14

A Minimal-Buffer Loss-Free Flow Control Protocol for ATM Networks

Siavash Khorsandi and Alberto Leon-Garcia
University of Toronto
Department of Electrical and Computer Engineering,
University of Toronto, Toronto, Ont., Canada M4Y 1R5,
Email: khorsand@comm.utoronto.ca

Abstract

Flow control is essential in ATM networks to prevent service quality degradation during periods of network congestion. In particular, best effort services such as data file transfer have stringent cell loss requirements. Use of backpressure in conjunction with buffer reservation is a promising approach. Buffer reservation on per-VC basis require large buffers. Besides, fairness and high network utilization cannot be supported at low buffer sizes. In this paper, we propose a new scheme based on buffer reservation for groups of VCs rather than for individual VCs. A general framework for credit-based link-by-link flow control is developed by disassociating buffer reservation from credit allocation. Optimality conditions for a credit allocation mechanism are found and an adaptive credit allocation algorithm is designed. The buffer requirement of this scheme is close to one round trip time worth of cells per group. Simulation results indicate that fairness is maintained in a robust manner.

Keywords

Flow control, Credit-based, ATM

1 INTRODUCTION

In ATM networks due to statistical multiplexing, service quality degradation will occur during periods of network congestion [Chiabaut 1994]. Best-effort services such as data file transfer have stringent cell loss requirements. In gigabit networks, the end-to-end feedback based congestion control mechanisms cannot react quickly enough to short-term congestion due to delay between the sender and receiver [Maxemchuck 1990]. Use of selective backpressure based on a distributed flow control protocol is an alternative approach [Mishra 1992, Kung 1993]. In order to avoid possible deadlock and unfairness problems it is necessary to reserve buffers for each active connection or flow [Tanenbaum 1989]. In the credit-based flow control, by combining selective backpressure per flow and buffer reservation mechanisms, it is possible to guarantee that cells are never dropped due to congestion. Furthermore, quick reaction to release of resources results in a better network utilization.

A possible approach for buffer reservation is to strictly partition the available buffer space at each node among active flows. This scheme requires a memory equal to one round-trip delay worth of cells for each connection. To reduce the memory size, it is possible to periodically

L. Mason et al. (eds.), *Broadband Communications*

evaluate the requirements of individual connections and dynamically reallocate buffers to meet those requirements [Ozveren 1994, Kung 1994]. These schemes still require a memory equal to several round-trip delays worth of cells for every group of connections.

In this paper, we propose a credit-based flow control algorithm which differs from the previous approaches mainly in buffer reservation process. We perform the buffer reservation for groups of VCs rather than for individual connections. As a result, the buffer requirement of the scheme is close to the minimum of one round trip delay worth of cells per group. Two levels of flow control is applied. One to control the aggragate flow of a group and the other to control the flow of individual VCs inside a group. Hence, credit allocation of individual VCs can be optimized independent form their actual buffer usage to achieve good transient response and high network throughput.

The rest of this paper is organized as follows. In Section 2, we develop the framework of the proposed flow control mechanism and the concept of group-level buffer reservation is established. In Section 3, the adaptive credit allocation mechanism for flow controlled connections is discussed and a rate-based algorithm is presented. In Section 4, we study the properties of the proposed scheme. Finally, Section 5 contains the simulation results in which transient and steady state behavior of the protocol is studied.

2 FLOW CONTROL MECHANISM AND GROUP-BASED BUFFER RESERVATION

2.1 Basic operation of credit-based flow control

The operation of the credit-based flow control protocol is shown in Figure 1 for a single hop of an ATM virtual circuit (VC). The protocol is applied on a hop-by-hop basis. At the upstream node, U, every VC is allocated a number of credits, C_i , $\forall i$. Each time a cell is forwarded on this VC by U, it increases a counter that keeps track of the number of outstanding cells, $O_i(t)$. At any time, the number of available credits of a VC is equal to $C_i - O_i(t)$. As long as the number of available credits remains positive, the upstream node can forward cells on that VC. At the downstream node, D, the data for each VC is queued separately. Every τ seconds, the number of cells forwarded from each VC's buffer is acknowledged to the upstream node where the number of outstanding cells is reduced by the same amount.

Let 'd' be the link propagation delay between U and D and R to be the round-trip time (RTT) defined as $2 * d + \tau$ both in seconds. We also denote the raw link rate (between U and D) in cells/second by B. Then, the maximum long-term average transmission rate of a VC is equal to $\lambda_i^s = \min(B, \frac{C_i}{R})$. This is referred to as *maximum sustainable rate* of the connection. We also define $G(t)$ and $H(t)$ as the cumulative cell departure process at the upstream and downstream nodes respectively (Figure 1). During the transmission of a cell, $G(t)$ and $H(t)$ increase at a constant rate of B cells/second.

In conjunction with the credit-based mechanism, a per VC fair queuing scheme is necessary to provide each VC with a fair share of the bandwidth. The primary goal in fair queuing is to serve sessions in proportion to some prespecified service shares, independent of the queuing load presented by the sessions [Parekh 1993].

2.2 Adaptive credit allocation

In static credit allocation, C_i remains constant during the life time of a connection. In order to prevent buffer overflow at D, each VC must then have a reserved buffer space for C_i cells

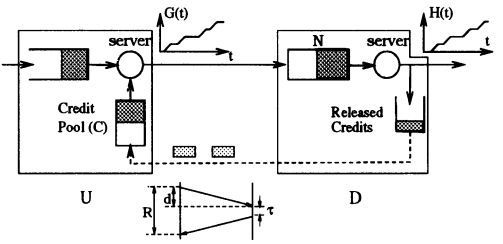

Figure 1 The operation of a credit-based flow-controlled connection.

which can be in the order a RTT worth of cells which is wasteful and expensive. Since the bandwidth requirement of the connections is not constant, in this work we periodically evaluate the requirements of individual connections and adjust their credit allocation to meet those requirements. This could be done only among those connections that can dynamically share the buffer at D. We define a set of such VCs as a *group* and it is denoted by $\mathcal{G}$. If the cells are buffered at a shared memory or at the input of the switch at D, a group includes all the VCs in the same input link. Otherwise, the VCs going from U to D may form several groups depending on their output port at D.

The time between two credit adjustments, τ_c, is called a *control interval.* The credit allocation of a VC during kth control interval is denoted by $C_i(k)$. Otherwise, all the previous formulations still apply. The time at the beginning of kth control interval at U is denoted by t_k. Also, for brevity we denote $t_k + d$ by t_k^d which is the start of kth control interval at D (Figure 2-b).

2.3 Group-level buffer reservation

In per VC buffer reservation, the available buffer space for a group, N, is strictly partitioned among VCs in that group and the credit allocation of a connection corresponds to its buffer allocation. Hence, we have $\sum_{i \in \mathcal{G}} C_i(k) \leq N$. As we will demonstrate later, this approach suffers from slow-start phenomenon and requires a buffer size of at least two RTT worth of cells.

Our approach is based on buffer reservation at group level. The buffer allocated for a group at D is not pre-partitioned among the VCs by the upstream node. The credit allocation of VCs does not necessarily correspond to their buffer allocation and hence the restriction on $\sum_{i \in \mathcal{G}} C_i(k)$ is removed, that is, $\sum_{i \in \mathcal{G}} C_i(k) \lg N$. This allows the credit allocation of VCs to be optimized independent from their actual buffer utilization. Besides, full buffer sharing is provided by allowing the actual buffer utilization of connections to be determined through contention among them. However, it may also result in buffer overflow at D since the availability of a credit may no longer imply the availability of a buffer space. To prevent this, a group-level flow control is applied. Figure 2-a demonstrates the operation of the proposed flow control mechanism.

The group-level flow control regulates the aggregate flow of a group. Therefore, lossless transmission is provided regardless of individual VCs credit allocations. The group-level flow control works similar to VC-level flow control. The credit allocation of a group is denoted by C_g which is equal to its buffer size, N. The number of outstanding cells of a group at time t is also denoted by $O_g(t)$. The number of available credits of the group is equal to $C_g - O_g(t)$. The upstream node can forward a cell on a VC belonging to $\mathcal{G}$ only if the number of available credits for that group is positive. As the acknowledgements for individual VCs in the group arrive, O_g is also re-

duced. Therefore, the group-level flow control does not require any extra transfer of information between nodes. In the next section, we will address the adaptive credit allocation for VCs.

3 CREDIT ALLOCATION PROCESS

Although the credit allocation of a VC does not necessarily correspond to its buffer allocation, it sets an upper bound on its buffer utilization. The credit allocation of VCs must be optimized to achieve objectives such as preventing a possible deadlock, maintaining fairness, maximizing network throughput and achieving good transient response.

If a VC is targeted to transmit at the maximum sustainable rate of λ_i^s, we must have $C_i(k) \geq \lambda_i^s R$. The credit allocation of a VC is equal to the maximum number of cells that it can transmit over a RTT. To proceed, we define the following average rates:

$$\lambda_i^R(k) = \frac{G(tk+R) - G(t_k)}{R}, \quad \mu_i^R(k) = \frac{H(t_k^d + R) - H(t_k^d)}{R}, \quad \gamma_i^R(k) = \frac{H(t_k^{-d} + R) - H(t_k^{-d})}{R} \quad (1)$$

As depicted in Figure 2-b, $\lambda_i^R(k)$ and $\mu_i^R(k)$ denote the average transmission rate over a period of R seconds at U and D respectively starting with the kth control interval. Also, $\gamma_i^R(k)$ is the average rate of credit arrivals at U during the same period. We also define $\hat{\mu}_i^R(k)$ as the maximum value of $\mu_i^R(k)$ given that the connection is able to transmit at its maximum sustainable rate, λ_i^s, at U. For simplicity of formulation, we assume that a RTT is an integer multiple of the control interval, that is, $R = m \cdot \tau_c$ where m is a positive integer. Nevertheless, the algorithm itself is robust and does not depend on this assumption. It can be seen that $\gamma_i^R(k) = \mu_i^R(k-m)$. We also define $n_i(t)$ to be the number of cells of VC $\#i$ in the buffer at D at a given time t.

Lemma 1: Under credit-based flow control with adaptive credit allocation, the following conditions are necessary to maximize network throughput:

a) $$C_i(k) \geq \hat{\mu}_i^R(k) \cdot R \ , \ \forall i$$

b) $$\sum_{i \in \mathcal{G}} [C_i(k) - \hat{\mu}_i^R(k) \cdot R]^+ \leq N - B \cdot R \quad (2)$$

where $[x]^+$ is equal to x if $x \geq 0$ and is 0 if $x < 0$.

Proof: a) The proof is straightforward. $\hat{\mu}_i^R(k) \cdot R$ is the number of cells that can be forwarded at D on this connection over the next R seconds. In order for the VC to utilize the available bandwidth, its credit allocation must be equal or greater than $\hat{\mu}_i^R(k) \cdot R$. On the other hand suppose $C_i(k) < \hat{\mu}_i^R(k) \cdot R$. The throughput of an active connection will be throttled by the flow control mechanism while available bandwidth at D is underutilized.

b) An active connection utilizing more than $\hat{\mu}_i^R(k) \cdot R$ cells over a RTT will eventually build up a queue of size $C_i(k) - \hat{\mu}_i^R(k) \cdot R$ at D. Taken over all the connections, a queue size of $\sum_{i \in \mathcal{G}} [C_i(k) - \hat{\mu}_i^R(k) \cdot R]^+$ will be built at D. If this is greater than $N - B \cdot R$, the total number of available group credits at U over a RTT is less than $B \cdot R$. Therefore, full link capacity cannot be utilized even if there are active connection with available bandwidth at D. □.

Combining conditions (2-a) and (2-b), the credit allocation of a connection is determined as follows:

$$C_i(k) = \hat{\mu}_i^R(k) RTT + N_i^+(k) \quad (3)$$

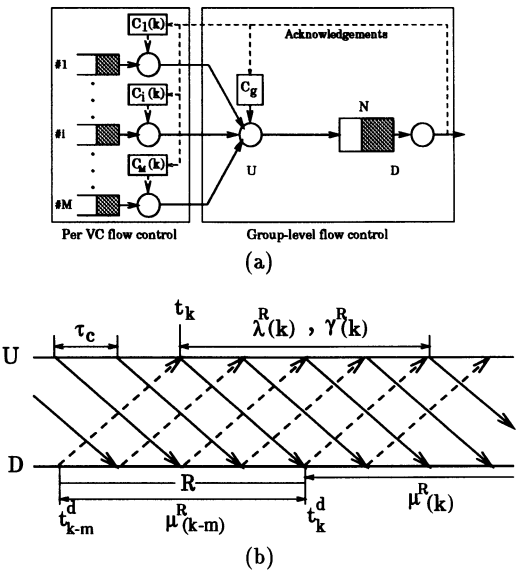

Figure 2 (a) Operation of the proposed flow control mechanism. (b) Time diagram of flow control procedure. We have omitted the subscript i for simplicity.

where $N_i^+(k)$ is called the credit *overallocation* of connection i. From (2-b), we have $\sum_i N_i^+(k) \leq N - B \cdot R$. Equation 3 is our credit allocation *law*. At the beginning of a control interval k, $\hat{\mu}_i^R(k)$ needs to be predicted based on the current and past observations of the cell departure rate at D.

The number of available credits at t_k is $C_i(k) - O_i(t_k)$ and the number of credits received over a RTT is $\gamma_i^R(k)R$. Hence, if the credit allocation remains equal to $C_i(k)$, the maximum number of cells forwarded by U over this period is $C_i(k) - O_i(t_k) + \gamma_i^R(k)R$. By replacing $C_i(k)$ from (3), we have:

$$\lambda_i^R(k) \leq \frac{1}{R}[\hat{\mu}_i^R(k)R + N_i^+(k) - O_i(t_k) + \gamma_i^R(k)R] \tag{4}$$

To complete the credit allocation process in (3), we need to calculate $N_i^+(k)$ and $\hat{\mu}_i^R(k)$. We will address these two problems later in this section.

An alternative formulation

It is instructive to present an alternative formulation of credit allocation process as a rate control mechanism. We show that the control law in (3) is equivalent to a rate-based flow control which aims to keep the buffer occupancy at a level less than or equal to $N_i^+(k)$. We have

$$n_i(t_k^d + R) = n_i(t_k^d) + \lambda_i^R(k)R - \mu_i^R(k)R. \tag{5}$$

From Figure 2-b, it is easy to show that

$$n_i(t_k^d) = O_i(t_k) - \gamma_i^R(k)R \tag{6}$$

Substituting (6) in (5) and setting $n_i(t_k^d + R)$ at less than or equal to $N_i^+(k)$, equation (4) is obtained.

3.1 Credit overallocation

The overallocation is necessary for a connection to keep a backlog in the buffer at D. If the departure rate from the buffer increases, the backlog is drained and the rate increase will be sensed at U by an increase in the rate of credit (acknowledgements) arrival. The credit allocation of the connection is then increased accordingly. The increase in the rate of credit arrivals over a RTT depends on the size of credit overallocation. For a fixed $N_i^+(k)$ over a RTT, maximum increase in the rate of the connection is equal to $\frac{N_i^+(k)}{R}$. We use the following formula to update the credit overallocation of a VC in the beginning of each control interval:

$$N_i^+(k) = \eta \cdot \gamma_i^{\tau_c}(k-1) \cdot R + \frac{N^+}{\#active\ VCs} \tag{7}$$

where $N^+ = [N - (1+\eta)B \cdot R]^+$ is the buffer space equally divided among active VCs, and $\gamma_i^{\tau_c}(k-1)$ is the average rate of credit arrival at U over the past control interval. A VC i is considered active if it has a backlog at U or if $\lambda_i^R(k-m) > 0$. The first term in (7) enables a VC to increase its rate to $(1+\eta) * 100\%$ of its current rate in one RTT. The rate increase is significant if the current rate is large. Hence, the first term favors high rate VCs. The second term, on the other hand, lets a VC to increase its rate by a fixed portion of link rate every RTT which is especially significant for low-rate VCs. Also, the second term ensures that every VC has a minimum buffer reservation. This protects the network from a possible deadlock.

3.2 Rate prediction

To predict the departure rate $\hat{\mu}_i^R(k)$, we use exponential averaging which is a first order autoregressive filter:

$$\hat{\mu}_i^R(k) = (1-\alpha)\hat{\mu}_i^R(k-1) + \alpha \cdot \gamma_i^{\tau_c}(k-1) \tag{8}$$

where $\gamma_i^{\tau_c}(k-1) = \mu_i^R(k-m-1) - \mu_i^R(k-m)$ is the latest observation of the average departure rate over a control interval. The predictor is controlled by parameter α which is the weight given to the latest observation relative to the past history. Although more complex predictions can be used, this predictor has proved effective and is simple enough to be implemented in ATM switches.

Parameter α

Parameter α can be dynamically determined using Kalman filtering if the perturbation in the departure rate is modeled as a white Gaussian noise, $\mu_i^R(k) = \gamma_i^{\tau_c}(k-1) + w_k$, and the variance of noise is known [Keshav 1991]. However, this is not usually the case. An alternative approach is to use an adaptive scheme to adjust α based on the prediction error. Intuitively, the value of α must be small when the transmission rate is steady to give more weight to the past history and filter out transient changes in $gamma_i^{\tau_c}(k)$. However, when the transmission rate is quickly

changing, α must be large to give the emphasis to the recent observations since past observations are too old.

Speed of Convergence

Starting with an initial state $\hat{\mu}_i(0) = 0$, a VC can increase its transmission rate only by $\frac{N_i^+(k)}{R}$ every RTT. To increase the speed of rate increase, the following actions are taken:

a- Instead of starting with $\hat{\mu}_i(0) = 0$, the departure rate of a VC can be predicted using the transmission rate of other active VCs in the same virtual path. An alternative is to start from a fixed nonzero initial state or develop a statistical method to predict the available transmission rate through the network.

b- If $\gamma_i^{\tau_C}(k-1) > \hat{\mu}_i^R(k-1)$ and $\lambda_i^R(k-1) > \hat{\mu}_i^R(k-1)$, this indicates the VC is in ramp up mode, that is, it requires higher transmission capacity which is available both at U and D. Hence, the prediction of departure rate is scaled up by a factor of β:

$$\hat{\mu}_i^R(k) = (1-\alpha)\hat{\mu}_i^R(k-1) + (\alpha+\beta)\gamma_i^{\tau_c}(k-1) \tag{9}$$

4 PROPERTIES OF THE PROPOSED SCHEME

4.1 Credit allocation

Due to separation of credit allocation from buffer reservation, we have:

1- It is possible to have $C_i(k) < O_i(t_k)$ which means that in-use credits of a VC can be revoked and assigned to other VCs. Consequently, as acknowledgements arrive for these credits, they can be immediately used by other VCs.

2- Credit overbooking is possible, that is, $\sum_{i\in\mathcal{G}} C_i(k) > N$. Hence a buffer space can be simultaneously assigned to several VCs. While cell loss is prevented by group-level flow control, the actual buffer usage is determined by contention among VCs. In the absence of contention, a VC can use a larger share of the buffer resulting in a better transient response and better resource utilization.

4.2 Buffer requirement

The minimum buffer requirement of a group is obtained by substituting (3) in (2-b) and taking the equality, $N = B_l \cdot RTT + \sum_{i\in G} N_i^+(k)$. Combining this with (8), we have $N = B_l \cdot RTT + \eta RTT \sum_{i\in G} \gamma_i(k-1) + N^+$. In the worst case, we have $\sum_i \gamma_i(k-1) = B_l$, then

$$N = (1+\eta)\cdot B \cdot R + N^+ \tag{10}$$

A typical value for γ is 0.2 which allows a VC to immediately increase its rate by 20%. The optimal value of N^+ depends on the network condition. However, if we let $N^+ = 0.3B \cdot R$. Then the total buffer requirement of a group is equal to $1.5B \cdot R$, that is, one and half round trip delay worth of cells. This scheme can work with buffer sizes very close to the minimum value of $B \cdot R$. A minimum buffer size of $B \cdot R$ is required to allow a group to fully utilize the link capacity with no cell loss occurring due to buffer overflow.

4.3 Stability and fairness

It can be shown that the flow control mechanism is stable for $\alpha < 1$ and that an steady state is achieved in a finite time. Besides, since the group-level flow control does not interfere with the scheduling mechanism, it can be proved that the bandwidth is fairly divided among VCs. Due to space limitation, we omit the proof of these statements.

4.4 Comparison with a per-VC buffer reservation scheme

An adaptive credit allocation scheme based on per VC buffer reservation is proposed by ATM-Forum in [Chiabaut 1994]. Using our notations, the credit allocation is done as follows:

$$C_i(k) = \hat{\mu}_i^R(k)R + \frac{N^+}{\#active\ VCs}$$

However, this credit allocation has to be modified based on the following two constraints: $C_i(k) \geq O_i(k)$ and $\sum_{i \in \mathcal{G}} C_i(k) \leq N$. The first constraint does not allow $C_i(k)$ to be arbitrarily reduced and the second one does not allow it to be arbitrarily increased. Due to these constraints, the credit allocation algorithm has to be invoked every time an acknowledgement message arrives. This results in an increased computational complexity compared to our scheme where credit adjustment is performed once in a control interval and the size of control interval can be arbitrarily set. Besides, this scheme requires a memory of at least $2B \cdot R$.

5 SIMULATION RESULTS

Extensive simulations have been carried out to evaluate the transient and long-term performance of the proposed scheme. To demonstrate the importance of buffer reservation at group level, we have also simulated the per VC buffer reservation (PVBR) scheme described in Section 4.4. using the same prediction algorithm and similar set of parameters.

The following choice of parameters are made: τ is set to $R/10$ to limit the bandwidth usage of acknowledgement messages to 2% of the link rate, τ_c is set to be equal to τ in order to make our scheme compatible with the PVBR, and η and β are both set to 0.2. In our experiments the distance between every two ATM switch is 330km and the transmission links operate at 1.2Gbps. The simulations are done for various buffer sizes. However, for PVBR scheme, the total buffer allocation of each group is always set at $2B \cdot R$.

5.1 Transient response

A three-hop network scenario shown in Figure 3 is used to evaluate the transient behavior of the protocol. The performance of four source-destination pairs, SD1-SD4, is monitored. Cross traffic is generated by high-speed and low-speed sources in bunches of 10. Low speed sources are added to test the sensitivity of the protocol to presence of large number of active VCs. Peak rate of high-speed and low-speed sources are B and $0.01B$ respectively. Destination node D1 receives 1,2 and 3 hop traffic. For SD1 and SD2, C-D1 link is the bottleneck while for SD3 and SD4 the bottleneck is B-C link.

Connections SD1 to SD4 are activated according to the pattern of Figure 4. Therefore, SD1 and SD2 will have enough time to fill up the group buffer before SD3 and SD4 become active. This is a particularly difficult scenario to handle since the buffer is occupied by the connections

which face congestion two hops farther down the line. As the last source to become active, SD4 is expected to receive the worst performance.

Credit allocation

The process of credit allocation of connection SD4 at node A is shown in Figure 5. The credit allocation starts from an initial state and gradually adapts to the network dynamics. When the buffer size is reduced from $2B \cdot R$ to $1.5B \cdot R$, the allocation process goes through an oscillatory period but still stays around the steady state range of allocation. On the other hand, the credit allocation in PVBR scheme has to start from zero and remains oscillatory. In the PVBR, a VC cannot increase its allocation before other VCs' allocations are decreased. Meanwhile, these VCs continue to use a larger share of bandwidth. Larger delay in adapting to the network dynamics results in oscillation.

Network throughput

Figures 6 and 7 compare the throughput of SD4 and the overall throughput for various cases. The reference point is the proposed scheme with a buffer size of $2B \cdot R$ in both cases. By decreasing the buffer size, the throughput also decreases in a transient period. On the other hand, too large a buffer is also problematic. Using a buffer size of $10B \cdot R$, the overall throughput is reduced up to 30%. The reason lies in the fact that due to a large buffer size, connections SD1 and SD2 continue to consume a large portion of the bandwidth while they face congestion at node C. However, for smaller buffer sizes, SD1 and SD2 are stopped faster and the bandwidth is used by SD3 and SD4 which does not face congestion and hence the network throughput increases. This effect is more explicitly shown in Figure 8 which depicts the overall cumulative throughput after 10 and 15 RTTs as a function of buffer size. The peak throughput is achieved for a buffer size of around $2B \cdot R$. In general, optimal buffer size depends on the number of connections and the network scenario.

Fairness issue

Two similar connections, SD3 and SD4, should receive the same amount of bandwidth when both are backlogged. Figure 9 shows the difference in the throughput of SD3 and SD4 after SD4 becomes active. The proposed scheme demonstrates a robust form of fairness. The throughput SD3 and SD4 remain very close to each other regardless of the buffer size. However, under PVBR the difference between their throughput is up to 80%.

5.2 Long-term performance

The average network throughput and burst delay over a long term have been measured in a statistically symmetric multihop network scenario as a function of the buffer size and average network load.

Source model

We have used On-Off source model with exponentially distributed on and off periods and the average burst size is 100 Kbytes. Each link is shared by 40 VCs and we change the utilization factor of the sources in order to adjust the network load.

We have constructed a statistically symmetric multihop network (Figure 10) where every two switches are 330km apart, each ATM switch is connected to ten local terminals, and all the virtual connections traverse four network hops. Therefore, each link is shared with 1,2,3 and 4 hop traffic.

The average burst queuing delay is shown in Figure 11. The long-term average network

throughput, ρ, in Figure 11-a is 0.6 and in Figure 11-b, it is 0.85. As it is seen the network throughput of $\rho = 0.85$ is supported with a group buffer size of as low as $1.2B \cdot R$. In terms of average burst delay, it is seen that the proposed scheme with a buffer size of $1.3B \cdot R$ works better than PVBR scheme with a buffer size of $2B \cdot R$. The minimum burst delay is achieved with a buffer size of $3B \cdot R$. In fact, in the heavy load condition, the delay tends to rise for buffer sizes beyond $3B \cdot R$ although the increase is not considerable.

6 CONCLUSION

In this paper, we developed a distributed credit-based flow control mechanism in which the buffer allocation is handled at group level while adaptive credit allocation is performed at VC level. The group level flow control mechanism provides strictly lossless transmission regardless of individual VCs credit allocation. As a result, credit allocation can be optimized independently. We developed a rate based credit allocation scheme and we proved that it is necessary to maximize the network throughput.

The technique developed in this paper can work with buffer sizes very close to one round trip delay worth of cells while maintaining the network throughput at a relatively high level. We investigated the transient and long-term performance of the technique. According to our results, in a multihop network scenario, peak network utilization is achieved with a buffer size of around $1.2B \cdot R$ per input link. We have also compared the proposed technique to an adaptive credit allocation technique based on per VC buffer reservation and the proposed scheme was clearly superior.

REFERENCES

J. Chiabaut (1994) Flow control for ATM networks, Bell-Northern Research, Technical document, Ref. CRL-94134

S. Keshav (1991) A control-theoretic approach to flow control, *SIGCOMM'91 Conference, Computer Communication Review,* Vol.21, No. 4, Sept. 1991

H.T. Kung, R. Morris, T. Charuhans and D. Lin (1993) Use of link-by-link flow control in maximizing ATM networks performance: simulation results, *IEEE Hot Interconnect Symposium'93*, Palo Alto, CA., August 1993

H.T. Kung, T. Blackwell and A. Chapman (1994) Credit-based flow control for ATM networks: credit update protocol, adaptive credit allocation and statistical multiplexing, *Computer Comm. Review*, Vol. 24, Oct. 1994, pp. 101-114

N. Maxemchuck and M. El Zarki (1990) Routing and flow control in high speed, wide area networks, *Proceedings of the IEEE,* Vol. 78, No. 1, Jan. 1990

P.P. Mishra and M. Kanakia (1992) A hop-by-hop rate-based congestion control scheme, *SIGCOMM'92*, pp. 112-123

C. Ozveren, R. Simcoe and G. Varghese (1994) Reliable and efficient hop-by-hop flow control, *Computer Comm. Review*, Vol. 24, Oct. 1994, pp. 89-100

A.K. Parekh and R.G. Gallager (1993) A generalized processor sharing approach to flow control in integrated services networks: the single-node case, *IEEE/ACM Transaction on Networking*, Vol. 1, No. 3, 1993, pp. 344-357

A. Tanenbaum (1989) *Computer Networks*, Prentice Hall, 2d edition

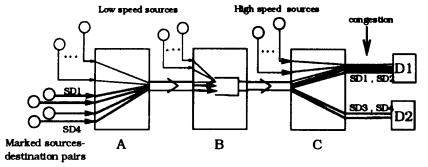

Figure 3 A network scenario used in evaluating the transient responses.

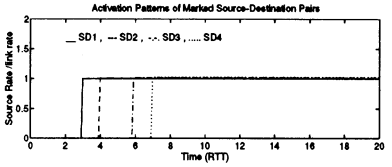

Figure 4 The activation patterns of marked source-destination pairs.

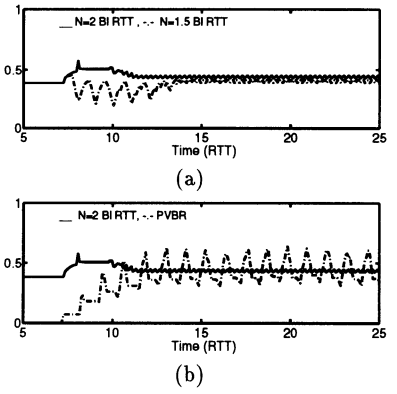

Figure 5 Credit allocation process for SD4 : a) for two buffer sizes $1.5B \cdot R$ and $2B \cdot R$ b) for the proposed and the PVBR schemes with buffer size $2B \cdot R$.

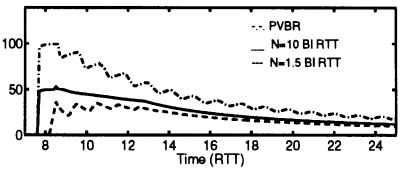

Figure 6 Percentage of loss in the cumulative throughput of connection SD4 relative to the proposed scheme with group buffer size of $2B \cdot R$.

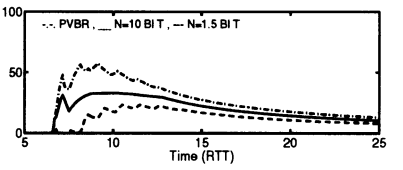

Figure 7 Percentage of loss in the overall cumulative throughput of marked connections relative to the proposed scheme with group buffer size of $2B \cdot R$.

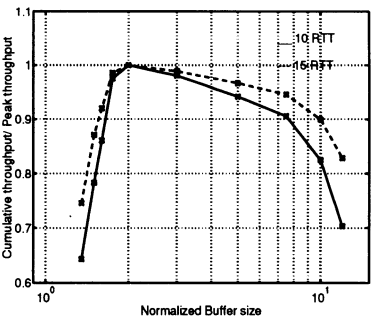

Figure 8 The overall cumulative throughput of the marked connections 10 and 15 RTTs after activation of SD1.

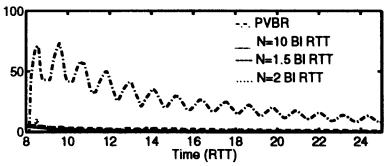

Figure 9 Percentage of difference between the cumulative throughput of connection SD3 and SD4 when both were active.

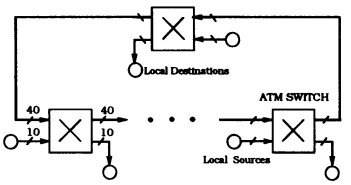

Figure 10 A symmetric multihop network scenario used in evaluating the long-term performance of the protocol.

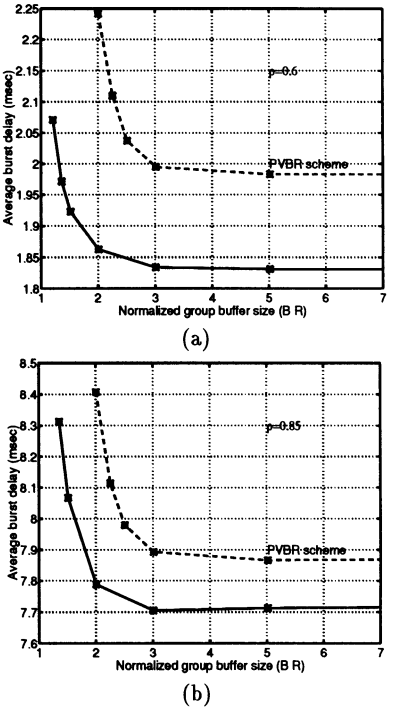

Figure 11 Comparison between the average burst delay of the proposed and the PVBR schemes as a function of group buffer size in a 4-hop symmetric network: (a) for a moderate activity factor of 0.6 (b) for a high activity factor of 0.85.