A New Rate-Based Switch Algorithm for ABR Traffic to Achieve Max-Min Fairness with Analytical Approximation and Delay Adjustment*

Danny H. K. Tsang and Wales Kin Fai Wong
Department of Electrical and Electronic Engineering
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
eetsang@ee.ust.hk

Abstract

In this paper, a new rate-based switch mechanism for ABR traffic in ATM networks, which aims to rapidly achieve max-min fairness allocation, is proposed. Simulation results show that the proposed scheme can out-perform both CAPC [AF94-0983] and ERICA [AF95-0178] in terms of response times and peak queue lengths. An analytical approximation of the performance is also introduced and its accuracy is found to be close to the simulation results. A variant of the proposed scheme is presented for handling the problem of different source-to-bottleneck separations. By using this scheme, the peak queue lengths at the switches can further be reduced without any degradation in throughput.

Keywords
Congestion Control, ABR Traffic, ATM, Fairness

1 Introduction

Since the traffic management group of the ATM Forum had adopted the rate-based approach for the Available Bit Rate (ABR) traffic in ATM networks, the original rate-based proposal was extensively modified [Jain95]. Many criteria are used to evaluate the performance of different rate-based flow control schemes. One of them is fairness. The definition that the ATM Forum adopted is called max-min fairness [Bart87, Jain95]. However, some of the proposed schemes cannot always achieve the max-min fairness allocation. Even if max-min fairness can be reached, the schemes may require a long time to do so.

In this paper, a new rate-based switch mechanism called Max-Min scheme, which aims to rapidly achieve max-min fairness allocation, is proposed. The basic idea is to divide the connections at each switch into two groups: constrained and unconstrained. With the use of the Resource Management (RM) cells, the bottleneck bandwidth values of different constrained connections can propagate to the switches along the paths of the connections. Bandwidth is allocated to the constrained connections based on the bottleneck information conveyed by the RM cells. The leftover bandwidth is then evenly distributed among all the unconstrained connections. It is shown through simulations that the proposed scheme can significantly reduce the transient response times as well as the peak queue lengths. In addition, the scheme is very simple and does not require any parameters to be set.

In almost all of the proposed schemes, the sources immediately modify their allowed cell rates (ACRs) upon receiving an RM cell. The immediate adjustment may lead to overloading of the bottleneck switch, and hence cell loss [DT95]. To deal with this problem, a rate-based scheme in which the sources may *not* alter their ACRs immediately is also proposed. It is shown through simulations that if delay adjustment is employed, the overloading of the bottleneck switch can be avoided and thus the peak queue lengths of the bottleneck switches can further be minimized.

The organization of this paper is as follows. Section 2 describes two switch algorithms presented in ATM Forum. Our proposed Max-Min scheme is introduced in Section 3. In Section 4, its performance are compared to that of the schemes discussed in Section 2. An analytical approximation of the performance for the Max-Min scheme is described in Section 5. Section 6 introduces a modification of the Max-Min scheme to deal with the problem of different source-to-bottleneck separations, and its performance is presented in Section 7. Section 8 concludes the paper.

2 Previous Work on Flow Control for ABR Traffic

2.1 Congestion Avoidance using Proportional Control (CAPC)

The idea of the Congestion Avoidance using Proportional Control (CAPC) [AF94-0983] 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:

$$delta = 1 - Rate/R0.$$

If $delta$ is greater than 0, the explicit rate for the switch,

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

ERS, is increased as follows:

$$ERS = ERS \cdot \min(ERU, 1 + delta \cdot Rup),$$

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:

$$ERS = ERS \cdot \max(ERF, 1 + delta \cdot Rdn),$$

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 newly received RM cell is updated to the minimum of the current value in the ER field and ERS. In addition, CAPC also marks the CI bit of the backward RM cells when the queue length exceeds some threshold. When the source receives a backward RM cell, its allowed cell rate, ACR, is increased by the value of additive increase rate (AIR) if the value of CI is equal to zero[AF94-0983]. 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 setting of many parameters. Incorrect setting of the parameters may lead to performance degradation. Furthermore, the use of queue length as overload indicator may lead to unfairness. It is because the sources that start up late are found to get lower throughput than those which start early [Jain95]. The scheme may also result in unnecessary oscillations [AF94-0882].

2.2 Explicit Rate Indication for Congestion Avoidance (ERICA)

To tackle the problem of using queue length as overload indicator, Explicit Rate Indication for Congestion Avoidance (ERICA) [AF95-0178] is proposed. The idea is to use the queue growth rate instead 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 desired target utilization is U, the overload factor can be computed as follows:

$$Overload\_Factor = N/(T * U * C).$$

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. Fairness is achieved by ensuring that every VC gets at least a fair share of bandwidth, FS, which is computed as follows:

$$FS = Target\_Cell\_Rate/Number\_of\_Active\_VC,$$

where $Target\_Cell\_Rate = U \cdot C$ and $Number\_of\_Active\_VC$ is the number of distinct VCs that were seen transmitting during the last measurement interval of N cell arrivals.

Combining the two factors, we get:

$$ERS(i) = \max(FS, CCR(i)/Overload\_Factor),$$

where $ERS(i)$ is the switch's recommended ER value for VC i. $CCR(i)$ is the current cell rate of VC i, which is available from the most recently received RM cell of VC i.

When a returning RM cell for VC i arrives 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 $ERS(i)$. When the source receives a backward RM cell, its ACR is always set to the value of ER in the received RM cell. It was shown in [DT96] that ERICA cannot always achieve max-min fairness.

3 Proposed Max-Min Scheme

The aim of the proposed Max-Min scheme [DT96] is to quickly achieve max-min fairness allocation when the network condition has changed. This can be done by using the information carried in the RM cells. Each switch maintains an information table for all active VCs that pass through it (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 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 [Bono95] 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 that the bandwidth of the connection is only limited by the bandwidth available at this node. Denote N as the total number of active connections and M as the number of constrained connections.

When the RM cell comes out from the source, its ER is set to PCR as depicted in Figure 3. When the switch receives a forward RM cell of VC j with ER field equal to ER_RM, the switch will do the following:

1. IF $ER\_RM = ER\_f(j)$ THEN GOTO step 6
2. $ER\_f(j) = ER\_RM$
3. 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 $constrained(j) = 0$
4. For all unconstrained connections i, let $CA(i) = \Lambda$, where
$$\Lambda = \frac{ABW - \sum_{constrained\_connection} CA(k)}{N - M} \quad (1)$$
5. $changed = 0$
6. For all unconstrained connections i
   IF $\min(ER\_f(i), ER\_b(i)) \leq \Lambda$ THEN
   $constrained(i) = 1$,
   $CA(i) = \min(ER\_f(i), ER\_b(i))$, and
   $changed = 1$
7. For all constrained connections k
   IF $\min(ER\_f(k), ER\_b(k)) > \Lambda$ THEN
   $constrained(k) = 0$ and $changed = 1$
8. IF $changed = 1$ GOTO step 4
9. END

ABW in (1) refers to the available bandwidth for ABR traffic. More detailed explanation of the above algorithms

can be found in [DT96]. Note that the computation of the term $\sum_{constrained\_connection} CA(k)$ can be more efficiently performed by considering the changes only. It is therefore not necessary to sum up $CA(k)$ every time when (1) is used.

As depicted in Figure 3, let $ER1$ be the ER value in the RM cell when arrived at the switch and CA be the current allocation for the VC at the switch. The new ER value for the outgoing RM cell, $ER2$, is computed as follows:

$$ER2 = CA. \tag{2}$$

When the RM cell reaches the destination, it is turned around by the destination 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 3). The resetting of the ER value in the RM cells allows more up-to-date bottleneck information from both forward and backward directions to reach the switches quicker and hence can improve the response time of the sources. Procedures similar to the above pseudocode are done when a backward RM cell is received at the switch, except that $ER\_f(j)$ is replaced by $ER\_b(j)$ in steps 1 and 2. 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 3).

When either the number of active VCs or the available bandwidth at the switch changes, steps 4 to 8 of the above pseudocode must also be executed in order to determine the new max-min fairness allocation. When a VC is terminated, its entry in the information table at the switches involved must be deleted. On the other hand, when a new VC is established, a new row in the information table at the switches involved needs to be created. 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 CAs for all VCs passing through the switches are recomputed using steps 4 to 8 in the above pseudocode.

4 Performance of the Max-Min Scheme

The performance of the Max-Min scheme, in terms of transient response times and peak queue lengths, are compared to CAPC and ERICA in [DT96]. Here, we present a summary of these results.

Figure 4 shows the simulation model [AF95-0395] that is implemented by using the simulation package BONeS [COMD93]. The source end system (SES) behavior is based on [AF95-0013]. 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 are shown in Table 2. The one-way propagation delay between the source or destination and its attached switch is $5\mu s$ while the one-way propagation delay between two switches is $50\mu s$. The sources we used are staggered one (i.e., the sources become active one by one). Ten random starting times are tested for every active VCs. The mean time of becoming active for VC1, VC2, VC3, VC4 and VC5 are $0ms$, $5ms$, $10ms$, $15ms$ and $20ms$, respectively with uniform distribution over intervals of width equal to Nrm cell times. The reason is to take into account of the different arrival times of the RM cells. The sources remain active once after startup until the end of simulation.

Each switch attempts to fully utilize the total available bandwidth (e.g., $150Mbps$ for switch 2). The connection is said to be active if the switch receives a cell from the particular connection. Different initial cell rates, ICRs, are used for comparison. The values of the parameters used in CAPC are based on [AF94-0983] and are shown in Table 3. For ERICA, the counting interval N is 30 cells, as suggested in [AF95-0178].

4.2 Performance Comparison

It is found in [DT96] that the response time of CAPC is much larger than that of ERICA. Therefore, 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 will concentrate on the case of $ICR = 0.5PCR$.

Tables 4 and 5 show the transient response times for the case of $ICR = 0.5PCR$ for ERICA and the proposed scheme, respectively. They show that the response times of the proposed scheme are much faster than that of ERICA.

In Table 6, the peak queue lengths at different switches for the two schemes are shown. It shows that a significant reduction in the peak queue length is achieved by the proposed scheme. It is vital in 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.

5 Analysis of the Max-Min Scheme

This section presents analytical approximations for the response time and the peak queue length of the Max-Min scheme. As mentioned in [Bono95], max-min fairness can be achieved by allocating the bottleneck bandwidths to the constrained connections, and then equally distributing the leftover bandwidth among the unconstrained connections. Therefore, by understanding how the bottleneck information flows from one switch to another, we can estimate the time that the RM cells must take before carrying the most up-to-date information back to the sources. After knowing the response times of different connections, the peak queue length at the switches can then be found.

Let us consider the following scenario. There is a group of VCs in a network. The max-min fairness allocation of VC i is λ_i. Denote $\Delta_{bs}(i)$ as the delay that the RM cells of VC i experience when traveling from the bottleneck switch of VC i to the source of VC i, and $\Delta_{RM}(i)$ as the mean time for the next RM cell of VC i to be seen at a particular switch. If the current input cell rate to the switch of VC i is $\alpha(i)$, then $\Delta_{RM}(i)$ is given by:

$$\Delta_{RM}(i) = \frac{Nrm}{2 \cdot \alpha(i)}. \tag{3}$$

Without loss of generality, we assume the new max-min fairness allocation satisfies the following condition:

$$\lambda_1 < \lambda_2 < \cdots < \lambda_m.$$

For the sake of simplicity, we also assume that the calculation of λ_i has to rely on λ_{i-1}. We will remove this assumption later in this section after the basic idea of this analytical approximation is introduced. Denote $\Delta_{i-1,i}$ as the delay that information on the bottleneck switch of VC $(i-1)$ can be delivered to the bottleneck switch of VC i, which includes the propagation delays, transmission delays and queueing delays at different network entities. Also denote τ_i as the response time of VC i for converging to its max-min fairness allocation, and S_i as the bottleneck switch of VC i. $Q_{i,old}$ is the queue length of S_i before the network condition changes and $Q_{i,new}$ is the queue length of S_i after all VCs converge to the new max-min fairness allocation. The initial value of $Q_{i,old}$ can be set to zero at the beginning. Let $\beta(i)$ be the new max-min fairness allocation for VC i. Also let $\Delta_{sb}(i)$ be the delay experienced by the cells of VC i when they travel from the source to S_i, and $\Delta_{bs}(i)$ be the delay that the RM cell of VC i could experience when they travel from S_i to the source. Both terms include the propagation delays, transmission delays and queueing delays at different network entities.

Now suppose the condition of the network changes. If we assume that there are RM cells traveling in the return path, the response time of VC 1, which has the smallest max-min allocation, is given by:

$$\tau_1 = \Delta_{0,1} + \Delta_{RM}(1) + \Delta_{bs}(1). \quad (4)$$

$\Delta_{0,1}$ is the time that the bottleneck switch of VC 1, S_1, becomes aware of the change in the network. $\Delta_{RM}(1)$ is required because the new allocation done at S_1 has to be carried by the RM cells and, on average, the switch has to wait for $\Delta_{RM}(1)$ for the next RM cell to arrive. The last term represents the delay that the RM cell could experience on its way to the source.

However, in case VC 1 is a starting VC, there is no RM cell in the return path. The minimum time for the first RM cell to return is its round trip time RTT_1. In addition, if the new max-min fairness allocation $\beta(1)$ for VC 1 is the same as the allocation before adjustment $\alpha(1)$ (*i.e.*, $\alpha(1) = \beta(1)$), we should take τ_1 as zero instead. In summary, the response time for VC 1 is given by:

IF $\alpha(1) \neq \beta(1)$ THEN
  IF VC 1 is a starting connection THEN
    $\tau_1 = RTT_1$
  ELSE
    $\tau_1 = \Delta_{0,1} + \Delta_{RM}(1) + \Delta_{bs}(1)$
ELSE
  $\tau_1 = 0.$

The queue length built up at S_1 by VC 1 after it changed from $\alpha(1)$ to $\beta(1)$ can be found by:

$$Q_{1,new} = \max\{Q_{1,old} + [\alpha(1) - \beta(1)] \cdot [\tau_1 + \Delta_{sb}(1)], 0\}. \quad (5)$$

$\alpha(1) - \beta(1)$ is the amount of mismatch between the current input rate and the desired input rate to S_1. $\tau_1 + \Delta_{sb}(1)$ represents the time that the cells transmitted at rate $\beta(1)$ can reach S_1. Since the queue length cannot be smaller than zero, we introduce a lower bound of zero here.

By the same approach, if VC 2 is not a starting VC and $\alpha(2) \neq \beta(2)$, the response time of the VC 2 is given by:

$$\tau_2 = \Delta_{0,1} + \Delta_{RM}(1) + \Delta_{1,2} + \Delta_{RM}(2) + \Delta_{bs}(2). \quad (6)$$

Since we assume that the calculation of λ_2 depends on λ_1, the first two terms denote the time that λ_1 can be found in S_1 and the third term denote the time required by the RM cell to carry λ_1 to S_2. In the calculation of $\Delta_{1,2}$ and $\Delta_{bs}(2)$, we can also estimate the queueing delay at S_1 by considering $Q_{1,new}$. After taking into account of different situations, the response time of VC 2 is given by:

IF $\alpha(2) \neq \beta(2)$ THEN
  IF VC 2 is a starting connection THEN
    $\tau_2 = RTT_2$
  ELSE
    $\tau_2 = \sum_{i=1}^{2}[\Delta_{i-1,i} + \Delta_{RM}(i)] + \Delta_{bs}(2)$
ELSE
  $\tau_2 = 0.$

Using the same approach as in (5), the queue length at S_2 built up by VC 2 can be found by:

$$Q_{2,new} = \max\{Q_{2,old} + [\alpha(2) - \beta(2)] \cdot [\tau_2 + \Delta_{sb}(2)], 0\}. \quad (7)$$

If $S_2 = S_1$, $Q_{2,old} = Q_{1,new}$ which is obtained by (5). Otherwise, $Q_{2,old}$ can be set to zero. Likewise, if we keep on iterating, the response time for VC j can be found by:

IF $\alpha(j) \neq \beta(j)$ THEN
  IF VC j is a starting connection THEN
    $$\tau_j = RTT_j \quad (8)$$
  ELSE
    $$\tau_j = \sum_{i=1}^{j}[\Delta_{i-1,i} + \Delta_{RM}(i)] + \Delta_{bs}(j) \quad (9)$$
ELSE
  $$\tau_j = 0. \quad (10)$$

Since for some connections j and k, their bottleneck may be at the same switch (i.e., $S_j = S_k$). The queue length at a switch should consider the changes in ACRs for all connections passing through it. Therefore, for all connections i passing through a bottleneck switch, its queue length after max-min fairness allocation is achieved can be found by:

$$Q_{length} = \max\{Q_0 + \sum_{for\ all\ i\ passing\ through\ the\ switch} [\alpha(i) - \beta(i)] \cdot [\tau_i + \Delta_{sb}(i)], 0\}. \quad (11)$$

Here Q_{length} is the queue length after all the connections converge to the max-min fairness allocation while Q_0 is the peak queue length before the network condition has changed. The initial value of Q_0 is zero as usual. τ_i is the response time calculated from either (8), (9) or (10). Since Q_{length} is never smaller than zero, we introduce a lower bound of zero as well.

We now remove the assumption that the calculation of λ_i has to rely on λ_{i-1}. In the early part of this section, we consider VC i as *one* connection. However, we can extend this idea by replacing VC i with a set of connections that share the same properties. For example, we can replace VC 1 by a set of connections whose max-min fairness allocation can all be found once their bottleneck switches become aware of the change(s). These connections shall be called *first-level*

connections. Their response times can be found by using (4), with different values of $\Delta_{0,1}, \Delta_{RM}(1)$ and $\Delta_{bs}(1)$ for different connections.

In addition, there may be connections whose max-min fairness allocation can only be determined after the allocation of a particular first-level connection is found. All the connections that share this property shall be called *second-level connections*. The response times for the second-level connections can be found by using (6), where $\Delta_{0,1}, \Delta_{RM}(1), \Delta_{1,2}, \Delta_{RM}(2)$, and $\Delta_{bs}(2)$ are replaced by their corresponding values for different connections.

Following the same approach, the response time for a jth-level connection, whose max-min fairness allocation must rely on the max-min fairness allocation of a $(j-1)$th-level connection, can be found by using (8), (9) or (10), with $\Delta_{i-1,i}, \Delta_{RM}(i)$ and $\Delta_{bs}(j)$ replaced by different corresponding values for different connections.

We now apply this analytical approximation and compare the result to the simulation results presented in the previous section. Here we assume that the network is only slightly congested. Since a VC is said to be active only when the switch receives the first cell from the connection, $\Delta_{0,1}$ is the time the first RM cell must take before reaching the switch. This includes the propagation delays, transmission delays and the queueing delays experienced by the first RM cell of the starting VC before it reaches the first bottleneck switch. $\Delta_{sb}(i)$ includes the propagation delays, transmission delays and the queueing delay at different switches. As in the simulation, we assume that there is no other traffic in the backward path, $\Delta_{bs}(i)$ is thus equal to the sum of the propagation delay and the transmission delay.

Tables 7 and 8 show the estimated response time and the peak queue length at different switches. In Table 8, the peak queue length is the maximum value of Q_{length} in (11) throughout the duration of the simulation. When compared to Tables 5 and 6, both Tables 7 and 8 show that the approximation is well within the confidence interval of the simulation results.

6 Max-Min Scheme with Delayed Adjustment (MMDA)

Consider Figure 1 in which there are a set of VCs passing through a bottleneck switch. The delays between the switch and the source end systems (SESs) of VCs a and b are $Delay\_a$ and $Delay\_b$ (where $Delay\_b \ll Delay\_a$), respectively. Suppose there is a change in the network condition such that SES a must decrease its ACR while SES b must increase its ACR so that the max-min fairness can be maintained. Since the response time of VC b is much faster than VC a (because $Delay\_b \ll Delay\_a$) , the switch is overloaded until ACR of SES a is decreased. This may lead to buffer overflow which causes cell loss at some switches. This in turns creates a "snow-ball" effect since the retransmission of packet as a result of cell loss would make the situation worse because a single cell loss requires the retransmission of the entire packet.

6.1 Basic Idea

The proposed Max-Min scheme with Delayed Adjustment (MMDA) is based on the Max-Min scheme. Since it is sometimes necessary to carry the delay information back to the sources, a new field in the RM cell, called *delay* is proposed. Similar to other time-related parameters, it is 3 bytes long [AF95-0554].

During call setup time, every switch participates in calculating the round trip time RTT and this RTT is used to determine other parameters. As mentioned in [AF95-0554], all the switches of the connection calculate the RM_Delay and add these to the estimation of RTT, where RM_Delay is the delay that RM cells could experience in the link and node at 99% confidence interval. Therefore, after the switch receives the partial result on RTT, it knows the delay the RM cell could experience when traveling from the source to the current node. At the end of the call setup, every switch along the connections knows the delay between itself and the source of that VC with good accuracy.

Each switch maintains an information table (Table 9), which is similar to the information table for the Max-Min scheme, for all VCs that pass through it. D_{sn} is the delay between the source and the node calculated during call setup time. Other entries have the same meaning as in the Max-Min scheme.

Basically, the forward RM cell will undergo the same procedure as in the Max-Min scheme (Section 3). However, the operations will be slightly different in the backward direction. When the switch receives a backward RM cell with ER field equal to ER_RM, the switch will do the following after step 8:

8 a For all connections $j \in J$, where J is the set of VCs with CAs that need to be changed,
IF CAs for some connections increase and for some decrease THEN

$$delay(j) = 2 \cdot \{\max_{i \in J}(D_{sn}(i)) - D_{sn}(j)\}.$$

Here $delay(j)$ is the value to be filled in the *delay* field of the return RM cell of VC j. When the source receives a RM cell, its ACR is set to the ER value after the delay time suggested in the *delay* field of the RM cell has elapsed.

7 Performance of the MMDA Scheme

In this section, the performance of the MMDA scheme is compared to the Max-Min scheme through simulations. Figure 5 shows the simulation model which is similar to that used in the Max-Min scheme. The only difference is the variation of $delay\_4$, the access delay for VC 4. The SES used is based on [AF95-0013] again. The one-way propagation delay between two switches is $250\mu s$ (MAN separation suggested in [AF94-0394]). The access delays for VCs 3 and 4 are $delay\_3$ and $delay\_4$ ($\gg delay\_3$), respectively. Let us define $d = delay\_4 - delay\_3$.

Initially, only VCs 1, 2, 3 and 4 are active. VC 5 becomes active at time=$0.02s$. The fair share allocation for the VCs before and after VC 5 turns active is shown in Table 10. Notice that the allocation for VC 3 increases while that of VC 4 decreases. Figure 2 shows the peak queue lengths at switch 2 (the bottleneck for VCs 3 and 4) for the Max-Min scheme and the MMDA scheme. It shows that there is a significant reduction in peak queue length and the reduction increases approximately linearly with the difference in delay.

8 Conclusion

In this paper, two new rate-based switch algorithms are proposed. The first one, called Max-Min Scheme, can quickly converge to the max-min fairness allocation, the fairness measure agreed by the ATM Forum. With this approach, the efficiency of the network can be maximized, which is verified by simulations. It is found that the response times of the VCs to converge to their max-min fairness allocations are the shortest when compared to both CAPC and ERICA. Because of the fast response times, the peak queue lengths built up at the bottleneck switches are also minimized.

Analytical approximations to calculate the response times and peak queue lengths are also introduced. The estimated values are compared to the simulation results. It is found that the estimated values are well within the confidence intervals of the simulation results.

To tackle the problem of different source-to-switch separations of different connections, another rate-based switch algorithm, called Max-Min scheme with Delayed Adjustment (MMDA) is proposed. The additional feature is to consider the switch-to-source delays, which can be obtained during call setup time. Under some situations, when the source receives an RM cell, it must adjust its ACR after some delay. With this, the amount of traffic going into the bottleneck switch can be maintained at the desired output rate. Simulation results show that overloading of the bottleneck switch can be avoided and a further reduction of peak queue length can also be achieved.

References

[AF94-0394] Lou Wojnaroski, " Baseline Text for Traffic Management Sub-Working Group," ATM Forum 94-0394R5.

[AF94-0882] R. Jain, et al, "Rate Based Schemes: Mistakes to Avoid," ATM Forum 94-0882.

[AF94-0883] R. Jain, et al, "The OSU Scheme for Congestion Avoidance Using Explicit Rate Indication," ATM Forum 94-0883.

[AF94-0983] A. W. Barnhart, "Explicit Rate Performance Evaluation," ATM Forum 94-0983R1.

[AF95-0013] S. Sathaye, "ATM Forum Traffic Management Specification Version 4," ATM Forum 95-0013.

[AF95-0178] R. Jain, S. Kalyanaraman, R Viswanathan and R. Goyal, "A Sample Switch Algorithm," ATM Forum 95-0178R1.

[AF95-0395] Y. Chang, N. Golmie and D. Su, "Comparative Analysis of the Evolving End System Behavior (Simulation Study)," ATM Forum 95-0395R1.

[AF95-0554] Lawrence G. Roberts, "Parameter and Vector Selection for ABR," ATM Forum 95-0554R3.

[Bart87] D. Bartsekas and R. Gallager, "Data Networks," Prentice Hall, 2nd Edition, 1987.

[Bono95] F. Bonomi and K. W. Fendick, "The Rate-Based Flow Control Framework for the Available Bit Rate ATM Service," IEEE Network Magazine, March/April 1995, page 25-39.

[COMD93] COMDISCO System Inc.,"BONeS DESIGNER Core Library Guide," June 1993.

[DT95] Wales Kin Fai Wong and Danny H. K. Tsang, "A Rate-Based Switch Algorithm with Delay Adjustment for ABR Traffic to Achieve Max-Min Fairness," IEEE ATM Workshop '95, October 30 - November 1, 1995, W4B-2.

[DT96] Danny H. K. Tsang, Wales Kin Fai Wong, Sheng Ming Jiang and Eric Y. S. Liu, "A Fast Switch Algorithm for ABR Traffic to achieve Max-Min Fairness," to appear in IEEE 1996 International Zurich Seminar on Digital Communications, February 19-23, 1996.

[Jain95] R. Jain, "Congestion Control and Traffic Management in ATM Networks: Recent Advances and A Survey," invited submission to Computer Networks and ISDN Systems.

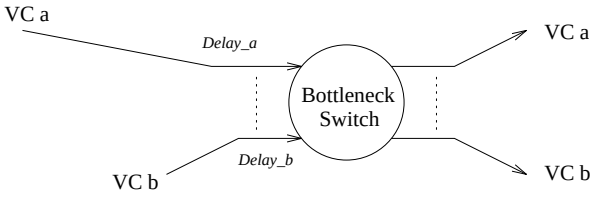


Figure 1: Example network

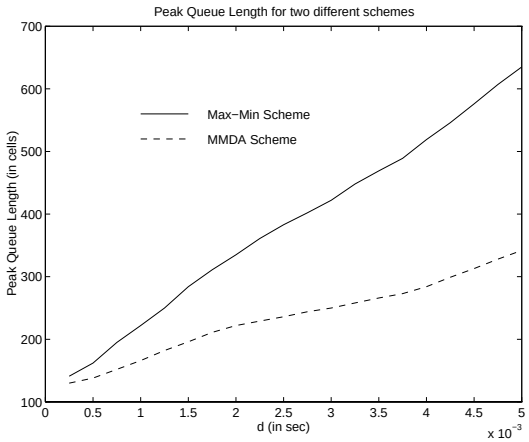


Figure 2: Comparison of peak queue length of switch 2

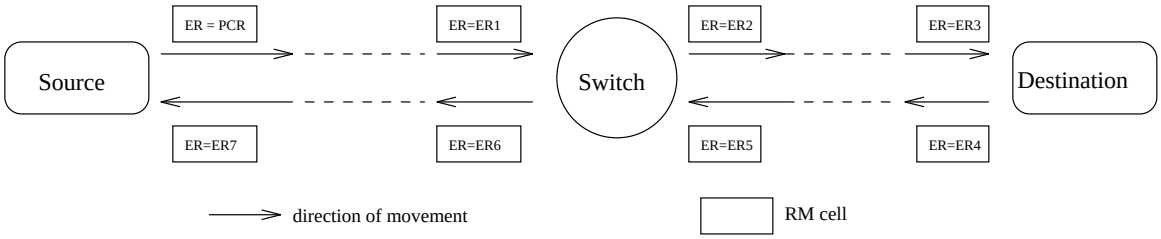
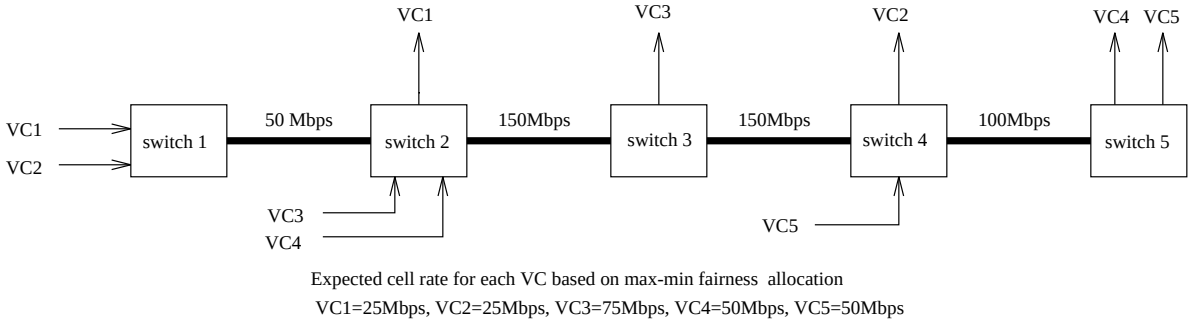


Figure 3: Flow of RM cells



Expected cell rate for each VC based on max-min fairness allocation
VC1=25Mbps, VC2=25Mbps, VC3=75Mbps, VC4=50Mbps, VC5=50Mbps

Figure 4: Simulation Model for the Max-Min Scheme

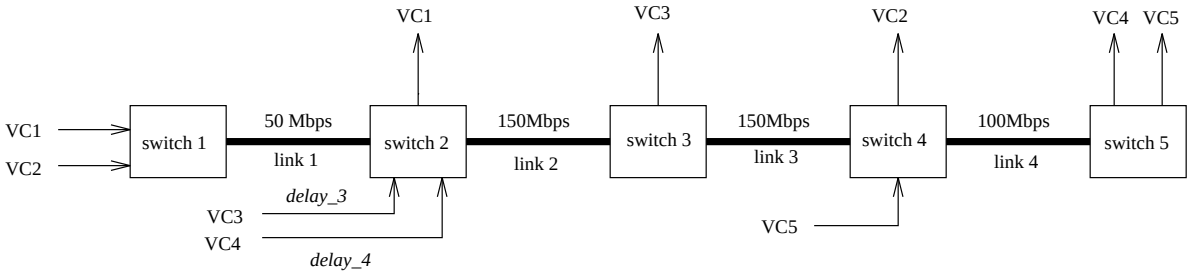


Figure 5: Simulation Model for the Max-Min Scheme and the MMDA Scheme

Table 1: Information Table for the Max-Min Scheme

VCI	*ER_f*	*ER_b*	*CA*	*constrained*
x	f1	b1	c1	0/1
y	f2	b2	c2	0/1

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

Table 4: Transient Response Time in μs 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 5: Transient Response Time in μs for the 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 6: 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

Table 7: Estimated Response Time in μs for the Max-Min Scheme

	ACR1	*ACR2*	*ACR3*	*ACR4*	*ACR5*
when VC1 is active	134	N/A	N/A	N/A	N/A
when VC2 is active	151.3	409.8	N/A	N/A	N/A
when VC3 is active	0	0	128	N/A	N/A
when VC4 is active	0	0	69.9	335.5	N/A
when VC5 is active	0	0	332.7	224.2	129.9

Table 8: Estimated Peak Queue Lengths in Cells

Switch 1	*Switch 2*	*Switch 3*	*Switch 4*
66.9	22.6	2	21.1

Table 9: Information Table for the MMDA Scheme

VCI	*ER_f*	*ER_b*	*CA*	*constrained*	D_{sn}
x	f1	b1	c1	0/1	D1
y	f2	b2	c2	0/1	D2

Table 10: Max-Min Fairness Allocation in Mbps

	VC1	*VC2*	*VC3*	*VC4*	*VC5*
Before VC5 becomes active	25	25	62.5	62.5	N/A
After VC5 becomes active	25	25	75	50	50