# A Fast Switch Algorithm for ABR Traffic to Achieve Max-Min Fairness*

Danny H. K. Tsang[1], Wales Kin Fai Wong[1], Sheng Ming Jiang[1] and Eric Y. S. Liu[2]

[1] Department of Electrical and Electronic Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
[2] Broadband and Data Strategies, Cable & Wireless, London, UK

**Abstract.** In this paper, a new rate-based switch mechanism is proposed for flow control of ABR traffic. By making use of the most up-to-date information from both the upstream and the downstream paths, the current bottleneck of the VC can quickly be found. The bandwidth allocation for different VCs can then be adjusted by using this new bottleneck information to achieve max-min fairness allocation [1]. We compare the proposed scheme to CAPC [2] and ERICA [3]. Simulation results showed that the transient response times of the sources are significantly reduced in the proposed scheme. Furthermore, the peak queue lengths of the switches are generally smaller.

## 1 Introduction

One of the challenges in developing Asynchronous Transfer Mode (ATM) is the support of connectionless traffic because connection-oriented ATM networks need to establish a connection before data can be transferred. Bandwidth negotiation is necessary such that appropriate resources can be allocated to the traffic to satisfy the user's declared Quality of Service (QoS). However, lacking of the prior knowledge of the characteristics of the connectionless traffic makes it very difficult to decide when and how much resources should be allocated. It is suggested that this kind of traffic should be supported by the Available Bit Rate (ABR) service. ABR traffic is used to fill in the bandwidth slack left by the scheduled traffic that has bandwidth and QoS guaranteed [4, 5].

Two flow control schemes for ABR traffic were under active discussion in ATM Forum [6]. They are credit-based flow control scheme [5] and rate-based flow control scheme [7]. After a prolong discussion, the traffic management group eventually adopted the rate-based approach. Since then, the original rate-based proposal was extensively modified [8]. Among the different approaches for rate-based flow control, the explicit rate (ER) approach requires the source to generate periodically resource management (RM) cells which traverse the path of the connection and eventually return back to the source. The switches along the

* Supported by Hongkong Telecom Institute of Information Technology grant HKTIIT93/94.EG01

path as well as the destination can use the ER field in the RM cells to carry congestion information about the path so that bandwidth allocation at the switches can be carried out. Upon receiving the RM cells, the source can then adjust its transmission rate based on the ER value in the received RM cells. Many of the proposed switch algorithms aim to achieve the max-min fairness allocation [1], a fairness criterion considered by ATM Forum. In this paper, a new switch mechanism which aims to rapidly achieve max-min fairness allocation is proposed. It is shown through simulations that the proposed scheme can significantly reduce both the transient response times of the sources and the peak queue lengths at the switches. In addition, the scheme is very simple and does not require any special parameters to be set.

The organization of the paper is as follows. In Section 2, two switch mechanisms previously proposed in ATM Forum are presented. In Section 3, a detailed description of the proposed scheme is given. Section 4 discusses the simulation results of both the transient response times of the sources and the peak queue lengths of the switches. Section 5 concludes the paper.

## 2 Previously Proposed Switch Mechanisms

### 2.1 Congestion Avoidance using Proportional Control (CAPC)

The basic idea of the Congestion Avoidance using Proportional Control (CAPC) [2] is to select a target rate, $R0$, at which the switch should operate. To achieve this, proportional feedback control is used along with the explicit rate approach.

The total input rate to the switch, $Rate$, is measured first. The rate adjustment factor, $delta$, is calculated as:

$$delta = 1 - Rate/R0 \tag{1}$$

If $delta$ is greater than 0, the explicit rate for the switch, $ERS$, is increased as follows:

$$ERS = ERS \cdot \min(ERU, 1 + delta \cdot Rup), \tag{2}$$

where $ERU$ is the maximum increase of $ERS$ (typically 1.5) and $Rup$ is the proportional constant for rate increase (typically 0.025 to 0.01). Otherwise, $ERS$ is reduced as follows:

$$ERS = ERS \cdot \max(ERF, 1 + delta \cdot Rdn), \tag{3}$$

where $ERF$ is the minimum decrease of $ERS$ (typically 0.5) and $Rdn$ is the proportional constant for rate decrease (typically 0.2 to 0.8).

When the switch receives a backward RM cell, the Explicit Rate ($ER$) field of the RM cell is updated to the minimum of the current value in the $ER$ field and $ERS$. In addition, CAPC also marks the Congestion Indication ($CI$) bit of the backward RM cells when the queue length exceeds a threshold. When the source receives a RM cell, its allowed cell rate, $ACR$, is increased by the value of Additive Increase Rate ($AIR$) only if the value of $CI$ is equal to zero

[2]. Furthermore, the final value of $ACR$ is always set to the minimum of the current $ACR$, the peak cell rate ($PCR$), and the value of the $ER$ field in the last received RM cell.

This scheme has several problems. First of all, the scheme requires the setting of many parameters. Incorrect setting of these parameters may lead to performance degradation. In addition, the use of queue length as overload indicator may lead to unfairness [8]. The scheme may also result in unnecessary oscillations [9].

### 2.2 Explicit Rate Indication for Congestion Avoidance (ERICA)

Instead of using queue length as the overload indicator, Explicit Rate Indication for Congestion Avoidance (ERICA) [3] uses the queue growth rate as the overload indicator. The switch measures the time $T$ for $N$ cell arrivals. If the available capacity of the link is $C$ cells per second and the target utilization is $U$, the overload factor can be computed as follows:

$$Overload\_Factor = N/(T * U * C) \tag{4}$$

At the end of the measurement interval of $N$ cell arrivals, the switch computes the overload factor and informs all the VCs passing through it to adjust their rates according to the overload factor. The scheme also takes fairness into consideration and is achieved by ensuring that every VC gets at least a fair share of bandwidth, $FS$, which is computed as follows:

$$FS = Target\_Cell\_Rate/Number\_of\_Active\_VC, \tag{5}$$

where $Target\_Cell\_Rate = U \cdot C$. The number of active VC is the number of distinct VCs that were seen transmitting during the last measurement interval of $N$ cell arrivals.

By combining the two factors, the switch's recommended $ER$ value for VC $i$ can be computed as:

$$ERS(i) = \max(FS, CCR(i)/Overload\_Factor), \tag{6}$$

where $CCR(i)$ is the current cell rate of VC $i$, which can be obtained from the most recently received RM cell of VC $i$.

When a backward RM cell of VC $i$ is received at the switch, the switch first computes $ERS(i)$ and then updates the $ER$ field of the RM cell to the minimum of the current value in the $ER$ field and the computed $ERS(i)$. When the source receives a returning RM cell, its $ACR$ is always set to the value of $ER$ in the received RM cell.

## 3 Proposed Max-Min Scheme

The proposed scheme operates as follow. Each switch maintains an information table for all active VCs that pass through it (e.g., see Table 1). VCI denotes the VC identifier. $ER\_f$ and $ER\_b$ respectively denote the $ER$ value of the most recent RM cell received in the forward and the backward directions. $CA$ is the current allocation for the VC at the switch. $Constrained$ is a boolean variable. When it is 1, the connection is a constrained one [10] and cannot achieve its fair share of bandwidth at this node because of the constraints imposed by its $PCR$ or by the limited amount of bandwidth available at other nodes along its path. Similarly, when $constrained = 0$, this implies the bandwidth of the connection is only limited by the bandwidth available at the considered node. Denote $N$ as the total number of active connections and $M$ as the number of constrained connections at the switch. The number of active VC is the number of distinct VCs that were seen transmitting during the last measurement interval of $N$ cell arrivals, as in ERICA [3].

**Table 1.** Information Table at the switch

| $VCI$ | $ER\_f$ | $ER\_b$ | $CA$ | $constrained$ |
|---|---|---|---|---|
| x | f1 | b1 | c1 | 0/1 |
| y | f2 | b2 | c2 | 0/1 |

When a RM cell is generated by the source, its $ER$ field is set to Peak Cell Rate ($PCR$) as depicted in Figure 1. When the switch receives a forward RM cell of VC $j$ with $ER$ field equal to $ER\_RM$, the switch will do the following:

i. IF $ER\_RM = ER\_f(j)$ THEN GOTO step ix
ii. $ER\_f(j) = ER\_RM$
iii. IF $\min(ER\_f(j), ER\_b(j)) \leq CA(j)$ THEN
$constrained(j) = 1$ and $CA(j) = \min(ER\_f(j), ER\_b(j))$
ELSE
$constrained(j) = 0$
iv. For all unconstrained connections $i$, let $CA(i) = \Lambda$, where

$$\Lambda = \frac{Available\_Bandwidth - \sum_{constrained\_connection} CA(k)}{N - M} \quad (7)$$

v. $changed = 0$
vi. For all unconstrained connections $i$
IF $\min(ER\_f(i), ER\_b(\mathrm{i})) \leq \Lambda$ THEN
$constrained(i) = 1, CA(i) = \min(ER\_f(i), ER\_b(i))$ and $changed = 1$

vii. For all constrained connections $k$
IF $\min(ER\_f(k),\ ER\_b(k)) > \Lambda$ THEN
$constrained(k) = 0$ and $changed = 1$
viii. IF $changed = 1$ GOTO step iv
ix. END

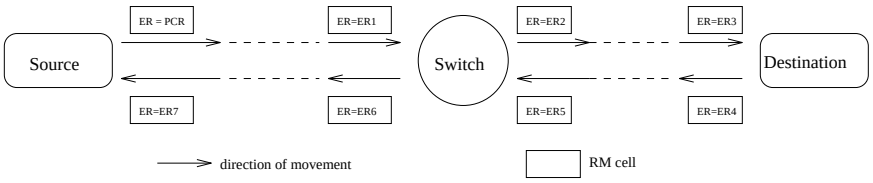


**Fig. 1.** Flow of RM cells

The algorithm works as follows. When a forward RM cell with $ER = ER\_RM$ for VC $j$ arrives at the switch, the switch checks whether $ER\_f(j)$ is equal to $ER\_RM$. If they are equal (step i), nothing needs to be done for this RM cell. Otherwise, $ER\_f(j)$ is set to $ER\_RM$ (step ii). If the minimum of the new $ER\_f(j)$ and $ER\_b(j)$ is less than $CA(j)$, this implies that the bottleneck of VC $j$ is elsewhere along its path. Therefore, $CA(j)$ is reduced to the minimum of $ER\_f(j)$ and $ER\_b(j)$, and $constrained$ is set to 1. Otherwise, $constrained$ is set to 0 (step iii). For all unconstrained connections $i$, $CA(i)$ is updated to $\Lambda$ (step iv) as in (7). Here, $\Lambda$ is the new current allocation for all unconstrained connections and $CA(k)$ is the current allocation for constrained connection $k$. For all unconstrained connections $i$, $\Lambda$ is compared to the minimum of $ER\_f(i)$ and $ER\_b(i)$ (step vi). If $\Lambda$ is larger, $constrained(i)$ is set to 1 and $CA(i)$ is set to the minimum of $ER\_f(i)$ and $ER\_b(i)$. It is because if $\Lambda$ is larger, the bottleneck is in fact elsewhere and thus the connection should be classified as a constrained one. The change in $\Lambda$ is due to either the change in the available bandwidth at some switch or the change in the number of active VCs in the network. Similarly, if the minimum of $ER\_f(k)$ and $ER\_b(k)$ is larger than $\Lambda$ for some constrained connection $k$ (i.e., the bottleneck for the VC is not elsewhere but at the current switch), $constrained$ is then set to 0 (step vii). If $changed = 1$ after steps vi and vii, further calculation of $\Lambda$ is necessary because of the change of some VC's constrained status. Therefore, steps iv to viii are repeated until $changed$ is 0 at the end of step vii. It is shown in [15] that the max-min calculation requires at most two iterations to converge.

The update of the $ER$ field is now discussed. As depicted in Figure 1, let $ER1$ be the $ER$ value in the RM cell when arrived at the switch and $CA$ be the newly computed current allocation for the VC at the switch. The new $ER$ value for the outgoing RM cell, $ER2$, is simply set to $CA$ since the computation of $CA$ in the algorithm has already taken both fairness and bottleneck information into account.

When the RM cell reaches the destination, it is turned around by the desti-

nation and the $ER$ value of the returning RM cell is reset to the minimum of $PCR$ and the destination's supported rate (i.e., $ER4$ in Figure 1). The resetting of the $ER$ value at the destination is important since it permits the independent flows of upstream and downstream congestion information via the RM cells to the switches. By making use of the most up-to-date congestion information from both the upstream and the downstream paths, the switches can know of the current bottleneck of the VC more quickly so that better bandwidth allocation can be performed.

When a backward RM cell is received at a switch, similar procedures as above are done except that $ER\_f(j)$ is replaced by $ER\_b(j)$ in steps i and ii. When the source receives the RM cell, it will set its $ACR$ to the $ER$ value in the received RM cell (i.e., $ER7$ in Figure 1).

When either the number of active VC or the available bandwidth at the switch changes, steps iv to viii of the above pseudocode must also be executed in order to determine the new allocation. When a VC is terminated, its entry in the information table at each of the switches involved must be deleted. On the other hand, when a new VC is established, a new row needs to be created in the information table at each of the switches involved. The initial values of $ER\_f$ and $ER\_b$ are set to $PCR$ while the initial constrained status is set to 0. The values of $CA$s for all VCs passing through the switches are recomputed using steps iv to viii of the above pseudocode.

In summary, by resetting the $ER$ field of the RM cells at the destination, the proposed scheme can significantly reduce the response times of the sources because the new congestion information carried by the forward RM cells can immediately be used at the switch to calculate the new $CA$s, which are then quickly carried back to the sources by the backward RM cells. With this, the sources can adjust to the max-min fairness allocation in much shorter times.

## 4 Performance of Max-Min scheme

Figure 2 shows the simulation model [11] which is implemented by using the simulation package BONeS [12]. In this network, there are two multi-hop VCs (VC2 and VC4) while the remaining VCs are single-hop. The source end system (SES) behavior is based on [4]. However, since no $NI$ field is used in CAPC, the operation based on $NI$ in the SES is disabled. Similarly, since no $NI$ and $CI$ fields are used in ERICA and the proposed scheme, the SES is modified such that the operations based on $NI$ and $CI$ are not carried out.

### 4.1 Simulation Settings

The values of the common parameters for the SES [4] are shown in Table 2. The one-way propagation delay between the source/destination and its attached switch is $5\mu s$ while the one-way propagation delay between two switches is $50\mu s$ (as suggested in [13] for LAN separation). The sources we used are staggered one (i.e., the sources become active one by one). VC1 starts at $0ms$. Ten starting

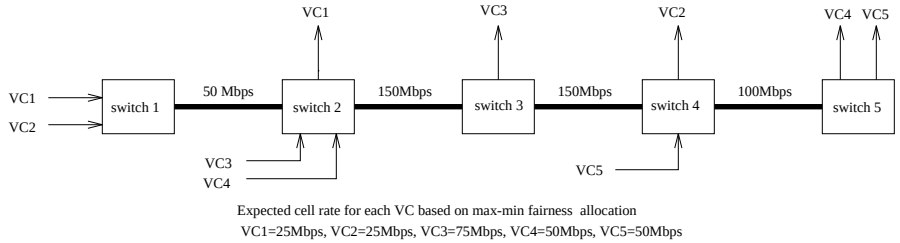


**Fig. 2.** Simulation Model

times are tested for each subsequently active VC. The mean time of becoming active for VC2, VC3, VC4 and VC5 are $5ms$, $10ms$, $15ms$ and $20ms$, respectively. The ten starting times are equally spaced and cover an interval of width equal to $Nrm$ cell times of the previously started VC. The reason is to take into account of the different arrival times of the RM cells. The sources remain active after startup until the end of simulation. Each switch attempts to fully utilize the total available bandwidth (e.g., $150Mbps$ for switch 2). Different initial cell rates, $ICR$s, are used for comparison. The values of the parameters used in CAPC are based on [14] and are shown in Table 3. For ERICA, the counting interval $N$ is 30 cells [3].

**Table 2.** Setting of Common Parameters for SES

| $PCR$ | $MCR$ | $Nrm$ | $RDF$ | $TOF$ |
|---|---|---|---|---|
| 150Mbps | PCR/1000 | 32 | 1024 | 2 |

**Table 3.** Setting of Parameters for CAPC

| $AIR$ | $Rup$ | $Rdn$ | $ERU$ | $ERF$ | $interval$ | $Qthreshold$ |
|---|---|---|---|---|---|---|
| PCR | 0.25 | 1.0 | 1.5 | 0.5 | 1ms | 100 cells |

### 4.2 Performance Comparison

Table 4 shows the max-min fairness allocation for different sources at different times. The values of $ACR$ for the different VCs are shown in Figures 3-5 when the VCs become active one by one. Figure 3 shows the values of $ACR$ for CAPC for three different values of $ICR$. The figures show that, for most cases, the $ACR$s of the VCs cannot converge to the steady-state values before the next VC becomes active. After all five VCs are active, it takes approximately 15 to $20ms$ for the VCs to achieve the max-min fairness allocation. The scheme has the longest response time when compared to the other two schemes.

**Table 4.** Max-Min fairness allocation at different times

| | $ACR1$ | $ACR2$ | $ACR3$ | $ACR4$ | $ACR5$ |
|---|---|---|---|---|---|
| when VC1 is active | 50 | N/A | N/A | N/A | N/A |
| when VC2 is active | 25 | 25 | N/A | N/A | N/A |
| when VC3 is active | 25 | 25 | 125 | N/A | N/A |
| when VC4 is active | 25 | 25 | 62.5 | 62.5 | N/A |
| when VC5 is active | 25 | 25 | 75 | 50 | 50 |

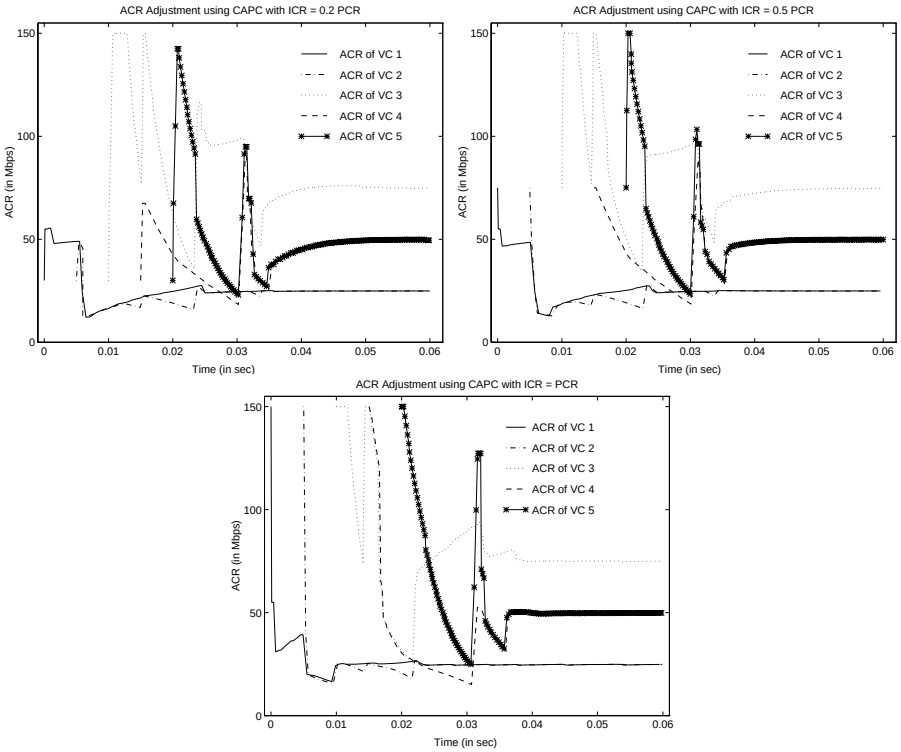


**Fig. 3.** ACR adjustment using CAPC under different ICRs

Figure 4 shows the values of $ACR$ for ERICA under three different $ICR$s. It shows that the response times of the sources in ERICA are better than those in CAPC. However, ERICA sometimes cannot converge to the max-min fairness allocation (e.g., in Figure 4 after VC4 becomes active in the interval between 15 and $20ms$). According to the max-min fairness allocation, after VC4 becomes active, $ACR$s for VCs 2, 3 and 4 should be $25Mbps$, $62.5Mbps$ and $62.5Mbps$ respectively.

The convergence problem of ERICA can be explained as follows. Before VC4 becomes active, $ACR$s for VCs 2 and 3 are $25Mbps$ and $125Mbps$, respectively.

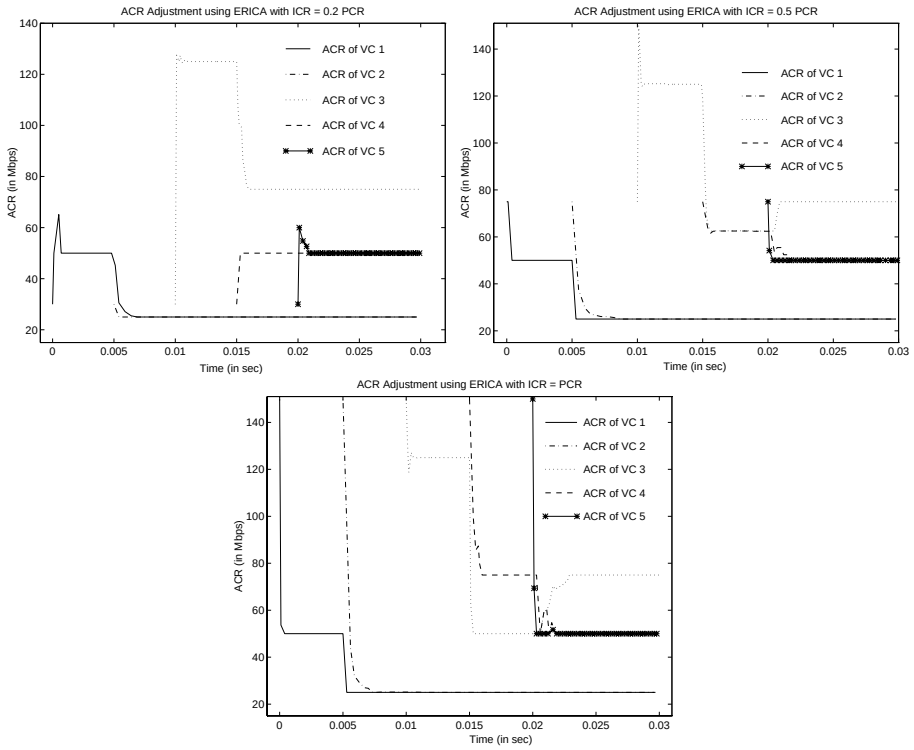


**Fig. 4.** ACR adjustment using ERICA under different ICRs

This means VCs 2 and 3 can fully utilise the output link of switch 2 (Figure 2). Therefore, once VC4 becomes active at $15ms$, the overload factor at switch 2 is over 1. For the case of $ICR = 0.2PCR$ (i.e., $ACR$ for VC4 is $30Mbps$ at $15ms$), only $ACR$ for VC3 is reduced because (i) $ACR$ for VC4 is allowed to increase to at least $50Mbps$, which is the fair share value suggested by (5), and (ii) since the bottleneck for VC2 is at switch 1, $ACR$ for VC2 is thus maintained at $25Mbps$. Therefore, the overload factor at switch 2 continues to be above 1 since only $ACR$ for VC3 is reduced. The reduction ends when $ACR$ for VC3 drops to $75Mbps$ and $ACR$ for VC4 increases to $50Mbps$. For the case of $ICR = PCR$, $ACR$s for VCs 2, 3 and 4 should all be decreased since the overload factor at switch 2 is over 1. However, since the bottleneck for VC2 is at switch 1, $ACR$ for VC2 is maintained at $25Mbps$. Moreover, $ACR$ for VC3, a single-hop connection, decreases faster than that for VC4 because VC3 reacts faster to network feedback [11]. $ACR$ for VC3 continues to decrease until it reaches $50Mbps$ (the fair share allocation) and $ACR$ for VC4 is reduced to $75Mbps$. At that time, the overload factor becomes one and steady state is achieved.

For the proposed scheme, max-min fairness allocation is always achieved and the response times of the sources are the shortest among the three schemes examined in this paper (see Figure 5). Moreover, the performance of the proposed scheme is independent of $ICR$ and is approximately the same for the different

cases of $ICR$s. Since the response times of the sources in CAPC are much larger than those in ERICA, our comparison will focus only between ERICA and our proposed scheme. In addition, since ERICA cannot achieve max-min fairness allocation for $ICR = 0.2PCR$ and $ICR = PCR$ in certain time intervals, we compare only for the case of $ICR = 0.5PCR$.

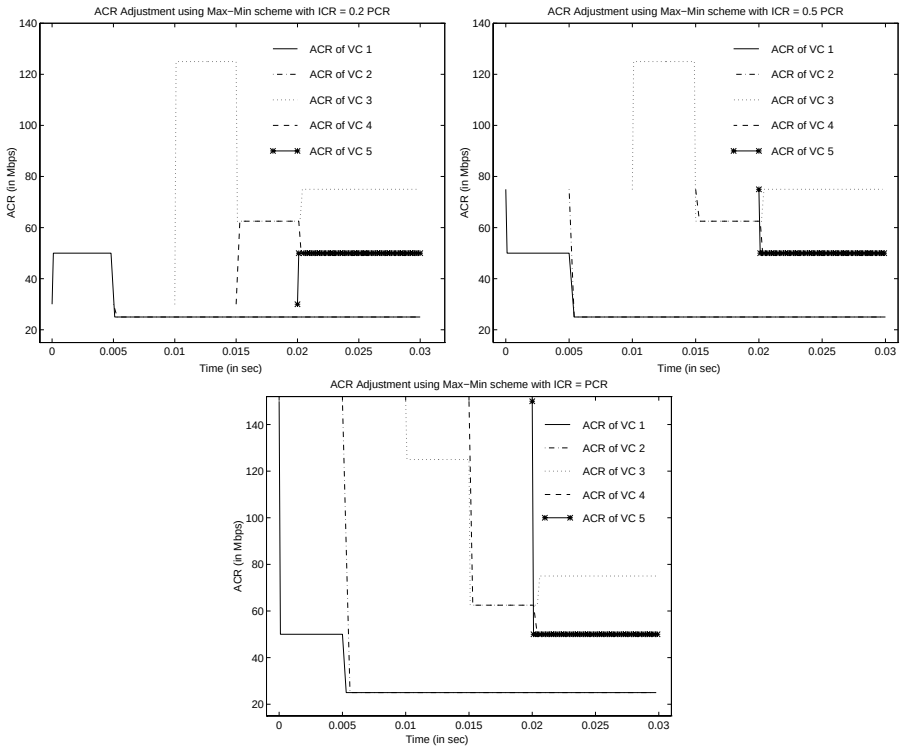


**Fig. 5.** ACR adjustment using Max-Min Scheme under different ICRs

Tables 5 and 6 show the transient response times of the sources for the case of $ICR = 0.5PCR$ for ERICA and the proposed scheme, respectively. They show that the response times of our scheme are much faster than that of ERICA.

In Table 7, the peak queue lengths at different switches for the two schemes are shown. It shows that a significant reduction in peak queue length is achieved by the proposed scheme. Small queue length is important for local area networks (LANs) because the buffer size of LAN switch is usually small. Better control of queue length can reduce the number of cell loss and therefore minimizes the performance degradation due to cell loss.

**Table 5.** Transient Response Time in $\mu$s for ERICA

| | *ACR1* | *ACR2* | *ACR3* | *ACR4* | *ACR5* |
|---|---|---|---|---|---|
| when VC1 is active | 405±0 | N/A | N/A | N/A | N/A |
| when VC2 is active | 531.7±268.4 | 3403.1±350 | N/A | N/A | N/A |
| when VC3 is active | 0±0 | 0±0 | 2791.6±25.8 | N/A | N/A |
| when VC4 is active | 0±0 | 0±0 | 2165.8±854.8 | 2161.8±886 | N/A |
| when VC5 is active | 0±0 | 0±0 | 1933.5±342.6 | 1463.3±514.2 | 484.7±84.6 |

**Table 6.** Transient Response Time in $\mu$s for Max-Min Scheme

| | *ACR1* | *ACR2* | *ACR3* | *ACR4* | *ACR5* |
|---|---|---|---|---|---|
| when VC1 is active | 134±0 | N/A | N/A | N/A | N/A |
| when VC2 is active | 159.2±77.8 | 408.9±1.6 | N/A | N/A | N/A |
| when VC3 is active | 0±0 | 0±0 | 128±0 | N/A | N/A |
| when VC4 is active | 0±0 | 0±0 | 69.9±23.4 | 338.5±1.1 | N/A |
| when VC5 is active | 0±0 | 0±0 | 322.3±46.3 | 227.3±54.6 | 131.2±1.2 |

**Table 7.** Comparison of Peak Queue Lengths in cells

| | *Switch 1* | *Switch 2* | *Switch 3* | *Switch 4* |
|---|---|---|---|---|
| ERICA | 131±4 | 53.8±2.8 | 2±0 | 54.2±7.6 |
| Max-Min Scheme | 66.9±4.6 | 22.6±3.5 | 2±0 | 21.9±1.6 |

## 5 Conclusion

A new switch mechanism that can quickly achieve the max-min fairness allocation is proposed. Max-min fairness is the fairness criterion considered by ATM Forum. In addition to always achieving the max-min fairness allocation, the proposed scheme also has several other advantages over the previously proposed schemes. One is the significant reduction of the transient response times of the sources. Another is the reduction of peak queue lengths at the switches. Furthermore, the proposed scheme is very simple and does not require any special parameters to be set. Therefore, the performance will not be degraded due to the improper setting of parameters.

## References


1. Bartsekas, D., Gallager, R.: Data Networks. Prentice Hall, 2nd Edition, 1987
2. Barnhart, A.W.: Explicit Rate Performance Evaluation. ATM Forum 94-0983R1
3. Jain, R.: A Sample Switch Algorithm. ATM Forum 95-0178R1
4. Sathay, S.: ATM Forum Traffic Management Specification Version 4. ATM Forum 95-0013

5. Kung, H.T., Morris, R., Charuhas, T., Lin, D.: Use of Link-by-Link Flow Control in Maximising ATM Networks Performance: Simulation Results. Proc. IEEE Hot Interconnects Symposium, '93
6. Saunders, S.: ATM Forum Ponders Congestion Control Options. Data Communications, March 1994, page 55-60
7. Ramakrishnan, K. K., Jain, R.: A Binary Feedback Scheme for Congestion Avoidance in Computer Networks. ACM Transaction on Computer Systems, Vol. 8, No. 2, May 1990, page 159-181
8. Jain, R.: Congestion Control and Traffic Management in ATM Networks: Recent Advances and A Survey. Invited submission to Computer Networks and ISDN Systems
9. Jain, R., et al.: Rate Based Schemes: Mistakes to Avoid. ATM Forum 94-0882
10. Bonomi, F., Fendick, K.W.: The Rate-Based Flow Control Framework for the Available Bit Rate ATM Service. IEEE Network Magazine, March/April 1995, page 25-39
11. Chang, Y., Golmie, N., Su, D.: Comparative Analysis of the Evolving End System Behavior (Simulation Study). ATM Forum 95-0395R1
12. COMDISCO System Inc.: BONeS DESIGNER Core Library Guide. June 1993
13. Wojnaroski, L.: Baseline Text for Traffic Management Sub-Working Group. ATM Forum 94-0394R5
14. Barnhart, A.W.: Providing improved Explicit Rate Performance for the LAN. ATM Forum 94-1111
15. Charny, A.: An Algorithm for Rate Allocation in a Packet-Switching Network with Feedback. MIT/LCS/TR-601. April 1994.

This article was processed using the LaTeX macro package with LLNCS style