Switch Mechanisms for the ABR Service: A Comparison Study

Dorgham Sisalem
Henning Schulzrinne

GMD FOKUS
Hardenberg Platz 2
10263 Berlin
Germany
Tél. : +49 30 25499 170
Fax : +49 30 25499 202
E-mail : hgs,dor@fokus.gmd.de

ABSTRACT. *While lots of effort is done on behalf of the ATM Forum and the ITU standardisation bodies in specifying the user network interface (UNI) for ATM's available bit rate (ABR) service the specification of the switch mechanisms will be left to the manufacturers. In this study some of the algorithms proposed for dynamically distributing the available bandwidth at the switches such as EPRCA, CAPC, the OSU scheme and a modified version of the MIT proposal are compared. In our investigations we will concentrate mainly on the fairness behaviour of the schemes, their complexity and their warm up periods.*

The work presented here is a description of our preliminary results. At this stage, we have mainly considered the fairness behaviour of the different schemes using a single test scenario.

MOTS CLÉS : *ATM; ABR; EPRCA; CAPC; MIT plan; OSU plan; Équité.*

KEYWORDS : *ATM; ABR; EPRCA; CAPC; MIT scheme; OSU scheme; Fairness.*

1 Introduction

Many applications, mainly handling data transfer, have the ability to reduce their sending rate if the network requires them to do so. Likewise, they may wish to increase their sending rate if there is extra bandwidth available within the network. This kind of applications is supported by an ATM layer service called the available bit rate service (ABR). Applications using the ABR service can expect the following quality of service commitments from the network:

1. The available bandwidth is fairly distributed among all active ABR connections.
2. A minimum cell rate that is agreed upon during the connection establishment phase.
3. Only a preset fraction of the sent cells can be dropped as long as the sending behaviour of the application conforms to the negotiated values.

To provide these quality of service requirements appropriate traffic control mechanisms have to be used.

In its September 1994 meeting the ATM Forum voted for a rate based congestion control mechanism to be used in ATM networks. With the chosen algorithm the ABR sources must send a so called resource management (RM) cell every *Nrm* sent data cells. The destination end system turns the RM cells around and sends them back to the source. Depending on the traffic situation in the network the intermediate switches can determine the fair bandwidth shares the ABR connections should use in order to avoid congestion. These values are then written into the backward RM cells in the explicit rate (ER) field. In accordance with the explicit rate noted in the RM cells the source can then increase or decrease its rate.

While lots of effort has been done in specifying the user network interface (UNI) specifying the switch mechanisms will be left to the manufacturers. In this study different switch algorithms are described and compared with each other. In the next section two explicit rate indication schemes and two congestion avoidance mechanisms are shortly introduced. Section 3 describes the used simulation environment. Simulation results showing the fairness behaviour of the different schemes are then presented in section 4.

2 Congestion Control and Avoidance in ATM Networks

In this study four switch mechanisms are investigated and compared with each other. The first two, the EPRCA and the MIT schemes, aim at estimating the fair bandwidth shares of the active connections based upon the information in the RM cells. The OSU as well as the CAPC schemes are more of congestion avoidance schemes that aim at driving the operation of the network toward the knee of the delay curve [1]. That is, these schemes aim at achieving an optimal overall performance with high bandwidth utilisation and low buffer usage and delay.

2.1 The Enhanced Proportional Rate Control Algorithm (EPRCA)

The EPRCA [2] scheme computes a mean allowed cell rate (MACR) using exponential weighted averaging. The fair bandwidth share of each connection is then set to a fraction of this calculated MACR if necessary, i.e., if the current cell rate indicated in the RM cells is larger than the fair share.

$$\begin{aligned} \text{MACR} &= (1-\alpha)\text{MACR} + \alpha\text{CCR} \\ \text{Fair Share} &= \text{SW\_DPF} * \text{MACR} \end{aligned}$$

Here, α is the exponential averaging factor, CCR is the current cell rate of the connection noted in the RM cells and SW_DPF is a multiplier called switch down pressure factor set close to but below 1. Various simulation studies [3] suggest that the appropriate values for α and SW_DPF are $1/6$ and $7/8$ respectively [4].

As a congestion indication this scheme uses the queue length. Whenever a preset threshold is exceeded the switch is said to be congested. However, this method was shown to result in unfairness [3]. This problem was then fixed by using the queue growth rate as the load indicator. With the method suggested by Siu and Tzeng [5] a switch is said to be congested if a queue growth is noticed after processing k cells. A threshold is used only to indicate the very congested state in which the sources are asked to rapidly decrease their rates.

2.2 The MIT Scheme

The original MIT scheme was proposed by Charny [6] in her master thesis at the Massachusetts Institute of Technology.

With this scheme the fair share is computed using an iterative procedure. Initially, the fair share is set to the link bandwidth divided by the number of active VCs. All VCs, whose rates are less than the fair share are called ”underloading VCs”. If the number of underloading VCs increases at any iteration, the fair share is recomputed as follows:

$$\text{Fair share} = \frac{\text{link bandwidth} - \sum \text{bandwidth of underloading VCs}}{\text{number of VCs} - \text{number of underloading VCs}}$$

This procedure is repeated until a steady state is reached in which the fair share and the number of underloading VCs does not change [4]. In her work [6] Charny shows that two iterations are usually sufficient enough to reach convergence.

While this proposal fullfills the fairness requirements of the switch mechanisms it requires for the fair share computation order n operations with n as the number of VCs for each received RM cell. For ATM switches supporting thousands of VCs the computational overhead becomes extremely large and can not be done by currently available hardware. Kalampoukas et al. [7] present therefore an allocation algorithm that is based on the one introduced by Anna Charny but requires only $\theta(1)$ operations. This is reached by keeping track of the available bandwidth for the ABR connections. For more detailed description of the required computations see [7].

2.3 The OSU Congestion Avoidance Scheme

This scheme was developed by Jain et al. [8] at the Ohio State University (OSU). In this scheme the switches measure their input rate over a fixed averaging interval that is counted in cells [9]. Thereby, the input rate is calculated as follows:

$$\text{Input rate} = \frac{\text{number of received cells}}{\text{length of measurement interval}}$$

With this calculated input rate a load factor can be computed which can then be used to indicate the over- or underload state of the switch.

$$\text{load factor} = \frac{\text{input rate}}{\text{Target rate}}$$

Here, the target rate indicates the rate level at which an optimal network performance can be achieved with high utilisation and low buffer usage. The target rate is usually set to a value between 85% and 95% of the link bandwidth. During the steady state the utilisation should be as high as the target rate and during the transient state with lots of bursty connections the utilisation could be even higher.

The sources are then asked to change their current sending rate inversely proportional to the calculated load factor. That is, with a load factor of 0.5 for example the sources should double their sending rates to avoid an underload situation at the switch.

To assure fairness and a minimal cell rate the active connections must at least receive the minimum fair share that is calculated as follows:

$$\text{Fair share} = \frac{\text{Target rate}}{\text{number of active connections}}$$

2.4 Congestion Avoidance using Proportional Control (CAPC)

Basicly, this algorithm developed by Andy Barnhart from Huges Systems [10] shows a great similarity to the OSU scheme described in sec. 2.3. Just as in the OSU scheme the switches set a target utilisation slightly below 1 and compute the input rate over averaging intervals. With the computed input rate a load factor (z) can then be calculated. The main difference lays in the way the explicit rates are determined.

During underload states $(z < 1)$ the fair share of the ABR connections is increased as follows:

$$\text{Fair share} = \text{Fair share} * \text{Min}(ERU, 1 + (1 - z) * Rup)$$

Here, Rup is a slope parameter in the range of 0.025 to 0.1 and ERU is the maximum increase allowed and is usually set to 1.5.

During overload states $(z > 1)$ the fair share of the ABR connections is increased as follows:

$$\text{Fair share} = \text{Fair share} * \text{Max}(ERF, 1 + (z - 1) * Rdn)$$

Here, Rdn is a slope parameter in the range of 0.2 to 0.8 and ERF is the minimum decrease allowed and is usually set to 0.5.

In addition to the load factor the scheme uses a queue threshold. Whenever the queue length is over this threshold, a congestion indication (CI) bit is set in all RM cells. This prevents all the sources from increasing their rates and allows the queue to drain out more quickly.

During the steady state with $(z = 1)$ this scheme achieves an oscillation-free steady state. As the frequency of the oscillation is a function of $(1 - z)$ having a load factor of 1 yealds an infinite oscillation period.

3 Simulation Environment

For testing the fairness performance of the different schemes we use a chain topology with five switches and 21 connections originating from seven sources, whereas each source is the origin of three similar connections. The distance between two neighbouring switches is 1000 km and the distance from a source to a switch is 0.01 km. All the links have a bandwidth of 45 Mb/s. Depending on the number of switches they have to pass three kinds of traffic can be distinguished:

- Long distance connections: Connections starting from source 0 traverse all four links. These connections will be denoted as VC 0.
- Medium distance connections: Connections starting from sources 1 and 4 traverse two links. Theses connections will be denoted as VC 1 and VC 4.
- Short distance connections: Connections starting from sources 2, 3, 5 and 6 traverse only one link. Theses connections will be denoted as VC 2, VC 3 and VC 5.

For the fairness investigations we used persistent sources. These are greedy sources that can always send with the maximum allowed cell rate and impose thereby the most stringent constraints on the network.

4 Simulation Results of the Fairness Performance

As a fairness criterion we used the max-min fairness definition as it was described in the ATM Forum traffic management specification 4.0 [11]. With this definition a connection's bandwidth request is either satisfied by the switch or set to the fair share calculated by the switch. This fair share is determined as follows:

$$\text{Fair share} = \frac{\text{link bandwidth} - \sum \text{bandwidth of satisfied VCs}}{\text{number of VCs} - \text{number of satisfied VCs}}$$

Using the max-min criterion to determine the appropriate bandwidth distribution we would get the bandwidth distribution presented in Tab. 1. For the fairness estimation we used the fairness index suggested by Jain [12] and defined as follows:

$$F = \frac{(\sum x_i)^2}{n \sum x_i^2} \quad \text{with } x_i = \text{ratio of the actual throughput to the fair throughput}$$

$$\text{and } n = \text{the number of connections}$$

For the fairness discussion we have investigated the case when the connections start with an initial cell rate (ICR) of (PCR/20) with PCR as the peak cell rate or the link rate.

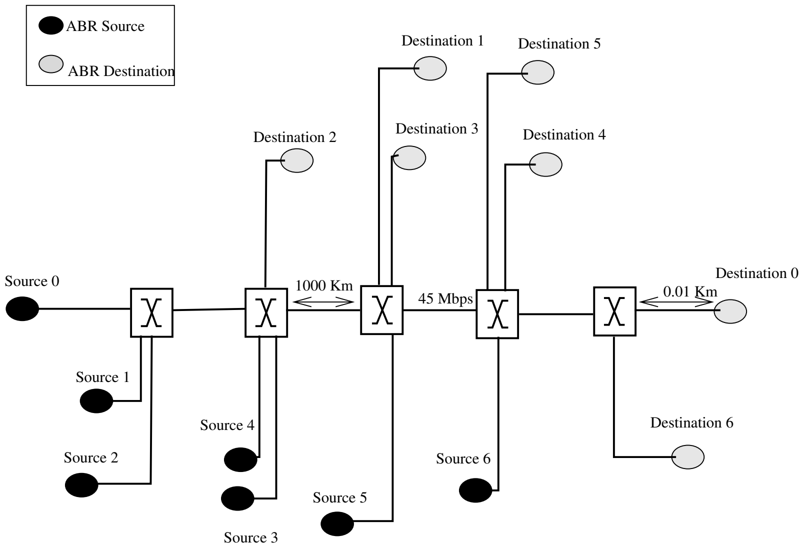

Figure 1: Test topology

VC 0	VC 1	VC 2	VC 3	VC 4	VC 5	VC 6
3.75	3.75	7.5	3.75	3.75	7.5	11.25

Table 1: Expected fair bandwidth share for the single VCs in Mb/s

4.1 The OSU Scheme

Here, we used a target rate of 95% of the available link bandwidth and would therefore expect to achieve a bandwidth distribution of only 95% of the values depicted in Tab. 1.

The results achieved for the OSU scheme depicted in Tab. 2 reveal some serious drawbacks of the algorithm. While the average rate achieved for a complete path, i.e., the average rate achieved by the three connections traversing the same path, is very close to our expectations the values of the single VCs vary strongly. This is of course surprising as all of the VCs traversing the same path are identical. Investigating the behaviour of the switches more closely revealed the following: the switches update the explicit rate of all RM cells that arrive during the same measurement period by the same load factor. So, if it happens that during a measurement interval RM cells from only a part of the VCs traversing a common path arrived at the switch the rate of only these connections will be adjusted with the current load factor. The rate of the other connections traversing this path will only be updated in one of the next measurement intervals with another load factor. This results then in different allowed rates at the sources. This problem can be solved by using larger measurement intervals. However, this reduces the reactivity of the algorithm and leads to longer transient periods. Actually, in our tests we already used a measurement interval of 60 cells which is already twice as large as the suggested value in [9].

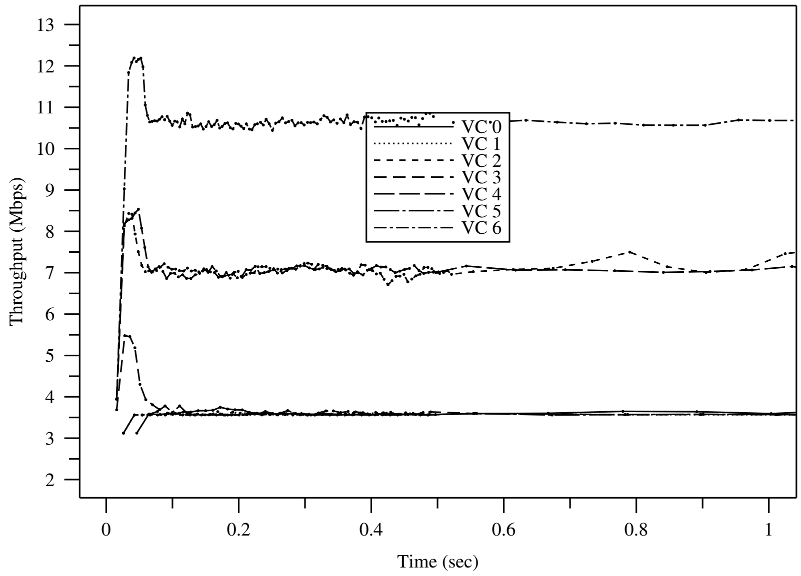

Figure 2: Average path throughput achieved with the OSU scheme

Another problem with the OSU scheme is its unfairness towards short distance traffic during overload periods. We have mentioned that the switches write the fair rate share in the backward RM cells. While this ensures that the rate is updated using the most current load factor it leads to some unfairness in the case of connections with different round trip times. As all connections contribute to the switch congestion in the same way we would expect the switch to reduce the rate of all connections using the same overload factor. However, due to the short round trip times of the short distance connections their RM cells arrive back at the switch much earlier than those of the long distance connections. Thereby, these sources will be forced to reduce their bandwidth with a very high overload factor. As the short distance connections reduce their rates the load factor will be reduced as well. With the smaller load factor the explicit rates of the later arriving RM cells of the long distance connections will be less reduced than was the case for the short distance traffic. In short, this means that the long distance connections can benefit longer from a high cell rate and utilise a bandwidth share larger than their fair share. This has also the negative effect of lengthening the wrap up period. On the other hand, during underload phases the short distance connections can increase their rates faster with a smaller overload factor. That is, in this situation the short distance connections can benefit from the small factor, increase their rate and thereby increase the load factor. When the RM cells of the long distance connections arrive after the longer round trip delay their rates can only be changed by the larger factor and the rate of the long distance connections can thereby only be increased slower than the rate changes of the short distance ones.

Connection	VC 0	VC 1	VC 2	VC 3	VC 4	VC 5	VC 6	F
1	3.63	3.57	7.5	3.65	3.58	6.06	8.8	
2	3.58	3.59	7.67	3.68	3.58	6.1	11.42	
3	3.60	3.57	6.2	3.66	3.59	9.22	11.92	
average	3.60	3.57	7.12	3.66	3.58	7.138	10.71	0.99

Table 2: Achieved bandwidth in Mb/s for the single VCs using the OSU scheme

4.2 The CAPC Scheme

With all of the sources starting with an initial cell rate (ICR) of only PCR/20 CAPC results in an ideal behaviour. All the VCs traversing a common path receive a similar bandwidth share, the bandwidth is fairly distributed and we can only notice a small transient period.

For the case of an initial cell rate as high as the peak cell rate we notice, however, a bug in the algorithm as it was described in [10]. Due to the long lasting congestion at the switches caused by the high initial cell rate the fair share value is reduced several times until its finally represented by 0. In this case the fair share value will remain 0 even if all the sources reduced their rates. A possible solution to this problem is to add a minimum threshold value for the fair share. This minimum threshold could be set as the available bandwidth divided by the number of connections seen in the last measurement interval. This might be combined with the setting of the congestion indication (CI) bit to force a rapid rate reduction at the sources. With this enhancement the switch would set the fair rate share during overload periods as follows:

$$\text{Fair share} = \text{Max}(\text{Fair share} * \text{Max}(ERF, 1 + (z-1) * Rdn), \frac{\text{bandwidth}}{\text{number of connections}})$$

Here, Rdn is a slope parameter in the range of 0.2 to 0.8 and ERF is the minimum decrease allowed and is usually set to 0.5.

Connection	VC 0	VC 1	VC 2	VC 3	VC 4	VC 5	VC 6	F
1	3.5	3.6	7.2	3.6	3.6	7.2	10.8	
2	3.6	3.6	7.2	3.6	3.6	7.2	10.7	
3	3.5	3.6	7.2	3.6	3.6	7.2	10.8	
average	3.5	3.6	7.2	3.6	3.6	7.2	10.8	0.99

Table 3: Achieved bandwidth share in Mb/s for the single VCs using CAPC

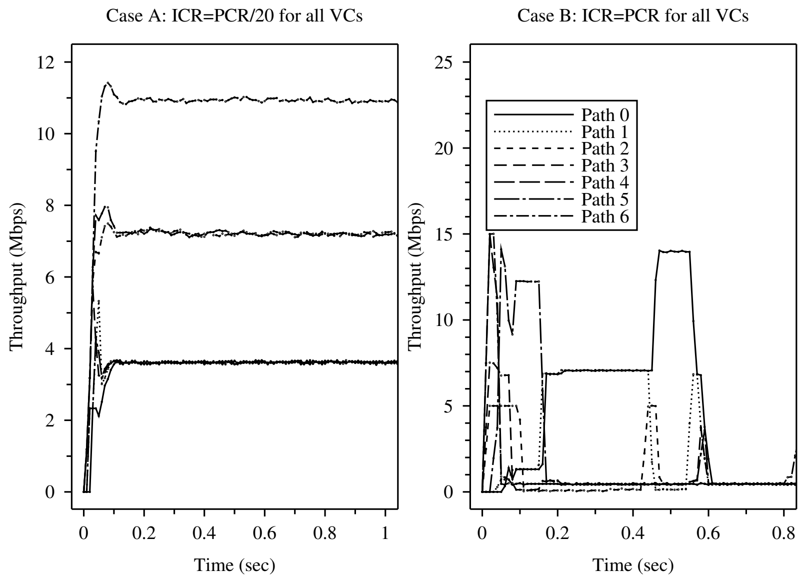

Figure 3: Average path throughput achieved with the CAPC scheme

As the CAPC scheme updates the received RM cells each measurement interval with the newly calculated load factor the same unfairness noticed with the OSU scheme for the case of connections with different round trip times applies here as well.

4.3 EPRCA

EPRCA estimates the fair bandwidth share of a connection by using the information in the RM cells. This results in very simple switches without the need of maintaining lots of information about the connections at the switch as was the case with the OSU scheme. However, the results depicted in Tab. 4 and Fig. 4 show that on the contrary to the bandwidth distribution achieved with the previous schemes, EPRCA results in a very oscillating distribution. Also, the algorithm requires several iterations until reaching a steady state resulting in a long transient period.

Connection	VC 0	VC 1	VC 2	VC 3	VC 4	VC 5	VC 6	F
1	3.2	3.7	7	3.6	4.1	7.7	11.5	
2	3.3	3.7	7	3.6	4.1	7.5	11.4	
3	3.5	3.8	7	3.8	3.9	7.5	11.6	
average	3.3	3.7	7	3.7	4	7.6	11.6	0.99

Table 4: Achieved bandwidth share in Mb/s for the single VCs using EPRCA

Another negative aspect of this scheme is its dependence on the used factors. The performance of the algorithm can be altered simply by using a different averaging factor. With a larger α MACR follows the changes in the connections rates more quickly. While this leads to faster responses it causes an even more oscillating bandwidth distribution. Using a smaller α results in a more steady behaviour but with a longer transient period.

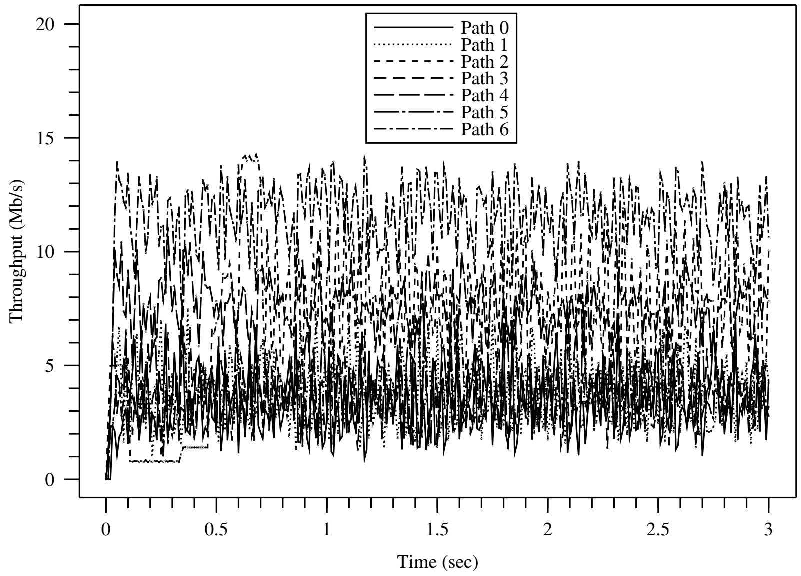

Figure 4: Average path throughput achieved with the EPRCA scheme

4.4 The MIT Scheme

Even though the MIT based scheme that was used here requires only $\theta(1)$ operations per received RM cell it is still more complicated than the previous algorithms described in this study, see sections 2.3, 2.4 and 2.1. However this added complexity results in a very fair bandwidth distribution, a stable steady state and high utilisation, see Fig. 5 and Tab. 5. Also, while this scheme requires more arithmetical operations most of the needed operations are simple addition and comparison operations that do not cause a high computational overhead.

This scheme as well the EPRCA scheme also solve the unfairness problems noticed in the OSU and CAPC schemes during the transient periods. The rates are not updated proportional to the CCR value noted in the RM cells. Instead, the switches write the calculated fair share directly in the RM cells. So, as soon as the switch has enough information about the connections and can determine the appropriate fair share all the connections will be updated in the same way. However, for a clearer picture about the fairness performance during transient periods more simulations using bursty sources and dynamically changing traffic patterns are needed.

Connection	VC 0	VC 1	VC 2	VC 3	VC 4	VC 5	VC 6	F
1	3.4	3.6	7.3	3.8	3.7	7.5	11.2	
2	3.4	3.6	7.3	3.8	3.7	7.4	11.2	
3	3.4	3.6	7.3	3.8	3.7	7.5	11.2	
average	3.4	3.6	7.3	3.9	3.7	7.5	11.2	0.99

Table 5: Achieved bandwidth share in Mb/s for the single VCs using the MIT scheme

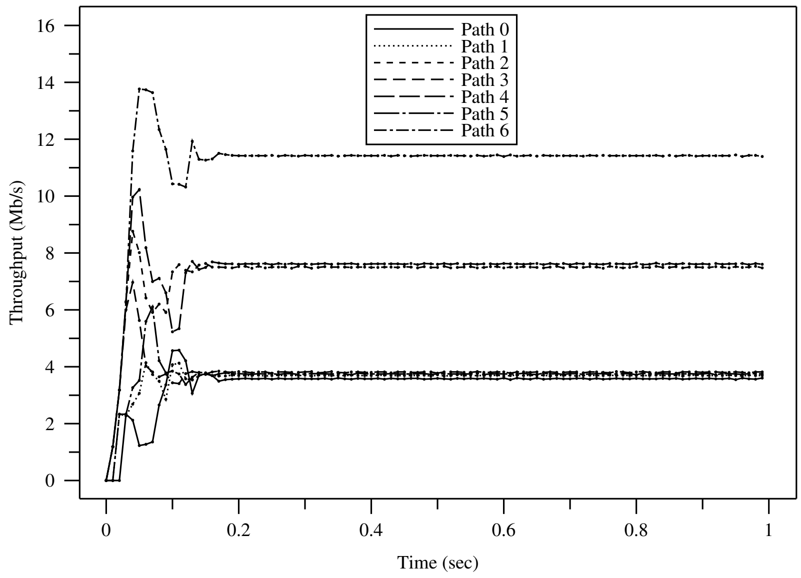

Figure 5: Average path throughput achieved with the modified MIT scheme

5 Summary and Future Work

This study is still work under progress. Until now only the fairness behaviour of the different algorithms was investigated using persistent sources. This gave us a first impression on the efficiency of the different schemes. For a more complete picture further investigations under varying loads and with different sources are required.

These preliminary results suggest that while congestion avoidance schemes are simpler they still have some problem that need further investigations. So, while determining the appropriate fair share directly might be more complicated it does seem to result in a better performance as we have seen in sec 4.4.

Further, the buffer requirements of the different schemes have to be examined especially under changing traffic conditions and bursty connections.

References

[1] K. K. Ramakrishnan and R. Jain, "A binary feedback scheme for congestion avoidance in computer networks with a connectionless network layer," in *SIGCOMM Symposium on Communications Architectures and Protocols*, (Stanford, California), pp. 303–313, ACM, Aug. 1988.

[2] A. Roberts, "Enhanced PRCA proportional rate control algorithm," Tech. Rep. 94-0735R1, ATM Forum, Aug. 1994.

[3] Y. Chang, N. Glomie, L. Benmohamed, and D. Su, "Simulation study on the new rate-based PRCA traffic management mechanism," Tech. Rep. 94-0809, ATM Forum, Sept. 1994.

[4] R. Jain, "Congestion control and traffic management in ATM networks: Recent advances and a survey." A PostScript version is available at http://www.cis.ohio-state.edu/ jain/papers.html, Jan. 1995.

[5] K.-Y. Siu and H.-Y. Tzeng, "Intelligent congestion control for ABR service in ATM networks," *ACM Computer Communication Review*, vol. 24, Oct. 1994.

[6] A. Charny, "An algorithm for rate allocation in a packet-switching network with feedback," Tech. Rep. MIT/TR-601, MIT, Cambridge, Massachusetts, May 1994.

[7] L. Kalampoukas, A. Varma, and K. K. Ramakrishnan, "An efficient rate allocation algorithm for ATM networks providing max-min fairness," in *HPN'95*, (Palma de Mallorca, Spain), IFIP, Sept. 1995.

[8] R. Jain, S. Kalyanaraman, and R. Viswanthan, "The OSU scheme for congestion avoidance using explicit rate indication," Tech. Rep. 94-883, ATM Forum, Sept. 1994.

[9] R. Jain, S. Kalyanaraman, R. Viswanthan, and R. Goyal, "A sample switch algorithm," Tech. Rep. 95-178R1, ATM Forum, Feb. 1995.

[10] A. W. Barnhart, "Explicit rate performance evaluations," Tech. Rep. 94-0983, ATM Forum, Oct. 1994.

[11] S. S. Shirish, "ATM Forum traffic management specification version 4.0," Tech. Rep. 94-0013R6, ATM Forum, June 1995.

[12] R. Jain, "Fairness: How to measure quantitatively?," Tech. Rep. 94-0881, ATM Forum, Sept. 1994.