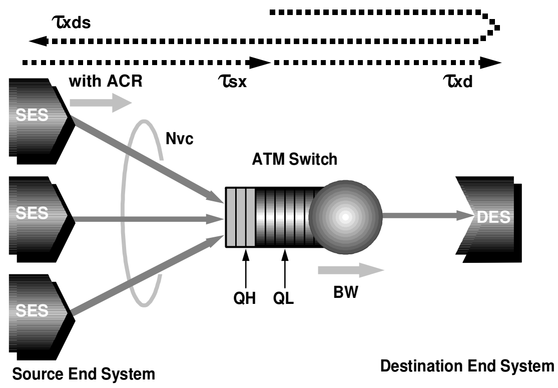

Figure 6: Analytic Model.

In all three types of switches, congestion is detected by threshold values associated with the queue length at switch buffer. The EFCI switch has high and low threshold values denoted as Q_H and Q_L. When the queue length at the switch exceeds Q_H, the switch detects its congestion and marks the EFCI bit in the header of data cells. When the queue length goes under Q_L, on the other hand, it is regarded that congestion terminates.

In EPRCA, the source end system sends RM cells proportionally to its rate. Furthermore, the rate is changed in response to backward RM cells returned from the destination end system, and the rate at which the backward RM cells are received is bounded by $BW/(N_{VC}\, N_{RM})$ when the switch is congested. Otherwise, it is identical to the transmission rate of the source end system. We, therefore, require an analytical treatment different from the one presented in [7], where the rate change is performed on a timer basis.

Let us introduce $ACR(t)$ and $Q(t)$, which respectively represent the cell transmission rate ACR of each source end system and the queue length at the switch observed at time t. In what follows, evolutions of $ACR(t)$

and $Q(t)$ in steady state are analyzed assuming that (1) the switch has infinite capacity of the buffer, and that (2) the source end system always has cells to be sent. Therefore, $ACR(t)$ becomes equivalent to CCR (Current Cell Rate), which is the actual cell transmission rate.

3.2 EFCI Switch

In this subsection, we analyze a dynamical behavior of $ACR(t)$ and $Q(t)$ by focusing on the EFCI switch. Forward RM cells are not considered explicitly in this analysis, so model is equivalent to the PRCA described in Subsection 2.2. Forward RM cells can easily be taken into account, however, by replacing BW in our analysis with BW', which is defined as

$$BW' = BW \frac{N_{RM}}{N_{RM}+1}. \quad (7)$$

In numerical examples, BW' will be used in comparisons with switches that require forward RM cells.

3.2.1 Determination of $ACR(t)$

Figure 7 shows a pictorial view of $ACR(t)$ and $Q(t)$ which have a steady-state periodicity. The initial point of one cycle is defined at the time when the congestion indication is received at the source end system. In the EFCI switch, it takes a time τ_{xds} for the congestion indication to reach the source end system after the queue length at the switch becomes Q_H. We divide one cycle into four phases according to behaviors of $ACR(t)$ and $Q(t)$ depicted in Fig. 7. For simplicity of presentation, we introduce $ACR_i(t)$ and $Q_i(t)$ as

$$\begin{aligned} ACR_i(t) &= ACR(t - t_{i-1}), \quad 0 \le t < t_i \\ Q_i(t) &= Q(t - t_{i-1}), \quad 0 \le t < t_i, \end{aligned}$$

where t_i is defined as the time when phase i terminates. The length of phase i is represented by

$$t_{i-1,i} = t_i - t_{i-1}.$$

We note here that a more strict treatment is required for representing system behaviors dependent on system parameters. This treatment will be presented in the next subsection. Here, however, we assume that the system behaves as in Fig. 7. The initial allowed cell rate $ACR_i(t)$ $(1 \le i \le 4)$ can be obtained as follows [31, 32].

Phase 1: $ACR_1(t)$

$$ACR_1(t) = ACR_1(0) e^{-\frac{ACR_1(0)}{RD} t}. \quad (8)$$

Note that MCR is not considered in this above equation ($MCR = 0$ is assumed).

Phase 2: $ACR_2(t)$

$$ACR_2(t) = \frac{a_1 e^{-a_1 t} + a_2 r e^{-a_2 t}}{c_1 (e^{-a_1 t} + r e^{-a_2 t})}, \quad (9)$$

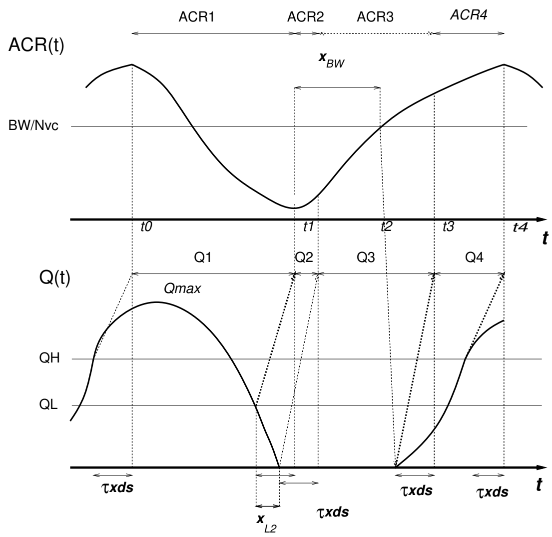

Figure 7: Pictorial View of $ACR(t)$ and $Q(t)$ in an EFCI Switch.

where a_1 and a_2 are roots of the equation

$$a^2 + c_2 a + c_1 c_3 = 0,$$

and c_1, c_2, and c_3 are given by

$$c_1 = -\frac{1}{RD}; c_2 = \frac{BW}{RD\, N_{VC}}; c_3 = \frac{BW\, AIR}{N_{VC}}.$$

The initial transmission rate $ACR_2(0)$ determines r in Eq. (9) as

$$r = \frac{a_1 - c_1\, ACR_2(0)}{c_1\, ACR_2(0) - a_2}.$$

Phase 3: $ACR_3(t)$

$$ACR_3(t) \cong ACR_3(0) e^{\beta t}, \quad (10)$$

where β is a root of the equation

$$\beta = \frac{N_{RM}\, AIR}{RD\, \log(\frac{RD}{RD - N_{RM}})} e^{-\tau \beta}.$$

Phase 4: $ACR_4(t)$

$ACR_4(t)$ is obtained in the same way as $ACR_2(t)$ since the rate at which the source end system recives RM cells is the same as in Phase 2.

3.2.2 Evolution of $ACR(t)$ and $Q(t)$

In what follows, we present the evolution of $ACR(t)$ and $Q(t)$. To do this, we should first determine the initial values $ACR_i(0)$ and the length of each phase $t_{i,i+1}$. Given the initial rates in Phase 1, $Q_1(t)$ is obtained as

$$\begin{aligned} Q_1(t) &= Q_1(\tau_{xds}) \\ &+ \int_{x=\tau_{xds}}^{t} (N_{VC}\, ACR_1(x - \tau_{sx}) - BW) dx. \quad (11) \end{aligned}$$

The length of Phase 1 $t_{12}(=t_1)$ is given as

$$t_{12} = x_{L1} + \tau_{xds},$$

where x_{L1} is obtained by solving the equation $Q_1(x_{L1}) = Q_L$. In what follows we will use a convention $x_{L1} = Q_1^{-1}(Q_L)$ for brevity.

We need a careful treatment For Phase 2 and later phases because some of these phases seem not to depend on the parameters. First, let us introduce x_{L2} as

$$x_{L2} = Q_2^{-1}(0) - Q_2^{-1}(Q_L).$$

That is, x_{L2} is the time for the queue length to reach 0 after it goes below Q_L. Furthermore,

$$x_{BW} = ACR_2^{-1}(\frac{BW}{N_{VC}}) - t_1,$$

which defines the time when the aggregate ACR reaches BW. We should consider the following four cases depending on x_{L2}, x_{BW}, and τ_{sx}. Case 1 in the below.

Case 1: $x_{L2} \le \tau$, $x_{L2} < x_{BW} + \tau_{sx}$
Case 2: $x_{L2} \le \tau$, $x_{L2} \ge x_{BW} + \tau_{sx}$
Case 3: $x_{L2} > \tau$, $x_{L2} < x_{BW} + \tau_{sx}$
Case 4: $x_{L2} > \tau$, $x_{L2} \ge x_{BW} + \tau_{sx}$

Due to lack of space, only Case 1 and Case 4 are explained. In Case 1 (Fig. 7), we have

$$\begin{aligned} t_2 &= t_1 + x_{L2} \\ Q(t) &= 0, \quad t_2 + \tau_{sx} < t \le t_3 + \tau_{sx}. \end{aligned}$$

In this equation, t_3 is given by

$$t_3 = t_2 + x'_{BW} + \tau,$$

where x'_{BW} is the time when the aggregate ACR reaches BW. That is,

$$x'_{BW} = ACR_3^{-1}(\frac{BW}{N_{VC}}) - t_2.$$

Furthermore,

$$Q(t) = \int_{t_2+\tau_{sx}+x'_{BW}}^{t} (N_{VC}\, ACR_3(x-\tau_{sx}) - BW)dx,$$
$$t_2 + \tau_{sx} + x'_{BW} < t \le t_3 + \tau_{sx}$$

Finally, we obtain equations for Phase 4 as

$$\begin{aligned} t_4 &= t_3 + t_{H4} + \tau_{xds} \\ Q(t) &= \int_{t_3+\tau_{sx}}^{t} (N_{VC}\, ACR_4(x-\tau_{sx}) - BW)dx, \\ & \quad t_3 + \tau_{sx} < t \le t_4 + \tau_{sx}, \end{aligned}$$

where t_{H4} is given by

$$t_{H4} = Q_4^{-1}(Q_H).$$

In Case 4, the queue length never reaches 0, so the switch is always fully utilized. Since neither Phase 2 nor 3 appears in this case, we have

$$t_4 = Q_4^{-1}(Q_H) + \tau_{xds}.$$

$Q_4(t)$ is then obtained from Eq. (11).

Finally, by setting

$$\begin{aligned} ACR_1(0) &= ACR_4(t_4) \\ Q_1(0) &= Q_4(t_4 + \tau_{xds}), \end{aligned}$$

and calculating these equations iteratively, we can obtain the dynamical behavior of EPRCA with the EFCI switch in steady state.

3.2.3 NUMERICAL EXAMPLES

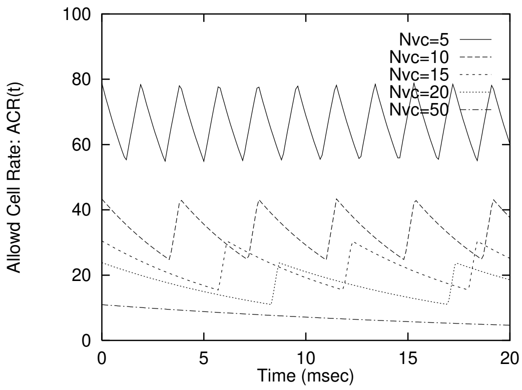

Figure 8: Effect of N_{VC} on $ACR(t)$ in an EFCI Switch ($\tau = 0.02$ msec).

Threshold values Q_H and Q_L are both set to 500, and the bandwidth of bottleneck link is set (assuming a 150 Mbits/sec ATM link) to 353.208 cells/msec. For other control parameters, the values suggested in [16] (also see Appendix A) are used throughout this paper unless other values are specified. Equation (7) is used to obtain BW; that is, the overhead for forward RM cells is taken into account for comparison with other switches.

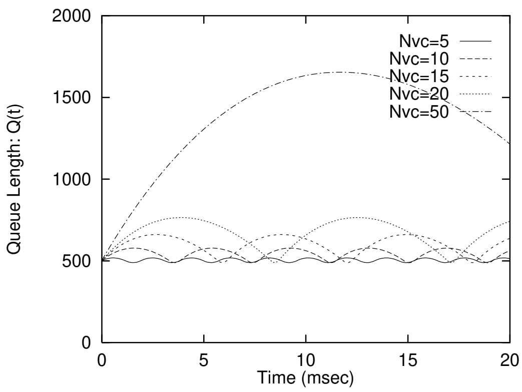

Figure 9: Effect of N_{VC} on $Q(t)$ in an EFCI Switch ($\tau = 0.02$ msec).

Figures 8 and 9 show how the $ACR(t)$ and $Q(t)$ are affected by the number of connections (N_{VC}). We choose the propagation delay $\tau_{sx} = \tau_{xd} = 0.005$ msec (2 km between source and destination) as a typical value for a LAN environment. In these figures, it is easily seen that the maximum queue length becomes large and the cycle is lengthened as N_{VC} increases. When the number of connections is fairly small, however, the maximum queue length in the LAN environment can be limited to an acceptable value.

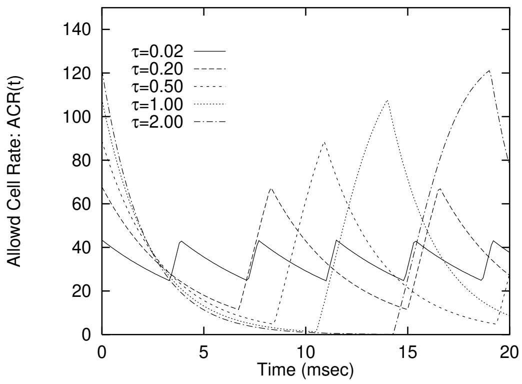

Figure 10: Effect of Propagation Delay on $ACR(t)$ in an EFCI Switch ($N_{VC} = 10$).

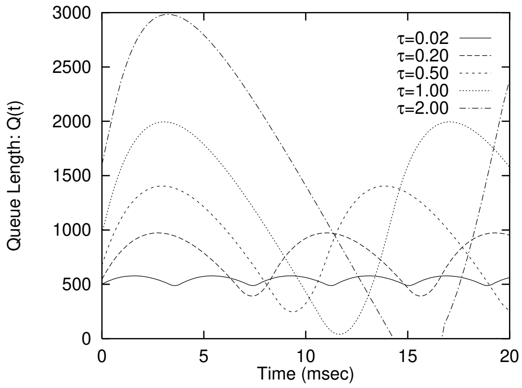

Figure 11: Effect of Propagation Delay on $Q(t)$ in an EFCI Switch ($N_{VC} = 10$).

$ACR(t)$ and $Q(t)$ for different values of propagation delays are compared in Figs. 10 and 11, which show that a larger τ causes slower congestion notification and results in an increased maximum queue length. Furthermore, under utilization occurs when τ is more than 2.0 msec (correspond to about 400 km). We can therefore conclude from these numerical results that the EFCI switch should be used in rather small networks.

As noted in Subsection 2.2, another problem of the EFCI switch appears at larger N_{VC}. Figure 12 shows that the maximum queue length grows almost linearly. As will be shown in the following subsections, the maximum queue length can be decreased considerably by introducing BES or EDS switches. When we consider

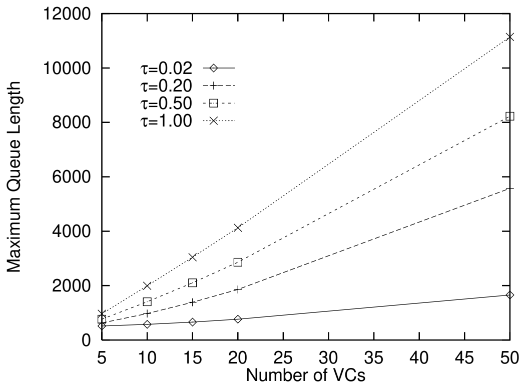

Figure 12: Effect of N_{VC} on the Maximum Queue Length in an EFCI Switch.

an initial transient state, however, that is, when all connections start its cell transmission at same time, the growth of the maximum queue length becomes unacceptable even with BES or EDS switches, as has been described in Subsection 2.3. This problem is investigated extensively in [33].

3.3 BES Switch

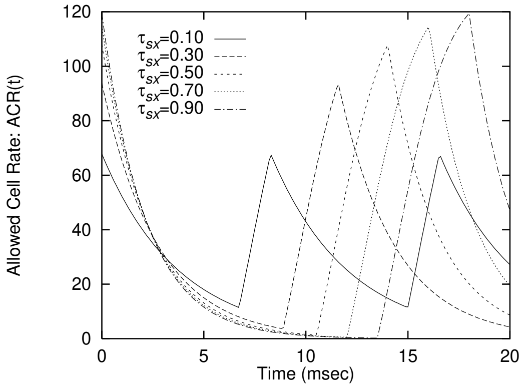

Figure 13: Effect of Switch Location on $ACR(t)$ in a BES Switch ($N_{VC} = 10, \tau_{sx} + \tau_{xd} = 1$).

To see the effect of the BECN-like capability of the BES switch more clearly, we show numerical examples taken from [32] in Figs. 13 and 14, where the distance of the switch from the source (τ_{sx}) is changed. In these figures, $\tau_{sx} + \tau_{xd}$ is fixed at 1.0 msec and N_{VC} is 10. We can easily observe the effect of the BECN-like quick congestion notification of the BES switch. When $\tau = 2.0$ msec, the maximum queue length can be reduced from 3,000 to 2,000 compared with the EFCI switch (Fig. 11).

3.4 Prioritized EDS Switch

The effect of N_{VC} on $ACR(t)$ and $Q(t)$ in the PEDS (prioritized EDS) switch, in which backward RM cells are given higher priority than other cells, are illustrated in Figs. 15 and 16 for $\tau = 0.02$ msec.

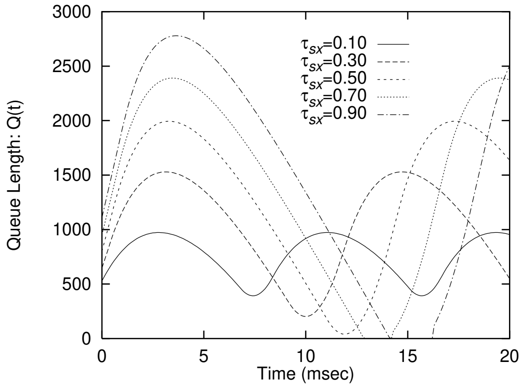

Figure 14: Effect of Switch Location on $Q(t)$ in a BES Switch ($N_{VC} = 10, \tau_{sx} + \tau_{xd} = 1$).

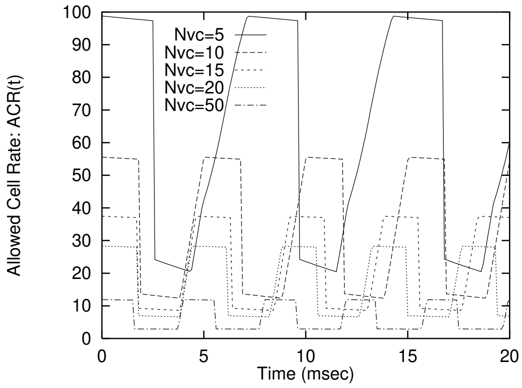

Figure 15: Effect of N_{VC} on $ACR(t)$ in a PEDS Switch ($\tau = 0.02$msec, $ERF = 15/16$).

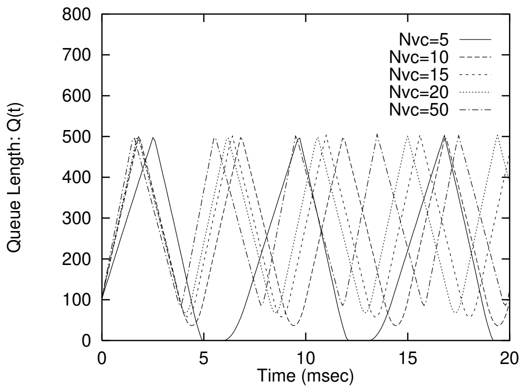

Figure 16: Effect of N_{VC} on $Q(t)$ in a PEDS Switch ($\tau = 0.02$msec, $ERF = 15/16$).

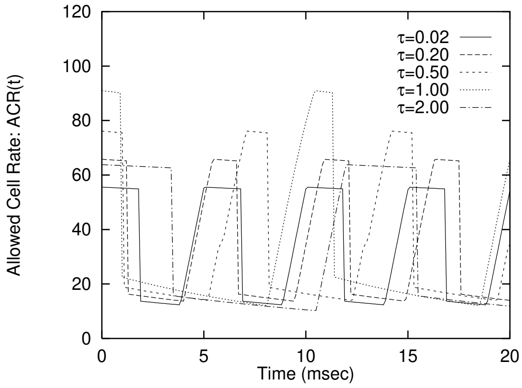

Figure 17: Effect of Propagation Delay on $ACR(t)$ in a PEDS Switch ($N_{VC} = 10$).

The effects of propagation delays are illustrated in Figs. 17 and 18 for $N_{VC} = 10$. In obtaining these figures, DQT was set to 500 and both Q_H and Q_L are set to 100. In all cases, the effect of major rate reduction is apparent because the queue length can be drastically decreased below that achieved BES switches or with EFCI switches. We notice, however, that the link is not fully utilized when τ becomes large (Fig. 18). This is due to inappropriate control parameter settings: $ERF = 15/16$ and $MRF = 1/4$. We need, therfore, set control parameters carefully to achieve better performance with the PEDS switches.

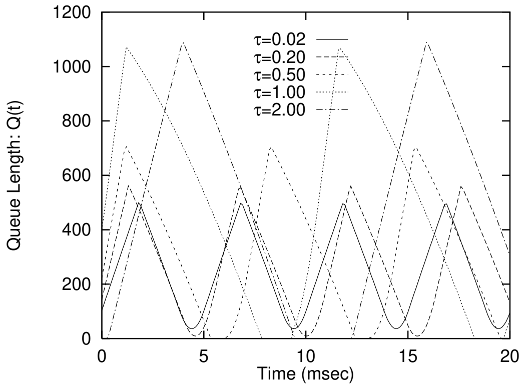

Figure 18: Effect of Propagation Delay on $Q(t)$ in a PEDS Switch ($N_{VC} = 10$).

4 Conclusion

Rate-based congestion control is a promising approach for incorporating ABR traffic into high-speed networks based on ATM technology. Many schemes for rate-based control approaches — from a "binary feedback" congestion indication scheme to an "explicit rate setting" scheme — in this area have been proposed and studied.

The historical overview in the first half of this paper pointed out advantages and disadvantages of rate-based schemes so far proposed in the ATM Forum. One important feature of these algorithms is their flexibility. That is, backward compatibility is always preserved in these schemes. One may choose cheaper end systems and simple switches such as the EFCI switches because of their cost-effectiveness. Those who prefer stability and efficiency, on the other hand, may use more expensive but superior switch such as the EDS switch. Even if different switch architectures are used, all switches (and networks) should interwork with each other. Rate-based congestion control methods are still being actively discussed, and the final voting for standard documentation to be included in User-Network Interface (UNI) Specification Version 4.0 is planned for the June–August 1995 meeting.

In the second half of this paper, we analyzed the performance of EPRCA, a representative rate-based approach. We provide analysis for three types of switches — EFCI, BES and EDS switches — suggested in EPRCA. Numerical examples show that EFCI switches, which are simple because they use a "binary feedback congestion indication", work well in the LAN environment if the number of active VC's is limited. Analysis of BES switches shows that the effect of their BECN-like capability depends on the distance between the source and the switch. The maximum queue length can be dramatically decreased by using EDS switches that have an explicit rate setting capability. More details about the analytic approach and numerical results can be found in [31, 32, 33].

Appendix A Control Parameters

This appendix lists control parameters and variables used by rate-based control algorithms. These parameters are summarized from [10, 13, 15, 16, 18].

Source End System Parameters

— General —

PCR	Peak Cell Rate; a maximum rate which ACR can set
MCR	Minimum Cell Rate; a minimum rate of ACR
ICR	Initial Cell Rate; an initial/reset value for ACR
AIR	Additive Increase Rate; rate increase permitted
MDF	Multiplicative Decrease Factor; $MDF = 2^{MD}$

— Timer-based Only —

UI	Update Interval; a period for which ACR is re-evaluated

— After PRCA —

N_{RM}	Number of Cells/RM; $N_{RM} = 2^N$

Source End System Variables

— General

ACR	Allowed Cell Rate, current maximum for CCR
CCR	Current Cell Rate, current rate of cell transmission

— After PRCA —

ADR	Additive Decrease Rate (cells/unit time)

Destination End System Parameters

— Timer-based Only —

RMI	Minimum Interval of RM cells transmitted

Switch Parameters Settings[16]

— BES and EDS Switches —

$MACR$	N/A	Congestion point rate computed by switch per queue, Ideally, should be the available bandwidth divided by the number of active connections
DQT	100	High queue limit to determine very congested

— EDS Only —

VCS	7/8	VC Separator
AV	1/16	Exponential Averaging Factor; for averaging ACR's
MRF	1/4	Major Reduction Factor; for major reduction
DPF	7/8	Down Pressure Factor
ERF	15/16	Explicit Reduction Factor

RM Cell Fields

— General —

CI	Congestion Indicator; 0 = no congestion, 1 = congestion

— After PRCA —

DIR	Direction of RM cell; forward or backward

— After EPRCA —

CCR	Current Cell Rate in effect when forward RM cell is generated
ER	Explicit Rate; initially set to PCR, and possibly modified by intermediate nodes along the path

References

[1] P. Newman, "Traffic management for ATM local area networks," *IEEE Communications Magazine*, pp. 44–50, August 1994.

[2] G. Woodruff and R. Ksitpaiboon, "Multimedia traffic management principles for guaranteed ATM network performance," *IEEE Journal on Selected Areas in Communications*, vol. 8, pp. 437–446, April 1990.

[3] "Special issue on 'Congestion control in high speed networks'," *IEEE Communications Magazine*, vol. 29, October 1991.

[4] H. T. Kung and A. Chapman, "Credit-based flow control for ATM networks: Credit update protocol, adaptive credit allocation, and statistical multiplexing," *The Proceedings of SIGCOMM '94*, vol. 24, pp. 101–114, October 1994.

[5] P. Newman, "Backward explicit congestion notification for ATM local area netwoks," *IEEE GLOBECOM '93*, pp. 719–723, December 1993.

[6] R. Jain, "Myths about congestion management in high-speed networks," in *Information Network and Data Communications, IV* (M. Tienari and D. Khakhar, eds.), pp. 55–70, Elsevier Science Publishers B.V. (North-Holland), 1992.

[7] N. Yin and M. G. Hluchyj, "On closed-loop rate control for ATM cell relay networks," *IEEE INFOCOM '94*, pp. 99–109, June 1994.

[8] C. Ikeda and H. Suzuki, "Adaptive congestion control schemes for ATM LANs," *IEEE INFOCOM '94*, pp. 829–838, June 1994.

[9] A. Berger, F. Bonomi, K. Fendick, and J. Swenson, "Control of multiplexed connections using backward congestion control," *ATM Forum Contribution 94-064*, January 1994.

[10] B. Makrucki et al., "Closed-loop rate-based traffic management," *ATM Forum Contribution 94-438*, May 1994.

[11] A. W. Barnhart, "Baseline model for rate control simulations," *ATM Forum Contribution 94-0399*, May 1994.

[12] J. C. R. Bennett and G. T. D. Jardins, "Failure modes of the baseline rate based congestion control plan," *ATM Forum Contribution 94-0682*, July 1994.

[13] L. Roberts et al., "Closed-loop rate-based traffic management," *ATM Forum Contribution 94-0438R1*, June 1994.

[14] J. C. R. Bennett and G. T. D. Jardins, "Comments on the July PRCA rate control baseline," *ATM Forum Contribution 94-0682*, July 1994.

[15] H. Hsiaw et al., "Closed-loop rate-based traffic management," *ATM Forum Contribution 94-0438R2*, September 1994.

[16] L. Roberts, "Enhanced PRCA (proportional rate-control algorithm)," *ATM Forum Contribution 94-0735R1*, August 1994.

[17] K.-Y. Siu and H.-Y. Tzeng, "Adaptive proportional rate control for ABR service in ATM networks," Tech. Rep. 94-07-01, UC Irvine, July 1994.

[18] L. Roberts and A. W. Barnhart, "New pseudocode for explicit rate plus EFCI support," *ATM Forum Contribution 93-0974*, October 1994.

[19] Y. Chang, N. Golmie, L. Benmohamed, and D. Su, "Simulation study of the new rate-based EPRCA traffic management mechanism," *ATM Forum Contribution 94-0809*, September 1994.

[20] K.-Y. Siu and H.-T. Tzeng, "Adaptive proportional rate control (APRC) with intelligent congestion indication," *ATM Forum Contribution 94-0888*, September 1994.

[21] K.-Y. Siu and H.-Y. Tzeng, "Limits of performance in rate-based control schemes," *ATM Forum Contribution 94-1077*, November 1994.

[22] H.-Y. Tzeng and K.-Y. Siu, "Comparison of performance among existing rate control schemes," *ATM Forum Contribution 94-1078*, November 1994.

[23] R. Jain, S. Kalyanaraman, and R. Viswanathan, "Simulation results: The EPRCA+ scheme," *ATM Forum Contribution 94-0988*, October 1994.

[24] R. Jain, S. Kalyanaraman, and R. Viswanathan, "The OSU scheme for congestion avoidance using explicit rate indication," *ATM Forum Contribution 94-0883*, September 1994.

[25] R. Jain, S. Kalyanaraman, and R. Viswanathan, "Current default proposal: Unresolved issues," *ATM Forum Contribution 94-1175R1*, November 1994.

[26] R. Jain, S. Kalyanaraman, and R. Viswanathan, "Transient performance of EPRCA and EPRCA++," *ATM Forum Contribution 94-1173*, November–December 1994.

[27] S. Sathaye, "ATM Forum traffic management specification," *ATM Forum Contribution*, December 1994.

[28] D. Bertsekas and R. Gallager, *Data Networks*. Englewood Cliffs, New Jersey: Prentice-Hall, 1987.

[29] N. Yin, "Fairness definition in ABR service model," *ATM Forum Contribution 94-0928R2*, November 1994.

[30] D. Hughes, "Fairness share in the context of MCR," *ATM Forum Contribution 94-0977*, October 1994.

[31] H. Suzuki, C. Ikeda, and M. Murata, "Transient state analysis and maximum buffer requirements for proportional rate control algorithm," *ATM Forum Contribution 94-0910*, September 1994.

[32] H. Ohsaki, M. Murata, H. Suzuki, and H. Miyahara, "Analysis of rate-based congestion control methods in ATM networks — part 1: Steady state analysis —," *submitted to GLOBECOM '95*, 1995.

[33] H. Ohsaki, M. Murata, H. Suzuki, and H. Miyahara, "Analysis of rate-based congestion control methods in ATM networks — part 2: Initial transient state analysis —," *submitted to GLOBECOM '95*, 1995.