

Network to Source Rate Based Flow Control

Dr. Lawrence G. Roberts

Abstract

A Backward Rate Flow Control (BRFC) technique for ATM Networks is described. The memory requirement hardware complexity traffic delay and line overhead are simulated analyzed and compared to the credit technique and various other variations of the rate based technique. The conclusion is that BRFC will provide loss free ABR data transmission without the higher hardware expense required by the credit techniques and with less memory and greater stability than the other rate techniques.

NOTICE: This contribution has been prepared to assist the ATM Forum. This document is offered to the Forum as a basis for discussion and is not a binding proposal on ATM Systems or any other company. The statements are subject to change in form and content after further study. Specifically, ATM Systems reserves the right to add to, amend or modify the statements contained herein.

Scope

A Backward Rate based Flow Control (BRFC) technique for ATM networks controls the data flow from bursty sources so as to minimize the average queuing delay, minimize the required buffer size, insure minimum or no cell loss, minimize control traffic overhead, and requires minimum or no hardware redesign. The technique is for each ATM switch to send directly backwards to the traffic source an OAM cell specifying a new multiple of the PCR rate to be used whenever congestion occurs or is relieved. This approach drastically reduces the round trip time delay from forward techniques. By virtue of directly specifying a new rate, it avoids oscillations, and therefore it achieves a new stable operating point with minimal buffer buildup. Since it can be implemented totally or primarily in software, it does not require the extensive specialized hardware and equipment redesign common to credit techniques.

Description

For ABR and VBR+ sources, a PCR is specified which is usually controlled by the traffic shaping mechanism on the NIC. If congestion occurs somewhere on the network and a queue starts building up, the switch observing the congestion (probably triggered by a threshold being exceeded) would construct a single OAM flow control cell. This cell contains a new multiplier of the PCR for the source to observe. This multiplier, M , is computed by observing the actual bandwidth available for this VC (referred to as A) and the actual bandwidth being used by this VC (B) and setting $M = u(A/B)$.

Typically, u should be set to 1 unless the queue is large. This insures that once the source receiving the new rate reduces its flow that the actual bandwidth being sent to the congestion point will be uA , and thus the utilization will be u , (typically 1). The OAM flow control cell should be sent back directly to the offending source. Note that this technique can be used to manage a collection of VCs; in this case OAM cells are sent to all sources making up the collection of VCs. When the source receives the OAM cell, it only need change the PCR being used for traffic shaping to the new $M \cdot \text{PCR}$ on that VC.

For network interface cards where the traffic shaping is implemented on the card, the speed of response will be a few microseconds before the rate of transmission is adjusted to the new PCR. The card would presumably set its PCR for ABR VCs to line rate to start with, and the network would reduce this as required. When network switch which has requested a bandwidth reduction for a VC (or a group of VCs), notes excess bandwidth available for this VC, then it should form a new OAM cell for the VC which has a new, higher $M = (A/B)$. This OAM cell would then be sent back to the source exactly as before, to notify them that their PCR can now be increased to the new level. Each network node will need to keep the cumulative M requested for each VC. Note that this is easily done by software, as opposed to the requirement for extensive VC tables maintained by hardware in credit techniques.

Network to Source Rate Based Flow Control

Dr. Lawrence G. Roberts

Analysis

The basic assumption underlying this flow control technique is that adequate buffer space is available in each network node congestion point to buffer the data traffic for the period of time required to notify the source of congestion and to have the change affect the traffic locally. For LAN environments, this will typically be 50-100 microseconds, which is 36-73 cells at OC-3 rates. This will currently fit quite reasonably within an ASIC if desired. For WAN environments, this delay might be 40ms, or at OC-3 rates, 1 Mbytes of memory. Since CBR, VBR and UBR do not need this much memory, only some of this memory will be required for ABR and VBR+ per OC-3, and thus the actual cost in a shared memory might be \$10 per OC-3, clearly not a prohibitive amount. Thus, backward rate flow control is quite feasible in terms of memory requirements. However, for forward notification techniques, the round trip delay is substantially increased and may not be economic or stable. By the time traffic has funneled through a LAN network into a WAN network, the huge bursts of multiple OC-3s hitting at once will already be flow controlled and compressed to manageable stream with the flow control specified herein. The buffers required for this will be less than 50 cells as indicated. If forward notification is used, the LAN buffers required are as large as the WAN buffers and the LAN switches become considerably more expensive. Also, due to the long time delays (ms, not usec), the stability of the control loop is no where near as predictable, and loss rates will certainly be far higher.

Rate vs. Bit Feedback

A second issue in flow control techniques is whether one bit is set to indicate congestion, or a new rate multiplier is computed to accurately set the traffic to the maximum rate supportable. If a single bit is used, the typical adjustment would take several cycles to adjust to a low enough rate, and then it may be up to 50% lower than required. After this, the source is turned on and off to keep the flow at the correct rate like a bang-bang servo. This clearly makes much more line overhead and typically only achieves 75% average utilization of the available bandwidth. On the other hand, by computing the correct rate the sources immediately adjust to the desired utilization level and the average utilization level can easily be held at unity. The drawback, if any, of specifying the rate is the added job for the switch to compute the new rate. This actually is quite simple if counters are available to measure the rate of traffic into the queue and the rate out of the queue. The rate available and the rate presented can be easily determined. If such counters are not available in first generation switches, then the rate can be estimated by examining the queue length periodically. No matter how it is arrived at, it is likely to be more accurate than techniques using a single congestion bit.

Traffic Statistics

The typical ABR traffic will be file or block transfer where a block of data is ready to transfer and the highest possible transfer rate is desired. The NIC will typically be taking data from the source's memory and sending at the full PCR (providing no other process needs the bandwidth). The file or block will typically be from 40 bytes to 100 megabytes (2 usec to 5 sec of transfer time at OC-3). The small one or two cell request or acknowledgment transfers are easily absorbed by any queue without needing any feedback. It is the longer transfers, which are the subject of flow control. This type of traffic is easily managed by rate flow control since it continues at the maximum allowable rate for a significant time, typically far longer than the 50-100 microseconds LAN feedback loop time. For the WAN case, the same situation exists: random length transfers exist which far exceed the loop time of 10-40ms. The shorter transfers are absorbed in the larger buffers. No matter what the loop time, the bursts that are shorter than the loop time are only controlled on an average rate basis to keep the queue length reasonable. The longer bursts can be directly controlled to effect correction before the buffer is exceeded.

Simulation Results

The BRFC technique has been simulated on the following basis:

Network to Source Rate Based Flow Control

Dr. Lawrence G. Roberts

Multiple (from 2-100) traffic sources are combined into one queue. Each source consists of random duration of maximum rate transmission followed by random silence periods. Transmission time was randomized from 0.1 to 0.5 of the loop delay time up to 2-10 times the loop time on different runs. Total traffic offered was tested at up to 20 times the available bandwidth.

Input cell rates, output cell rates and queue depth information was available in the simulated switch to decide upon flow control messages to be sent.

Loop delay before receiving confirmation message was variable. Traffic overhead caused by flow control messages was measured. The results of the simulation runs demonstrated that the average and maximum queue depth could be managed to keep it below two times the number of cells in the round trip loop time. Also, the message overhead could be maintained at 1-3%. A configuration of the LAN network is shown in Figure 3.1, where four VCs are flowing through the network converging one destination. Each source bursts on and off randomly at 155 Mb/s and the connection to the destination is also 155 Mb/s. The total peak rate is up to four times the output rate. The round trip loop time is assumed to be 54 nsec or 20 cells.

The backward rate control simulation results for 1000 cell times are shown in Figure 3.2. The queue depth was controlled by the flow control to a maximum of 24 cells, 120% of the loop time. The average queue depth was 12.7, or 64% of the loop time. The overhead of the flow control message traffic was 2.5% and the output rate was maintained at 98.5% of maximum.

With buffer sizes of 2-4 times the round trip delay time, there should be no cell loss with a tightly controlled process. If the overhead is desired to be reduced, the process can be slowed down and the overhead reduced to 1% or less with a corresponding increase in the buffer size required, as well as in the average queue size. If one is willing to decrease the average output rate to 90-95% rather than 98-100%, then the control task becomes easier and the maximum and average queue size can be reduced. For typical collections of heavy traffic, the output rate can be held at 100% with a small increase in queue depth.

Comparison with Credit Based Techniques Hardware Requirements

The credit technique requires counting cells on each VC, maintaining credit balances on each end of each link, and sending and receiving credit cells for each VC. Given that there may be 4,000 VCs on a link this is a major undertaking in both memory and ASIC hardware. Each link requires 32K of counter memory alone for 4,000 VCs. On the other hand, the backward rate technique requires no special hardware and can be operated on large groups of VCs that may be within a common QoS class. For the group (and there might be 10-20 groups) hardware counters on the cells arriving and leaving are desirable, not mandatory. Such counters can easily be accommodated inside ASICs and do not require large external memories.

Buffer Memory

Table 1 below identifies the memory utilization for the Credit Technique (FCVC) and the Backward Rate Technique (BRFC). Based on the information in [1] the assumptions for FCVC are 10 cells per VC plus shared memory for one times the loop delay. For BRFC the memory required is assumed to be three times the loop delay for the whole group of VCs. The table below is for one OC-3 port with 4000 VCs.

	LAN	WAN
Loop Delay	20	14,000

Network to Source Rate Based Flow Control

Dr. Lawrence G. Roberts

FCVC (Credit)	40,000	54,000
BRFC (Rate)	60	42,000

Table 1: Buffer Memory Requirements (in cells)

Delay

In reference [1] it is suggested that the delay for FCVC is two times the delay of directly sending the data as in BRFC. The argument is that data will be lost this is not true for BRFC. Also the queue build up for BRFC is very small compared to an uncontrolled system. This increase in delay is at least in part due to the slow buildup of a user sending in FCVC, whereas in BRFC the user can burst at his full (modified) rate immediately.

Overhead

FCVC described in [1] has a 10% overhead, compared with 1-3% for BRVC.

Conclusion

The BRFC fully controls the flow of data in ABR and VBR+ services without any cell loss and with minimal complexity, either in hardware or software. The credit technique (FCVC) is extremely hardware intensive, and in the LAN environment requires more memory, dramatically increasing the switch cost. There is no justification for such a complex technique. The forward rate techniques have a major problem for the LAN because they require all switch congestion points to have the buffer space (42K cells) required by the WAN. The techniques using one bit instead of a full rate specification have considerably higher overhead (due to the higher frequency of OAM cells required) and lower throughput, with no corresponding benefit.

Reference

[1] ATM Forum/94-0282 "Adaptive Credit Allocation for Flow-Controlled VCs", H.T. Kung et al.

ATM Forum 94-0448

PROJECT: ATM Forum Technical Committee Traffic Management Sub-working Group

SOURCE: Dr. Lawrence G. Roberts - ATM Systems

DISTRIBUTION: Traffic Management Sub-working Group