

Performance of Explicit Rate Flow Control in ATM Networks

Dr. Lawrence G. Roberts

Abstract

For the last 26 years packet networks have insured low data loss for bursty data traffic with flow control which the author introduced in the ARPANET in 1969 and as a standard in X.25 in 1975. However, from the time ATM was introduced commercially around 1992 until now, flow control has been totally missing. The credit flow control used historically in packet networks was far too expensive at ATM speeds and no alternative had been developed. Without flow control, ATM has been no more effective than a high speed circuit switch. Over the past two years the ATM Forum has studied and refined the authors proposal for explicit rate flow control. This recommendation which will be released in early 1996 will finally permit ATM to carry data traffic with high efficiency and with extremely low loss. This paper examines the performance of various possible versions of rate flow control permissible under this recommendation.

INTRODUCTION

The goal of a flow control technique is to control the source transmission in such a way as to maximize the network utilization while minimizing the size of queues in the network, network delays, burst transmission time, and data loss. As the speed of the data flows increase, given a constant speed of light, the difficulty of achieving these combined goals increases dramatically. For optimal performance where the source is allowed to start up from idle and send at maximum rate, the volume of data in the pipe before a remote switch can possibly detect the event and return control information is the pipe speed times the round trip delay. For cross country operation at OC-3 speed (155 Mbps) this amounts to 1.7 Mbytes of data, all of which must be absorbed or buffered by the network, even with theoretically perfect flow control. This is already a very large amount of data to buffer even with the ideal control system. Many of the control techniques studied required a substantial multiple of this buffer requirement so as to be uneconomic to build today. Thus, the control method must be extremely rapid and converge the source to the right rate very quickly in order to achieve the stated goals of highly responsive WAN operation. The explicit rate technique was originally proposed to the Forum by the author in mid 1994. This explicit rate method has been refined and adapted by the ATM Forum over the last two years. It fully accomplishes the responsiveness goal and is extremely close to theoretically optimal.

COMPARISON OF TECHNIQUES

Three different rate control techniques will be compared, examining the network utilization, buffer size, and other important differences. All are permitted as compliant with the ATM Forum's TM 4.0 recommendation, however performance differ widely.

BINARY RATE CONTROL (EFCI)

To be backward compatible with UNI 3.1 the ATM Forum permits binary rate control switches under TM 4.0. Most of the current generation of ATM switches set the EFCI bit when they are congested on forward data cells and this is their only flow control mechanism. When the EFCI bit reaches the VC's destination, it is turned around by marking a bit in a Resource Management (RM) cell (which circulate every 32nd cell). When the RM cell reaches the source, the rate is either increased slightly, or if there was congestion, is decreased slightly. Thus, if any switch in the forward data path is in congestion, this is fed back to decrease the source rate. The result is a continual oscillation of the source rate around the congestion point of the network.

Another form of binary control is possible where the switches mark the backward flowing RM cells directly, rather than marking the EFCI bit. Except for decreasing the delay from congestion to source rate change, this would work the same as EFCI bit marking. For the simulations shown below, the results would be identical since the switch with the bottleneck was near the VC destination.

Performance of Explicit Rate Flow Control in ATM Networks

Dr. Lawrence G. Roberts

The performance of a network with one or more binary rate switch's in a VC's path is strongly impacted by the Round Trip Time (RTT) along the VC's path. For a LAN network with an RTT below two milliseconds (a few miles in distance plus 600 cells queuing delay), the network utilization, delay and queue sizes are in the acceptable range for general computing. A simulation of an EFCI network with a RTT of 2 milliseconds is shown in Figure 1.

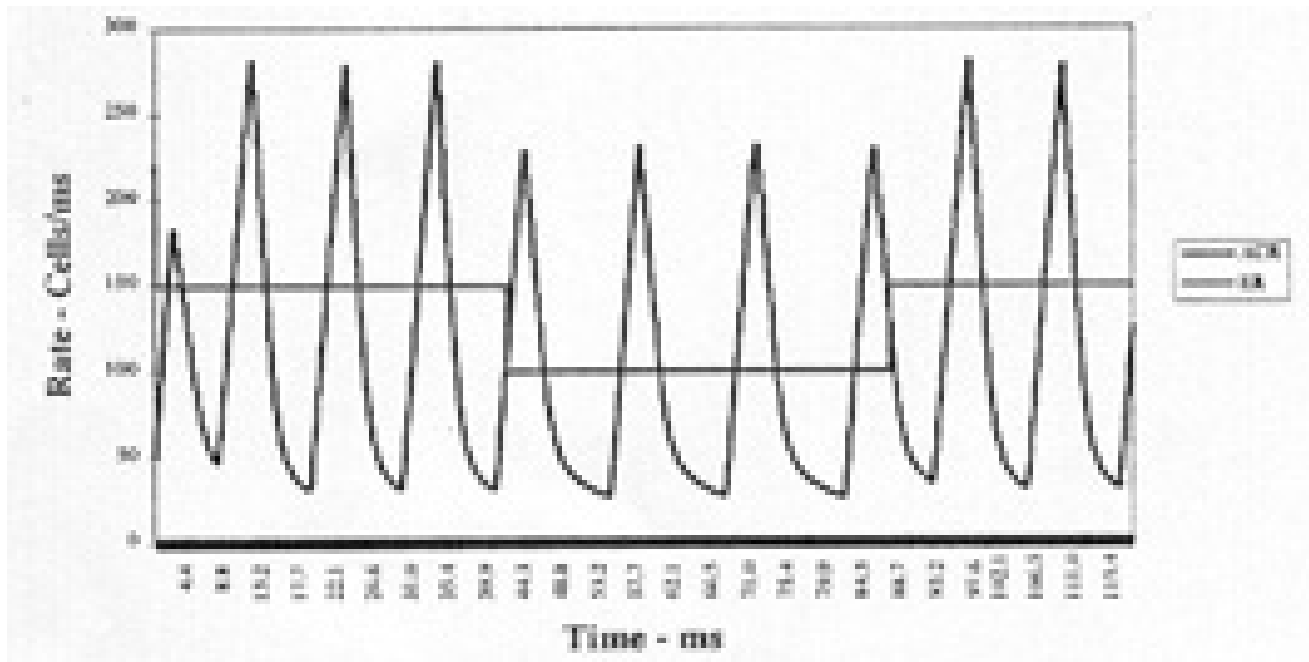


Figure1. Binary Rate Network

RTT=2 ms, TBE=100, RDF=I/16, AIR=I/64

Network Utilization = 83%

Max Queue Size = 332 cells

Average Queue Size= 104 cells

Block Transfer time = 34 ms

As can be seen in Figure 1, the rate oscillates with an amplitude of 166% of the average rate and a period of seven round trips. Already at 2 ms RTT, the utilization is down 17% and the buffer size is up to 332 cells. For larger RTT's these factors get progressively worse because as the oscillations get longer and longer and the period of congestion (queue buildup) and non-congestion (under-utilization of the network) get longer proportionately. Thus, using a binary rate switch outside the LAN area will most likely lead to major data loss as buffers overflow and serious data delays due to low idle startup rates and large queuing delays.

Binary switches also suffer from a very serious fairness problem arising from their not computing the average rate each VC must be controlled to at each switch. Thus, each switch marks all VC's as congested indiscriminately, whereas it should only mark those VC's which are operating at a rate equal to or higher than the local congestion rate. Those VC's which are already at a lower rate than the local congestion rate are being controlled elsewhere and should not be marked at every node. Each VC should only be marked at the node which is currently the bottleneck for that VC. Otherwise,

Performance of Explicit Rate Flow Control in ATM Networks

Dr. Lawrence G. Roberts

a VC going through six nodes will be marked congested 63/64 of the time, thus getting 3% of the throughput of a VC going through one node. However, once a switch computes this local congestion rate, it can easily mark the RM cells with this rate and then would be an explicit rate switch. Thus, it is unlikely anyone will build a binary switch without the fairness problem.

A third serious problem for binary switches is the oscillations themselves. For real-time traffic, where, for example, a video codec is compressing each video frame based on the then current rate request from the network, the 166% variation in rate will cause major problems for the video. Sometimes the frame may get coded at far too high a rate and sometimes at far too low a rate. The problem is that the rate is never correct, it only averages the correct rate. Thus, binary rate switches will not be suitable for real-time traffic.

Thus, binary rate switches are best suited for local area networks with only one or two switches in tandem, and for data traffic only. In this limited position, the older EFCI switches will do very well and should have no significant problems. They must be isolated from the WAN however, by an explicit rate switch which can turn around the EFCI marking locally before it traverses the long haul lines. This can either be a simple form of Virtual Source/Virtual Destination (VS/VD) where EFCI bits get saved and cleared and the next RM cell going toward the source gets marked with CI=1 indicating that there was congestion, or, at far greater expense, a full VS/VD. Without this protective wall, the EFCI problems would exist for all VC's which went out of the local network.

Iterative Explicit Rate Control

When the ATM Forum needed a detailed algorithm to simulate and compare to credit type networks in September 1994, the author submitted the following iterative technique to the Forum. The simulation results were sufficiently favorable that the Forum voted 10:1 to accept explicit rate control rather than credit control for its recommendation. This simple iterative technique still oscillates and is clearly not as effective as the VC counting method presented later. However, it has no fairness problem, requires dramatically smaller buffers than the binary rate technique, and operates at substantially better network utilization. In comparison to the credit proposals, it was far more simple and economic in the LAN, and permitted fair, economic WAN operation which no single credit proposal permitted. Static credit was altogether uneconomic for the WAN and adaptive credit was proven to be unfair like the binary rate.

The iterative method uses one register in each switch to keep an estimate of the local congestion level. This estimate is computed as the exponential average of the rates of all the VC's which were bottlenecked at this node. The rate of each VC is assumed to be the rate carried in the RM cell in the CCR field. The CCR field is filled in by the source as the rate at which it currently is permitted to send. However, this may not be the real rate the source is sending due to internal source blocking.

We now know, that this rate will almost always be an overstatement of the real rate since very few Network Interface Cards (NIC's) are intending to correct their permitted rate to their actual rate. However, assuming this CCR field to be accurate or that the rate is measured at the switch, the congestion rate estimate is computed as the moving average of the VC rates. Assuming that all the VC rates are eventually adjusted to average the correct rate, then the average of them all will be the correct congestion point estimate. The source sends out RM cells every 32 cells as well as whenever it restarts from idle. Each RM cell has an ER field for switches to mark the rate at which they could support the source operating. The source starts this field out at a maximum value and each congested switch on the return path lowers the ER rate to a value slightly less than its congestion estimate if that is lower than any prior switch has inserted.

Performance of Explicit Rate Flow Control in ATM Networks

Dr. Lawrence G. Roberts

Thus, a switch will let all rates increase until it is congested, and then it will mark this rates down to its estimate. The result is a fast rate oscillation (compared to a binary switch) for each VC at about two times the RTT per cycle with a magnitude of about 100% of the average value. A simulation of the iterative method is shown in Figure 2.

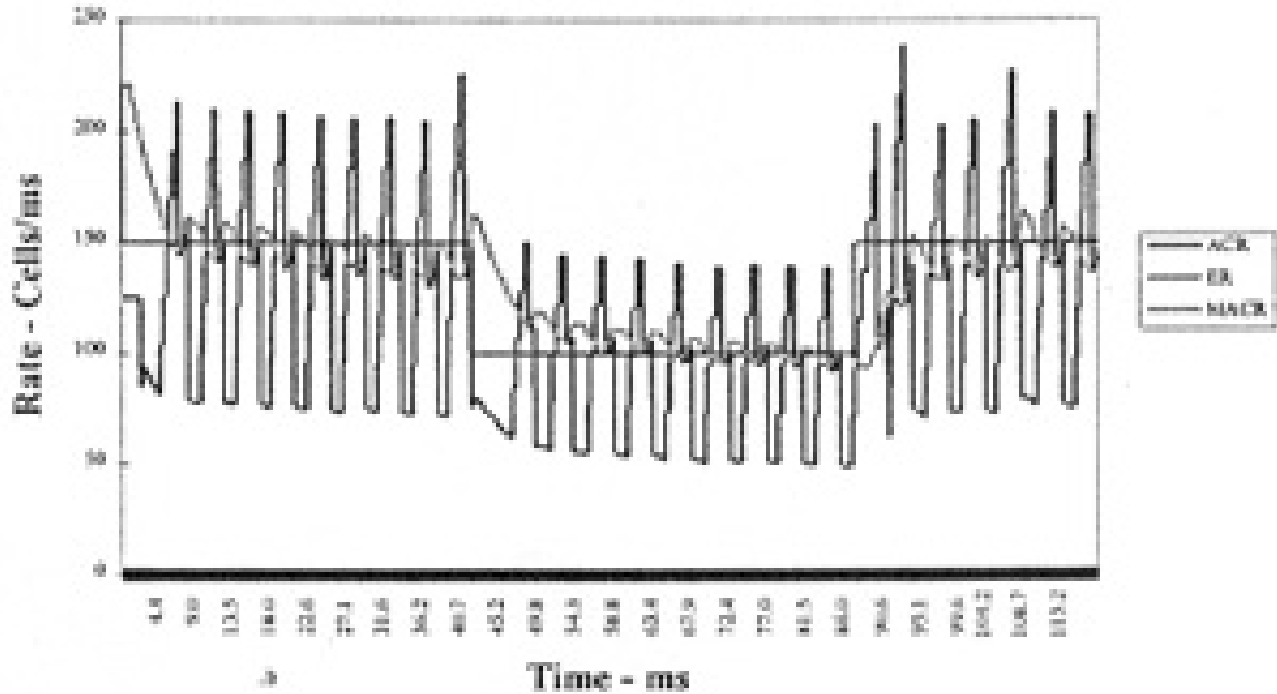


Figure 2. Inerativ Explicit Rate Network

RTT=2 ms, TBE=250, RDF=1/16, AIR=1/16
Network Utilization = 88%
Max Queue Size = 125 cells
Average Queue Size = 20 cells
Block Transfer time = 32 ms

The parameters and simulated network are the same as for the binary rate case except that the Transient Buffer Exposure (TBE) is the full value based on the same buffer size, rather than a reduced one required for a binary network, and AIR is adjusted as appropriate for an explicit rate operation.

The middle trace in the simulation with the small oscillations is the congestion estimate (MACR). It responds slowly to switch load changes (where the square (ER) load switches). As a result, when the estimate is not correct, it tends to create additional queue buildup, or utilization loss.

The iterative rate operation achieves average queue sizes 5 times less than the binary rate network and network utilization 5% higher. This improvement is sufficient to permit the iterative explicit rate switch to be used for WAN operations and still not require uneconomic buffer sizes, suffer cell loss, or result in uneconomic network utilization. Also, as with all explicit rate switches, it is totally fair, not

Performance of Explicit Rate Flow Control in ATM Networks

Dr. Lawrence G. Roberts

suffering from the binary switch fairness problem. It does still have substantial oscillations of the VC rates and these would severely limit real time operation.

VC Counting Explicit Rate Control

If, instead of estimating the congestion rate at a switch, it is computed accurately based on the total available bandwidth capacity for bottlenecked VC's divided by the actual number of bottlenecked VC's which are active, an accurate, up-to-date congestion rate can be computed every millisecond or so. Computing this rate takes considerably more memory cycles than the simple iterative approach, but it provides a considerably superior result. With this accurate congestion rate in hand, the ER field of each backward RM cell can be precisely set to the lower of the rate from the prior switch or the congestion level or, in the case that there is a queue buildup, to a value some percentage (like 10%) below the congestion level. In the latter case, when all VC's are set to at least 10% below the congestion level, then the queue will drain quickly. When it is drained, the rates are all increased to the congestion level until the next queue buildup. Alternatively, the ER's could always be set to 98% of the congestion level, but then queues would take longer to drain and the delays would be greater. Figure 3. shows a simulation of the same case as shown in the prior two ceases.

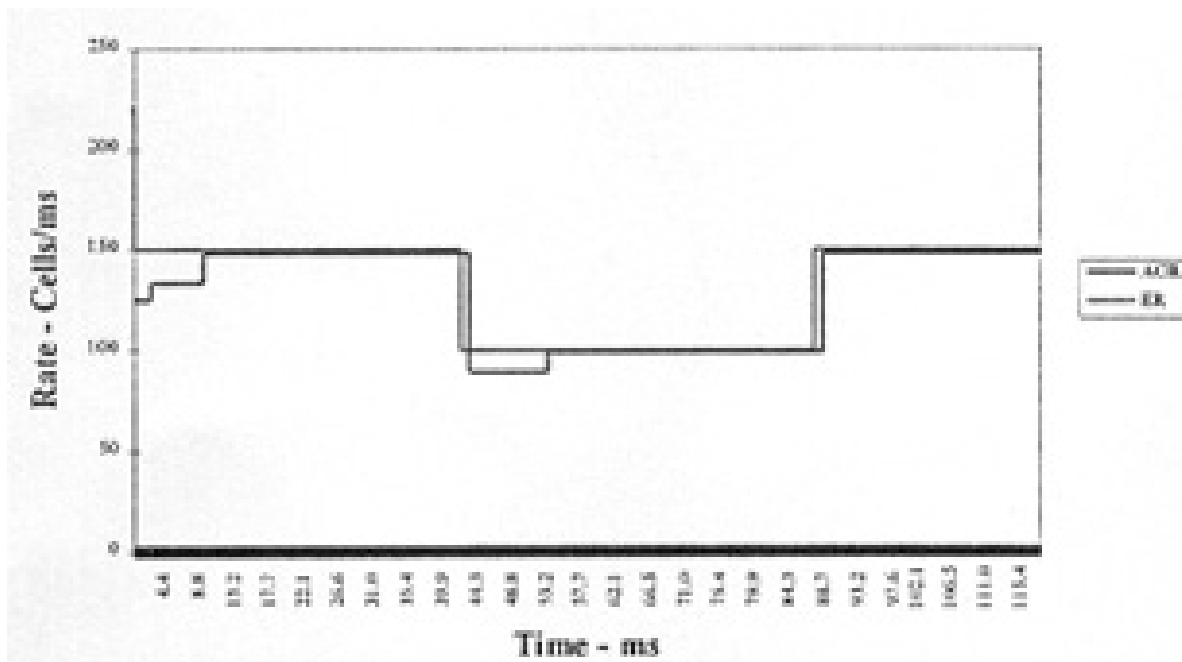


Figure 3. VC Counting Explicit Rate Network

RTT=2 ms, TBE=250, RDF=1/16, AIR=1

Network Utilization = 97%

Max Queue Size = 125 cells

Average Queue Size= 10 cells

Block Transfer time = 29 ms

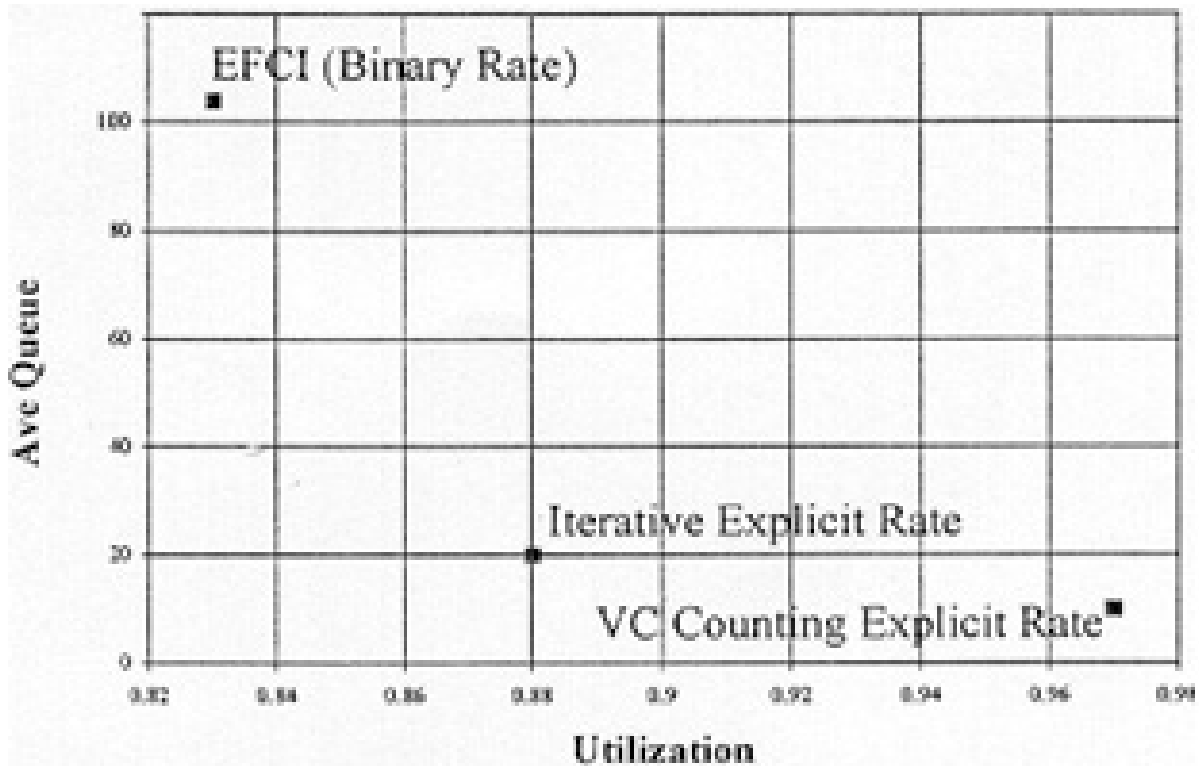
As can be seen in Figure 3, the VC rate (ACR) stays exactly on the switch congestion rate (ER) except for the delays in feedback due to the speed of light and the designed undershoot to clear the queue. The network utilization is 97% and is not significantly affected by the RTT since there is no oscillations. The utilization is 9% better than the iterative method and 14% higher than the binary

Performance of Explicit Rate Flow Control in ATM Networks

Dr. Lawrence G. Roberts

method. Also the average queue size is 50% of the iterative method and 10% of the binary method. These queue improvements are due to the quick feedback of the exact rate without oscillations and substantially improve the ability to support large WAN distances without having to reduce the startup rate.

Figure 4 shows the three cases in the two dimensions of network utilization and average queue size.



CONCLUSION

The flow control specified in the ATM Forum TM 4.0 is a major step forward over any flow control ever examined before. In the old days of large packet networks at low speeds, we did not need flow control as precise and powerful as the new explicit rate method. However, with today's network speeds at distances anything further than LAN distances, the new explicit rate is a critical and necessary improvement. As was shown, the binary rate method is still viable when restricted to the LAN, with only a few switches, and only for data. But the explicit rate technique can be used at WAN distances with high startup rates with as low a cell loss rate as desired (10-12 typically). As a result of its extremely good control and low delay (queue buildup), the VC counting explicit rate switches can also be used for real-time operation, for example for video.