Flow Control in a High-Speed Bus-Based ATM Switching Hub

Song Chong[a] Ramesh Nagarajan[b] Yung-Terng Wang[b]

[a]Department of Electronic Engineering, Sogang University, Seoul 100-611, Korea
[b]Performance Analysis Department, Bell Laboratories, Holmdel, NJ 07733, USA

Abstract

This paper studies flow control in a high-speed bus-based ATM switching hub for premises switching. The switching fabric is a dual-bus with slots for diverse port cards that interface to the external world. Due to the potentially large switching capacity needed, there can be a significant discrepancy in the switching fabric speed and the port card speed. This can result in buffer overflows at the receiving port buffer and consequently high losses in the switching fabric. Adequate flow control mechanisms are necessary in order to prevent buffer overflows and consequently losses. This paper first examines two different flow control strategies and highlights their strengths and weaknesses. Then we consider a third hybrid strategy which combines the strengths of the first two strategies. In the first flow control scheme, which we refer to as physical flow control, all streams destined to a receiver with buffer problems such as high occupancy and loss are shut down till the congestion is cleared. This scheme has the advantage that loss is severely limited but has the disadvantage that high-rate streams arriving at this buffer can in some circumstances starve lower-rate streams. In a second strategy, referred to as logical flow control, streams are implicitly selected based on their rates and shut down. This scheme has the advantage that the higher-rate streams will eventually be shut down more often and hence cannot overwhelm the lower-rate streams. However, loss is not easily controlled in this scheme. Finally, in a hybrid strategy, we combine physical and logical flow control with logical control activated first and physical control activated later when the logical control is unable to limit the buffer occupancy and loss. We show that this hybrid strategy has the desirable properties of the physical and logical control schemes and hence is the recommended choice for flow control in the setting of interest.

1 Introduction

The past years have witnessed a tremendous increase in the traffic volumes in both WANs like the Internet and also on-premise LANs like the Ethernet. This increase in traffic volume is due to new technologies, migration from a paradigm of central to distributed computing and a host of new applications. Also, the fast pace of technology growth is witnessing increasingly inter-disciplinary work in which groups of individuals from diverse groups/divisions come together for a project and disband. From a networking standpoint, this implies that the typical communities of interest (COI) such as a department no longer are the rule and in fact these COI are regularly changing [5]. This results in severe network management problems. In addition to traffic volumes and network management problems, there is also a bewildering variety of co-existing applications such as telephony, video and data networking. The seamless integration of these services poses an extreme challenge in both the premises network and the wide-area network. This has resulted in a dramatic shift from the present method of operation which typically involves routers and bridges to switching in the premises. This paper focuses on flow control in a high-speed bus-based ATM switching hub for the premises switching.

The switching hub architecture we consider [1] is a dual-bus-based one with bus slots supporting various port cards that interface to the external world. A typical configuration might be a bus running at multi-gigabit rates supporting diverse port cards with aggregate rates up to OC-12. Port cards are likely to have various interfaces ranging from TDM circuits like T1.5, Ethernet segments, ATM connections to desktops etc.. Since the bus is running at high speeds, there is a requirement for a high-speed buffer in the port cards to stream data to and from the bus. On the other hand, it is desirable to keep the amount of high-speed buffer small in order to limit the cost of the port card. It may now be obvious to the reader that the potential difference in the aggregate port and bus rates can make management of the high-speed buffers particularly difficult. We argue in the paper that with suitable flow control, the amount of high-speed buffer needed can be

0-7803-4443-X/97/$10.00

minimized (and hence cost) while maintaining negligible loss, high throughput and "fair" bandwidth sharing among streams.

The term flow control means different things to different people. In this paper, the term flow control is used to reflect control of streams *within the fabric* in order to manage the high-speed buffers. It does not reflect any network node to source (like in ATM ABR services) or end-to-end flow control between destination and source pairs. We first examine two different flow control strategies and highlight their strengths and weaknesses. Then we consider a third hybrid strategy which attempts to combine the strengths of the first two strategies. In the first flow control scheme, which we refer to as physical flow control, all streams destined to a receiver with buffer problems such as high occupancy and loss are shut down till the congestion is cleared. This scheme has the advantage that loss is severely limited but has the disadvantage that high-rate streams arriving at this buffer can in some circumstances starve lower-rate streams. In a second strategy, referred to as logical flow control, streams are implicitly selected based on their rates and shut down. This scheme has the advantage that the higher-rate streams will eventually be shut down more often and hence cannot overwhelm the lower-rate streams. However, as we will see, loss is not easily controlled in this scheme. In addition, the operation of the scheme requires significantly more bus bandwidth and high-speed buffer than that of the physical scheme. Finally, in a hybrid strategy, we combine physical and logical flow control with logical control activated first and physical control activated later when the logical control is unable to limit the buffer occupancy and hence loss. We show that this hybrid strategy has the desirable properties of the physical and logical control schemes and hence is the recommended choice for flow control in the setting of interest.

The rest of the paper is organized as follows. In Section 2, we detail the hub architecture in consideration and also the flow control problem. Section 3 considers various strategies for flow control. Performance of the various flow control strategies is presented in Section 4 and the conclusion appears in Section 5.

2 Switch Architecture and Problem Statement

We consider a switching hub architecture as shown in Figure 1 [1]. The switch fabric is a dual-bus architecture in which all port boards transmit on a transmit bus and receive from a separate receive bus. The transmit bus is looped back onto the receive bus through a loop-back circuit located at the far end of the bus. A typical configuration is a bus running at a multi-Gbps speed, supporting port cards with aggregate rates up to 622 Mbps. Port cards are likely to handle a variety of interfaces such as ATM connections, TDM circuits, Ethernet segments and so on. In particular, such an architecture (or like) is considered as an attractive solution for access hubs and backbone hubs in campus, private or corporate networks. Examples are found in [1] and are making appearance in the market place [2].

Access to the bus is achieved via an elevator-style polling mechanism among *active boards* [4]. Transmission on the bus is in units of envelopes which are ATM cells wrapped in some local switch fabric headers including several flags and addressing information. The hub could potentially be employed to switch variable-sized packets as well but our discussion in this report will focus only on fixed-size ATM cells. On each visit of the poll to a port card, it is assumed that only one envelope is served ($MAX = 1$). This is sufficient to maintain high throughput in a short bus because propagation delays are small and it is possible to completely pipeline envelope transmissions and arbitration via polling. The port cards interface to the bus via a high-speed chip which we refer to as BIC (Bus Interface Chip). The BIC is assumed to have simple high-speed FIFO staging buffers for transmission on the bus and receipt from the bus. A large amount of slow-speed memory, which we simply refer to as RAM, is assumed to be resident on the port card outside of the BIC and serves as the primary buffering area to and from the actual physical ports. Thus the function of the BIC memory space is to serve as a staging area for envelopes on the transmit side and as a rate-converter (from the bus transmission rates to the port transmission rates) on the receive side. Due to the large potential difference in rates between the bus speed and port rates, buffer overflows are a serious issue on the receive side of the BIC. This paper focuses on flow control mechanisms that enable one to limit queueing on the receive side of the BIC by essentially shifting the queueing to the large slow-speed memory on the sending port card.

Routing in the fabric is achieved based on a *logical addressing* scheme. An address is assigned to each "logical" egress point which represents either a port card, in which case it will be referred to as a *physical address* as well, a port or even an ATM address (VPI/VCI). Note that no source addressing is used in the hub, and hence all addresses refer to an egress point. As mentioned earlier, this logical (physical) address is part of the local envelope header. On the receive side, BICs use this address to filter envelopes destined to them. Note that multicast is accomplished by simply assigning the same logical address to multiple physical entities (ports or boards) and via a *single transmission*. Queueing on

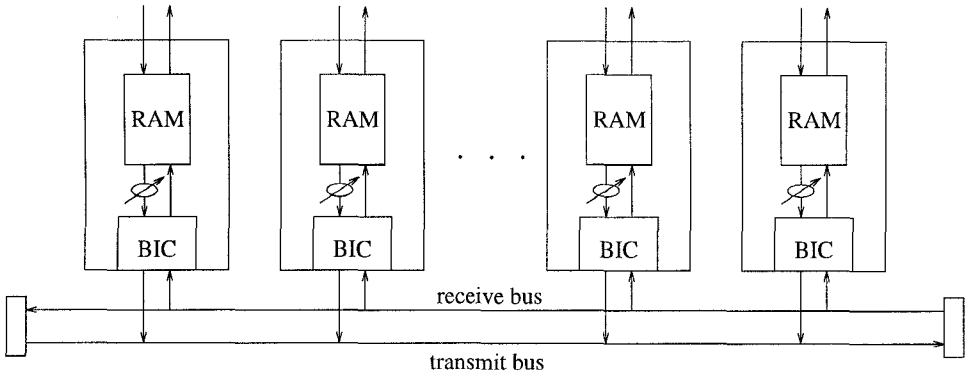

Figure 1: Switch architecture with bus and port boards.

the sending RAM buffer is by (logical) destination addresses which prevents head-of-line blocking when flow control is activated for particular address(es). Finally, the movement of data from the RAM to the BIC buffer in a send board is achieved via a simple round-robin (among destination queues) discipline.

There are several performance issues for a flow control mechanism and we examine some of them here. The first *goal of flow control* is to *prevent cell loss* at the receive BIC. If this were not the case, retransmissions would be necessary to recover from loss and maintain a lossless switch fabric. Retransmissions are wasteful of hub capacity, introduce excessive delays for delay-sensitive services and also require complicated mechanisms at the sender and receiver.

The second goal of flow control is to *maintain throughput.* Persistent flow control, which implies throttling the streams, can result in severe backlogs at the sending side. This in turn will require large amounts of bus and access bandwidth to the BIC during the non-flow-controlled states in order to clear the backlog. This excess bandwidth which may be well in excess of the average stream's bandwidth may or may not be available. When this excess bandwidth is not available, throughput drops below the capacity of the system.

The third goal of flow control is to preserve *fairness in bandwidth allocation* among the streams. Assuming no additional knowledge of the streams such as application behavior, QOS requirements etc., (even if this information were available, any scheme to use this information may be too complex to implement at the BIC level), a reasonable goal for fairness is MAX-MIN fairness, which simply stated guarantees that any stream, at a bottlenecked resource, gets its injected rate and if it receives less than this rate then no other stream that also receives less than its injected rate receives a greater share of bandwidth. Further, fairness is an issue only when the resource is overloaded and if the system is engineered to limit the overload duration, we may view MAX-MIN fairness simply as a means to temporarily penalize the high-rate streams while letting the lower-rate streams through. This is in contrast to schemes like rate-proportional fairness where a high-rate stream can completely block a low-rate stream.

The fourth goal of flow control is to maintain a *low delay* under flow control. Since the essential philosophy of flow control is to shift the queueing from the receive side of the BIC to the large slow-speed memory on the sending port card, it necessarily trades off loss for delay. Since the switch fabric is expected to be extremely fast compared to the individual stream rates, we expect that such a tradeoff has an inherent advantage and it will be important for us to ensure that this is indeed true under flow control.

The other important goals in flow control are to minimize *utilization of bus bandwidth, BIC buffer space and the BIC processor* for control purposes. In particular, when control signaling is done in-band, control messages have to contend with ordinary data for the resources and in some circumstances can starve the data streams. Finally, the flow control scheme should be *simple* enough to implement at the BIC and operate at high data rates up to the bus speed.

3 Strategies and Motivation

In the following, we consider three flow control strategies and examine their performance in light of the goals for flow control that we formulated in the previous section. Before we proceed to do that, we outline the basic tenets of all the flow control schemes to be considered:

- Flow control is of an ON/OFF type with activation shutting down the flow of data from the RAM to the BIC selectively by its logical destination at the sending port board. Note that the flow control limits access to the BIC buffer, but not access to the bus. This is assumed since selective control for

bus access based on logical addresses is too complex to implement in the BIC. In addition, with such a control, a flow being flow-controlled can starve the other flows as its envelopes wastefully occupy the BIC buffer.

- Flow control is triggered based on occupancy statistics of the receiving BIC buffer. An out-of-band flow control signal carries control messages containing congestion information from receiving boards to sending boards and is synchronized with the data signal.

- Flow control is deactivated by the BIC transmitting a "dummy" envelope to itself and upon receipt of the dummy envelope, the BIC automatically generates and transmits a control message just as if the envelope were transmitted by some other board. The control message contains the indication of recovery from congestion and the address of the uncongested BIC. Note that since no envelopes may arrive at a board that activated flow control, the dummy envelope operation is necessary to deactivate flow control.

- For transmission of the dummy envelopes, there is a separate high-speed staging buffer in the BIC, which we refer to as the control buffer. Dummy envelopes have priority over ordinary data envelopes in access to the bus. Among the boards with queued dummy envelopes, bus access is arbitrated via a round-robin discipline.

The above architectural choices were mostly made to simplify the flow control scheme and hence make it attractive to implement in the BIC.

Given the above assumptions, the three flow control strategies that we consider are:

- Logical flow control
- Physical flow control
- Hybrid flow control - combination of logical and physical controls

The philosophy of the logical control scheme is to isolate the streams responsible for congestion and control them only. In principle, it is an ideal scheme since it limits flow control to the responsible streams. However, ideal selection of streams is difficult. To do this, one can use measurements such as occupancies or arrival rates of the streams at the BIC buffer. Since we desire to keep the BIC buffer small, measurement of the occupancies may not be a reliable indicator particularly as a large number of streams are multiplexed. On the other hand, information on arrival rates is regarded as a better indicator but the burden of its measurement is far beyond implementation at the BIC. In the paper, we consider an approach that is simple to implement but achieves the essential philosophy of logical control. In this approach, streams are "implicitly" selected based on their rates and controlled and *no measurements* are required. During a control period, streams are shut down in the order their envelopes appear at the congested receiving board. By doing so, high-rate streams are more likely to be controlled than small-rate streams since the probability of a high-rate stream's appearance is greater than that of a low-rate stream's. The logical flow control algorithm is outlined in the below with the following terminology:

$addr$: Denotes a physical or logical address

$board_{addr}$: Denotes the physical address of a port board

log_{addr}: Denotes the logical address of a physical entity (eg., port or VC).

$Env[log_{addr}]$: The logical address of the destination of the envelope

$Flow_{sig}[addr]$: The address (physical or logical) which is congested

$Log_{map}[board_{addr}]$: A function that maps a physical board address to the logical destination addresses at that board

$Log_{cntrl}[\cdot]$: A list of logical addresses that are being flow controlled, i.e., the traffic to these addresses is being controlled

Que_{bic}: The queue occupancy at the receiving side of the BIC in envelopes

LOGICAL FLOW CONTROL:

Receiver protocol

Event: Arrival of $Env[]$
  If $(Que_{bic} > HTH)$
    Flow Control = ON
    Flow Control $Env[log_{addr}]$
    Controlled List $Log_{cntrl}[] \leftarrow Env[log_{addr}]$
Event: Departure
  If $(Que_{bic} < LTH)$ and (Flow Control == ON)
    Deactivate Flow Control $(Log_{cntrl}[])$
    Flow Control = OFF

Sender protocol

Event: Flow Control Signal Receipt $(Flow_{sig}[])$
  If $(Flow_{sig}[log_{addr}] == ON)$
    Flow Control log_{addr}

else
Deactivate flow control log_{addr}

The essential idea in logical control is to sequentially control logical address flows as the envelopes make their appearance at a congested receiving BIC, defined as queue occupancy exceeding a high threshold HTH, and then to collectively de-control them when the queue occupancy at the BIC drops below a lower threshold LTH.

As we will see in the next section, a drawback of the logical flow control scheme is that loss is not easily controlled since streams are shut down sequentially and in a probabilistic manner. For tight control of loss, however, physical control is necessary. For the implementation of physical control, all port boards in the fabric need to be addressed distinctively and this is accomplished via physical addresses as outlined earlier. The physical addresses can be in a distinct address space with the addition of one bit to distinguish between physical and logical addresses and reusing a common addressing field in the envelope header or some part of the logical address space can be devoted for addressing port cards.

It is also assumed that each BIC can map a physical address into a set of logical addresses that are defined for the board of that physical address. The physical flow control algorithm is similar in structure to the logical flow control algorithm and is shown in the following.

PHYSICAL FLOW CONTROL:
Receiver protocol
Event: Arrival of $Env[]$ at $board_{addr}$
If $(Que_{bic} > HTH)$
Flow Control = ON
Flow Control $board_{addr}$
Event: Departure
If $(Que_{bic} < LTH)$ and (Flow Control == ON)
Deactivate Flow Control $(board_{addr})$
Flow Control = OFF
Sender protocol
Event: Flow Control Signal Receipt $(Flow_{sig}[])$
If $(Flow_{sig}[board_{addr}] == ON)$
Flow Control $Log_{map}[board_{addr}]$
else
Deactivate flow control $Log_{map}[board_{addr}]$

Upon congestion, the physical scheme attempts to shut down all the logical streams at the same time whereas the logical scheme implicitly selects the streams based on their rates and shuts them down only. It will be seen that the physical scheme controls loss more tightly than the logical scheme. In the physical scheme, however, bandwidth sharing among streams can be compromised in some cases: it will be shown that a higher-rate stream can starve a lower-rate stream in some cases since the higher-rate stream dictates the control frequency of the lower-rate stream. On the other hand, the logical scheme consumes more bandwidth than the physical scheme for dummy-envelope transmission since a dummy envelope is generated for each logical address flow to be resumed. In the hybrid control scheme, we attempt to combine the respective strengths of the above two strategies. Logical control is activated first when the BIC occupancy reaches the high threshold HTH. Physical control is only activated when the logical control is unable to limit the occupancy and the occupancy reaches a higher threshold (TTH). Thresholds for this scheme are illustrated in Figure 2. For the implementation of the hybrid control, it is assumed that the flow control message can signal either a physical or logical address. The following shows the hybrid flow control algorithm.

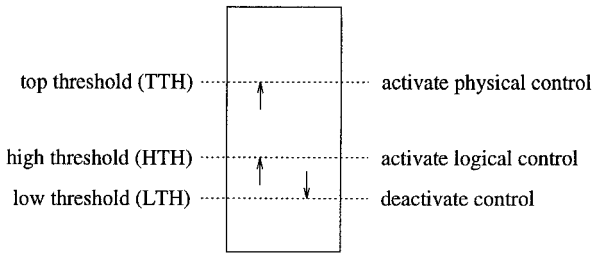

Figure 2: Thresholds for hybrid control in a receiving BIC buffer.

HYBRID FLOW CONTROL:
Receiver protocol
Event: Arrival of $Env[]$ at $board_{addr}$
If $(Que_{bic} > HTH)$
Flow Control = ON
Flow Control $Env[log_{addr}]$
else
If $(Que_{bic} > TTH)$
Flow Control $board_{addr}$
Event: Departure
If $(Que_{bic} < LTH)$ and (Flow Control == ON)
Deactivate Flow Control $(board_{addr})$
Flow Control = OFF
Sender protocol
Event: Flow Control Signal Receipt $(Flow_{sig}[])$
If $(Flow_{sig}[board_{addr}] == ON)$
Flow Control $Log_{map}[board_{addr}]$
else
If $(Flow_{sig}[log_{addr}] == ON)$
Activate flow control log_{addr}
else
Deactivate flow control $Log_{map}[board_{addr}]$

4 Flow Control Performance

We take a simulation approach for the performance study since the analytical modeling of the switching hub with flow control is difficult and is further complicated due to a potentially large state space. Before we proceed to a performance comparison of the control schemes in consideration, we examine the *nature of loss* in the switch fabric as a function of certain key system parameters when *no flow control* is applied. Consider a 4-Gbps bus with 20 port boards and a data transfer rate from the RAM to the BIC at each board, which we will refer to as fetch rate (FR), of 200 Mbps. The transfer rate from the BIC to the RAM, which we refer to as drain rate (DR), is also set to 200 Mbps. Assume that each receiving board has 20 logical destinations and each destination receives data from all the boards in the fabric. For simplicity, the size of the RAM at each board is assumed infinite. Since the main issue in this paper is the queueing performance at the receive side, the traffic pattern at the send side is of less concern and hence, unless otherwise specified, the aggregate incoming traffic is assumed uniformly distributed among all the sending boards. The function of the send BIC buffer is to serve as a staging area for envelopes for transmission on the bus. The capacity of the send BIC buffer should be large enough to sustain flow and thus maximize bus throughput for a given bus arbitration mechanism. Figure 3 shows bus throughput as a function of the size of send BIC buffer when offered load is equal to the bus speed. With a buffer capacity greater than or equal to 2 (envelopes), the traffic flow is sustained and hence full utilization of bus bandwidth is achieved. On the other hand, it is obvious that the larger the capacity of sending BIC buffers is, the more envelopes the receiving boards should accommodate while control is ON because the flow control in consideration acts between the RAM and the BIC buffer. Therefore, we set the size of send BIC buffer to 2 (envelopes) in the rest of section. Denote the total and receive-side capacities of a high-speed BIC buffer, respectively, by B and $RB = B - 2$ in envelope.

The traffic pattern to the receive BIC has the greatest impact on the BIC queueing. The greater the traffic focus, the more serious the congestion is. Note that the traffic pattern at receiving boards is essentially determined by two types of correlation: one in time and the other in space. The former implies the case that a single board is sending envelopes to a particular logical destination for a long period of time, the duration of which is referred to as length of destination correlation, and hence stressing the receiving board to which the destination belongs. The latter implies the case that a number of boards are simultaneously transmitting envelopes to the same logical destination. To model such destination correlations, a random process $\{L_i(k), i = 1, 2, \cdots, k = 1, 2, \cdots\}$ is defined where $L_i(k)$ determines the logical address of the destination for envelopes generated at board i during time-step k (time is assumed discrete). It is assumed that the $L_i(\cdot)$ are i.i.d. and the behavior of $L_i(\cdot)$ evolves as a Markov chain with a transition matrix that is circulant, i.e., each row circulates to the right by one element to form the next row. The first row of the matrix is defined by a vector $[1 - (M-1)p, p, \cdots, p]$ where M is the total number of logical addresses and p is the transition probability from one logical address to another. The mean length of destination correlation at a sending board is represented by the mean sojourn time of the chain in a given state which is $\frac{1}{(M-1)p}$.

Figures 3b and c show loss performance when the destination of envelopes is determined by the above chains. In the first scenario with the results shown in Figure 3b, we examine loss as a function of offered load and length of destination correlation, assuming that $RB = 98$ (envelopes). As expected, loss increases substantially as offered load increases. The offered load of 2.4 to 3.92 Gbps implies that the utilization of each receiving board is from 0.6 to 0.98 respectively since the total load is equally distributed among the 20 receiving boards with $DR = 200$ Mbps. For a given load, loss performance deteriorates greatly as more envelopes are consecutively routed to a same destination by increasing the mean length of destination correlation from 1 to 40 (envelopes).

In the second scenario with the results shown in Figure 3c, we examine the effect of the size of receive BIC buffer on reducing loss. In this subsequent scenarios, we fix the offered load at 3.92 Gbps. By increasing the size of receive BIC buffer, loss is reduced to a certain extent. However, the improvement is limited especially in the presence of strong destination correlation at the sending boards. Considering the cost of high-speed memory, it is obvious that increasing receive-buffer capacity is not a panacea for the prevention of loss. In the last scenario, we consider batch arrivals and examine the impact of batch size on loss. A natural example of batch arrivals occurs in practice with TCP/IP-based applications riding over a ATM transport. In the "Classical IP over ATM" service being defined by IETF (Internet Engineering Task Force), the maximum size of an IP packet is 9.18 Kbytes which implies that the hub can see a batch arrival of up to 190 envelopes [3]. In the simulation, we assume batches of envelopes (generated by a large IP packet) arriving at each sending board head-

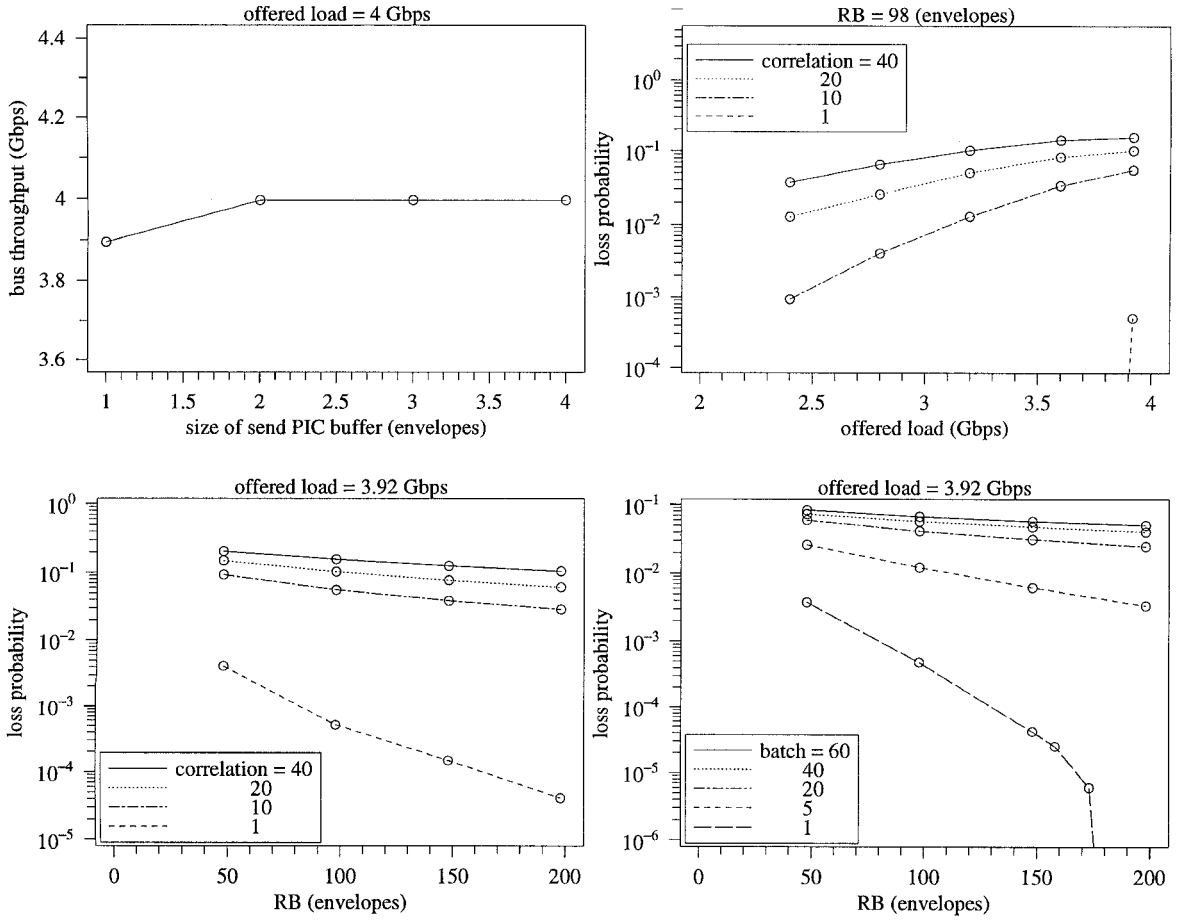

Figure 3: No flow control: (a) bus throughput as a function of size of send BIC buffer, (b) loss as a function of offered load with correlated destination, (c)(d) loss as a function of size of receive BIC buffer with correlated destination and batch arrival.

ing to the same destination. Distinct batches, however, have independently and uniformly distributed destinations. We also assume that the size of a batch is governed by a uniform distribution. Figure 3d shows the loss probability as a function of mean batch size with different receive-buffer capacities. Obviously, loss performance is improved as the batch size decreases or the receive-buffer capacity increases. However, the improvement by buffering is again seriously limited as in the previous scenario. Flow control is, however, a means to reducing the loss by shifting the queueing from the receive side of the BIC to the large slow-speed RAM on the sending port board. However, it is important to not sacrifice other performance metrics while exercising flow control as indicated in Section 2.

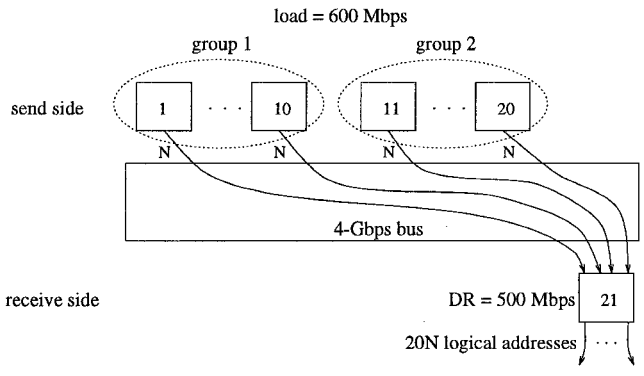

Figure 4: A hot-spot scenario.

To compare the performance of flow control schemes, we consider a hot-spot scenario in a 4-Gbps, 21-board switching hub with $FR = DR = 500$ Mbps, as depicted in Figure 4. The boards indexed from 1 to 20 transmit data to a hot-spot, board 21. For convenience, we refer to boards 1 to 10 as group 1 and boards 11 to 20 as group 2. The total offered load is set to 600 Mbps and the fraction of total load gener-

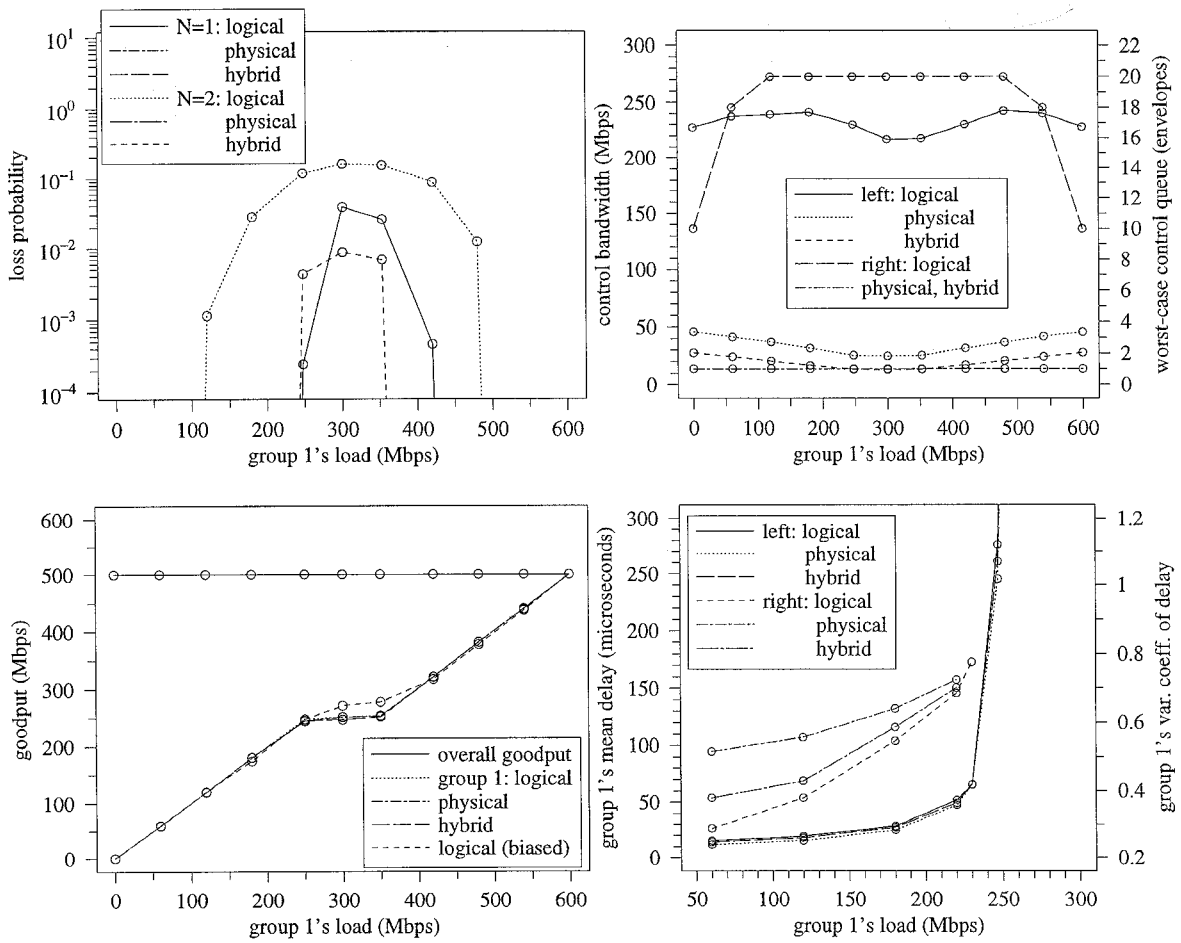

Figure 5: Performance comparison of flow control schemes in a hot-spot scenario: (a) loss, (b) use of hub resources, (c) goodput and bandwidth sharing, (d) switching delay ((b)-(d) are with N=1).

ated by groups 1 and 2 is varied as a parameter. We assume that each sending board sources N independent streams and the streams have disjoint logical destinations. Within each group, the load is uniformly distributed among the logical streams. The size of the BIC buffer is chosen aggressively small at $B = 45$ (envelopes). As an example, we design the thresholds such that $(LTH, HTH) = (5, 10)$ for the logical and physical schemes and $(LTH, HTH, TTH) = (1, 6, 16)$ for the hybrid scheme. Figure 5 summarizes the performance comparison of the three control schemes in this hot-spot scenario. In the experiments, we vary group 1's load from 0 to 600 Mbps and accordingly, group 2's load from 600 to 0 Mbps. In Figure 5a, we observe that the logical scheme cannot easily control loss since in this scheme control is activated incrementally on a per-received-logical-address basis. Figure 6 illustrates such behavior at the hot-spot receive BIC buffer when $N = 1$ and the groups 1 and 2's loads are both equal to 300 Mbps. For the same reason, loss increases as the number of logical streams multiplexed increases from 20 ($N = 1$) to 40 ($N = 2$). Another observation is that loss tends to decrease as group 1 and 2 become incomparable in their rates. This is intuitively true because in this case a group of streams has a dominant rate and hence control is likely to act on only this group. On the other hand, with the physical scheme, no loss occurs for any mix of groups 1 and 2's loads. In the physical scheme, the total number of "transit" envelopes that need be accommodated by the hot-spot buffer during control ON periods is the sum total of envelopes both being transmitted on the bus and those waiting for transmission in the send BIC buffers. In the above example with 20 sending boards and a 2-envelope BIC buffer capacity and a bus length (propagation delay) shorter than 1-envelope transmission time, the sum is bounded by $20 * 2 + 1 = 41$ envelopes. With a spare buffer capacity of 33 envelopes ($= RB - HTH$), the physical scheme is able to accommodate all of the transit envelopes. The occupancy dynamics at the receive buffer

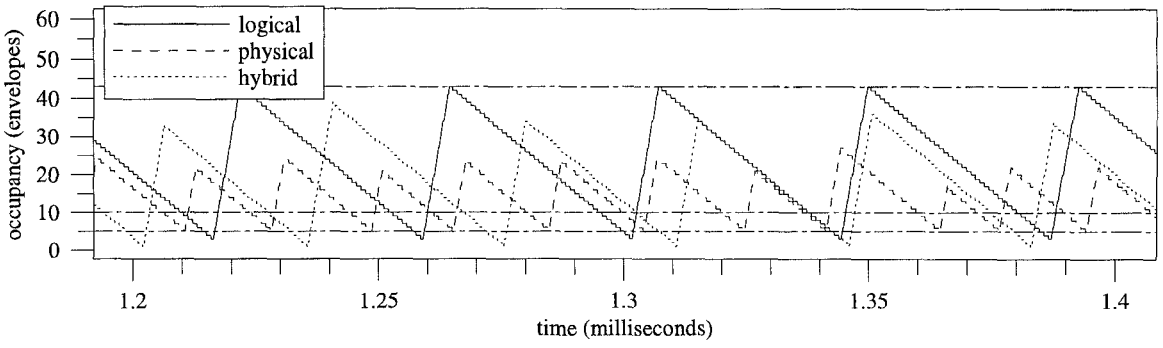

Figure 6: Occupancy dynamics at the hot-spot BIC buffer with different control schemes when $N = 1$ and groups 1 and 2's loads are equally 300 Mbps.

under the physical scheme are also illustrated in Figure 6. Compared with the logical case, the physical case shows a much smaller occupancy overshoot over HTH, thus resulting in no buffer overflow. Also, in the physical case, the minimum occupancy appears to be $LTH - 1$ because all the streams are resumed almost immediately after the occupancy falls down to LTH by sending the send side a common control message indicating the recovery from congestion at that physical destination. In contrast, in the logical case, since the streams receive the uncongestion message individually, it takes additional time for all the sources to be resumed and hence the occupancy goes much further below LTH. With a large number of streams, therefore, the BIC occupancy in the logical case occasionally drops to zero so that the drain bandwidth can not be fully utilized.

In Figure 5b, we compare the utilization of bus bandwidth and the control-buffer space requirement for the transmission of dummy envelopes in the different control schemes. The results shown in Figures 5b, c, d were obtained with 20 streams. The transmission of dummy envelopes, in this scenario, with the physical scheme used 25 to 46 Mbps of bus bandwidth on average and required only one-envelope worth of buffering, whereas the logical scheme utilized 216 to 240 Mbps of bus bandwidth and required up to a 20-envelope buffer. It is obvious that the logical scheme utilizes much more hub resources than the physical scheme since the logical scheme generates as many dummy envelopes as the number of flow controlled streams. The goodput performance of groups 1 and 2 is given in Figure 5c. First, under both logical and physical schemes, the overall goodput is nearly equal to the drain rate of the hot-spot board for any fraction of the total load, which implies that no wastage of bandwidth. Second, both physical and logical schemes achieve MAX-MIN fairness in bandwidth allocation among streams. Further, streams within a group get a equal share of the MAX-MIN share allocated to the group. An important note is that in the logical scheme, unless streams are resumed in a random order, unfair bandwidth allocation among streams can happen. An example with such a bias in bandwidth allocation is also shown in Figure 5c. In the example, we intentionally resumed streams in order from board 1 to 20 each time the occupancy at the receive buffer fell down to LTH, and it turns out that the streams within group 1 receive more allocation than mandated by the MAX-MIN fair allocation, thereby starving the streams within group 2. Finally, in Figure 5d, the switching delay incurred by the control schemes is compared. The switching delay is defined to be the time spent by an envelope from arrival at a send RAM to departure from a receive BIC buffer. Since group 1 cannot get more than 250 Mbps in bandwidth, the switching delay exponentially increases as group 1's load approaches 250 Mbps. The mean switching delay of a stream receiving its requested rate is in the order of a few tens of microseconds with negligible difference between control schemes. The logical scheme leads to a relatively larger variation in the switching delay than the physical scheme. However, the difference is again negligible since the absolute value of delay is fairly low.

Next, we consider the scenario depicted in Figure 7 to reveal an intrinsic advantage of logical flow control in maintaining the throughput of low-rate streams. In this scenario, two logical streams, the high-rate one from board A with destination address 0 and the low-rate one from board B with destination address 1, share the 200-Mbps drain-rate at board C. On the other hand, the 500-Mbps fetch-rate at board B is shared by stream 1 and the other $(N - 1)$ streams via a round-robin discipline. The $(N - 1)$ streams are assumed persistent and hence there are always envelopes destined for addresses 2-N at the RAM so that the stream 1 periodically gets fetch opportunities at an average rate of $500/N$ Mbps. The rate of streams 1 and 0 is set to

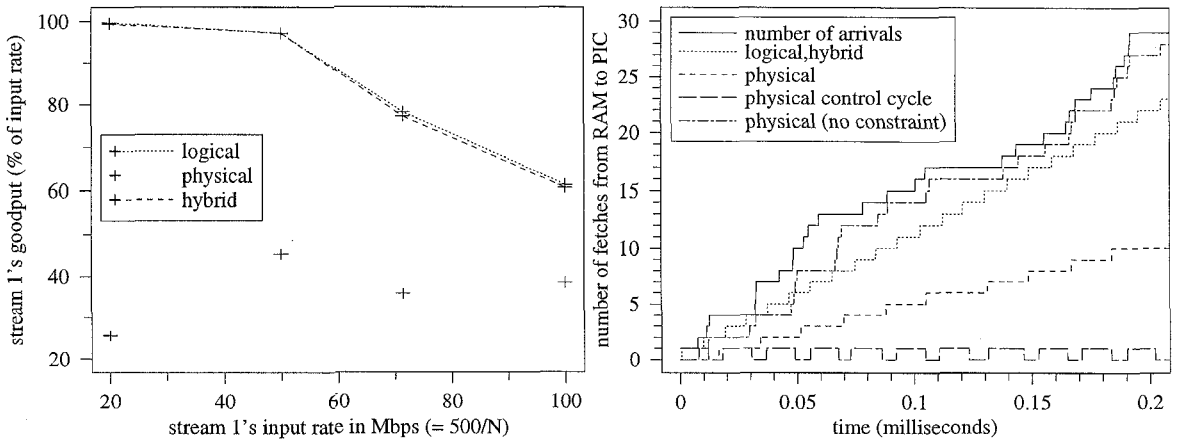

Figure 8: Performance of a low-rate stream subject to fetch-rate constraint with different control schemes: (a) stream 1's goodput at different input rates (b) stream 1's fetch trajectory when its input rate is 50 Mbps (N=10).

$500/N$ Mbps and $300 - 500/N$ Mbps respectively, and the fraction of stream 0 and 1 traffic of the total is varied with N. Shown in Figure 8a is the goodput perfor-

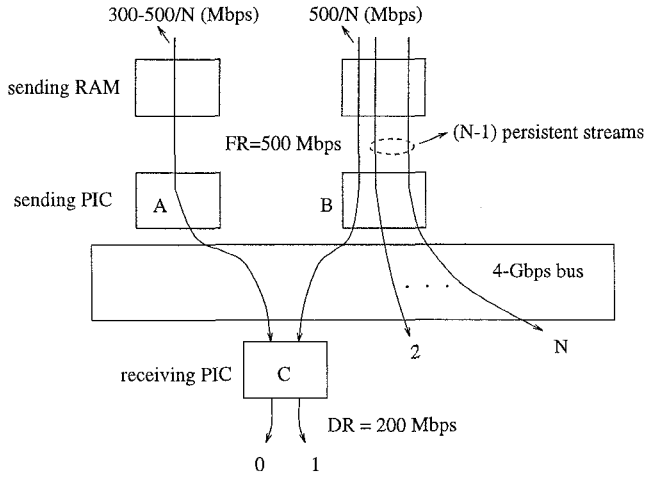

Figure 7: A hot-spot scenario with fetch-rate constraint.

mance of stream 1 with different control schemes when the input rate is varied from 20 to 100 Mbps. Ideally speaking, for any input rate in this range, the low-rate stream 1 should be able to receive the injected rate since the stream is allowed to be fetched at the input rate and the equal share of bandwidth at the bottleneck is 100 Mbps which is greater than or equal to the input rate. However, it is found that the high-rate stream 0 greatly starves the low-rate stream 1, resulting in as low a goodput of stream 1 as 25 to 45 % of the input rate. This is because stream 1 traffic can be moved from the RAM to the BIC only if the control is turned OFF so that this might result in loss of fetch opportunities which occur periodically. This behavior of stream 1 at the fetch point is illustrated in Figure 8b as the case with physical control and the associated control cycle is also plotted at the bottom. In contrast, if we remove all the persistent sources, the stream 1 gets the entire fetch bandwidth when the control is in OFF state and the fetch opportunities are enough to catch up to the input arrival rate as shown in the trajectory with physical control and without fetch-rate constraint in Figure 8b. Any ON/OFF type of flow control can potentially starve the streams with fetch-rate constraint since the streams require a higher rate than their input rate for fetch duing control OFF periods. We argue that such starvation should be limited to the streams responsible for the congestion. As found in the above, the problem with the physical control is that the low-rate stream is overwhelmed by the high-rate stream in sharing bandwidth, although it is less responsible for the congestion. The logical control scheme has an intrinsic advantage in such a scenario since the high-rate stream is more likely controlled than the low-rate stream. As shown in Figure 8a, with the logical scheme, the constrained low-rate stream was able to get more bandwidth as the low-rate to high-rate stream bandwidth ratio decreases, and achieve almost 100% of its input rate with a ratio of 0.2 (i.e., 50-Mbps stream 1 and 250-Mbps stream 0) or less. The fetch trajectory of the 50-Mbps stream 1, plotted in Figure 8b, shows that the stream 1 is hardly controlled and hence fetched almost periodically at 50-Mbps rate.

Finally, we examine the performance of the hybrid scheme with the same hot-spot scenarios and show that this scheme has the desirable properties of both the physical and logical schemes. We found earlier that

with the physical scheme, the occupancy at the hot-spot buffer has a very small down-swing below LTH since the streams are resumed rapidly physically. This observation leads us to design the thresholds for hybrid control such that (LTH, HTH, TTH) = (1, 6, 16) with lowered LTH. If we lower TTH to HTH, the hybrid scheme reduces to the physical control scheme, whereas if we increase TTH to the BIC buffer size, the scheme is equivalent to logical control. Thus, in designing TTH, one can always tradeoff between the properties of the physical and logical controls. In the first hot-spot scenario, the hybrid scheme with the above design led to much better loss performance than the logical scheme (see Figure 5a) because of both extra physical-control capability and lowered LTH. On the other hand, with increased number of streams from 20 to 40, the hybrid scheme performed worse than the physical scheme in preventing loss. In practice, however, one can always tune TTH to prevent loss for a given condition. In terms of hub-resource utilization, the hybrid scheme uses the least bus bandwidth as well as control buffer (see Figure 5b) since the sources are resumed physically and the control activation/deactivation is less frequent than with the physical scheme as observed in Figure 6. The hybrid scheme also achieves MAX-MIN fairness in bandwidth allocation among streams when there is enough fetch-rate available at the send side (see Figure 5c) and low switching delay (see Figure 5d). In addition to the above desirable properties, the hybrid scheme was able to maintain the throughput of the low-rate stream as effectively as with the logical scheme (see Figure 8a) when there are fetch-rate constraints.

5 Conclusions

This paper examined the issue of flow control in a high-speed bus-based ATM switching hub. The switching fabric is a dual-bus with slots for diverse port cards. Due to the potentially large switching capacity needed, there can be a significant discrepancy in the switching fabric speed and the port card speed. This can result in buffer overflows at the receiving port buffer and consequently high losses in the switching fabric. Adequate flow control mechanisms are hence necessary to maintain a lossless switching fabric.

We first examined two different flow control strategies and highlighted their strengths and weaknesses. Then we considered a third hybrid strategy which combined the strengths of the first two strategies. In the first flow control scheme, which we refer to as physical flow control, all streams destined to a receiver with buffer problems such as high occupancy and loss are shut down till the congestion is cleared. This scheme has the advantage that loss is severely limited but has the disadvantage that high-rate streams arriving at this buffer can in some circumstances starve lower-rate streams. In a second strategy, referred to as logical flow control, streams are implicitly selected based on their rates and shut down. This scheme has the advantage that the higher-rate streams will eventually be shut down more often and hence cannot overwhelm the lower-rate streams. However, as we will see, loss is not easily controlled in this scheme. In addition, the operation of the scheme requires significantly more bus bandwidth and high-speed buffer than that of the physical scheme. Finally, in a hybrid strategy, we combine physical and logical flow control with logical control activated first and physical control activated later when the logical control is unable to limit the buffer occupancy and loss. We show that this hybrid strategy has the desirable properties of the physical and logical control schemes and hence is the recommended choice for flow control in the setting of interest.

References

[1] G. David Bergland et al., "A Technology Platform for Providing Broadband Communications Services," *AT&T Technical Journal,* Nov./Dec. 1993, pp. 48-56.

[2] Northern Telecom's Magellan Passport. *http://www.nortel.com* .

[3] M. Laubach, "Classical IP over ATM," IETF RFC 1577, January 1994.

[4] S. Chong et al., "Architecture and Performance of a Novel ATM Switching Hub: Part I (Physical and MAC layers)," *Preprint,* March 1996.

[5] Fore
Systems Case Study: Andersen Consulting - Technology Park, *http://www.baynetworks.com /Corporate/Solutions/Consulting/andersen.html,* 1996.