FLORAX- Flow-Rate Based Hop by Hop Backpressure Control for IEEE 802.3x

Duke Lee, Sinem Coleri,Xuanming Dong, and Mustafa Ergen
Ph.D. Candidates
{duke, csinem, xuanming, ergen}@eecs.berkeley.edu
Phone: 510-64{2-5649,3-5887,3-5887,3-5889} Fax: 510-642-6330
Department of Electrical Engineering and Computer Science
University of California, Berkeley
Berkeley, CA 94720

Abstract

The standard IEEE 802.3x is introduced to manage XON/XOFF flow control mechanism in full duplex gigabit Ethernet. In a IEEE 802.3x network, a downstream station can send XOFF messages to upstream stations to stop them from sending data until a specified time slot has passed or the downstream station receives a XON message. The development of efficient XON/XOFF flow control mechanism is still an active research topic. This paper describes FLORAX, a IEEE 802.3x XON/XOFF control scheme with proposed modifications. It introduces a hop-by-hop backpressure control based on flow rate in order to fully utilize the performance of large scale LANs. The control strategy of FLORAX is based on applying backpressure depending on the bandwidth of the flows, where flow is defined to be a source - destination MAC address pair. FLORAX allows the use of Service Level Agreement (SLA) for quality of service. FLORAX is fair in the sense that it distributes the bandwidth among flows during congestion. The simulation study is performed to compare the performance of earlier approaches with the FLORAX. Experiment results demonstrate that FLORAX achieves fair bandwidth distribution for the duration of transmission, efficient use of link bandwidth, and reduction in packet loss.

1. Introduction

LANs at the edges of the network are growing in size with help of new technological advances in the high performance switches, new standards such as Virtual LANs, and the standardization of full duplex gigabit Ethernet (IEEE802.3x [1]). For instance, at the University of California at Berkeley, campus wide network is being converted into a single large scale LAN.

In high bandwidth large scale LANs, hop-by-hop flow control is essential for full utilization of the bandwidth due to the following reasons. First, the end-to-end congestion control of TCP using sliding window does not scale well in gigabit range. Second, the round trip time of most of the TCP flows is short since most of the traffic occurs inside the extended high speed LANs. This results in short mean round trip time, and causes fast window openings giving rise to burstiness and packet loss in TCP [5]. Third, the rate mismatch brings about bottleneck at the switch and the low speed link may cause packet drop in this switch [7].

Hop-by-hop flow control, also known as backpressure flow control, is implemented in the data-link layer of the OSI reference model. Before buffer gets full, a congested switch temporarily prevents input switches from sending packets in order to prevent data packet drop. In particular, in IEEE802.3x standard, congestion control is invoked by triggering XON/XOFF flow control messages. The XOFF flow control message is sent to the upstream when buffer exceeds the upper threshold. When a switch receives an XOFF signal, it pauses sending packets until it receives an XON signal from the same switch or until the time in XOFF message expires. The XON signal is triggered when the buffer at the congested switch descends below the lower threshold.

In this paper, we discuss previous proposals for improvement of the backpressure flow control of 802.3x and compared to Flow-Rate Based Hop-by-Hop Flow Control (FLORAX) in terms of throughput and quality of service. In FLORAX, we define the flow id to be source-destination MAC address pair. The rate based backpressure implemented by estimating the throughput of each flow. The flow-based control also enables per-flow quality of service. For instance, an operator is able to set up an agreed bandwidth per-flow specified in service level agreement (SLA).

2. Definitions

In this paper, the fairness of a congestion control is defined to be fair distribution of bandwidth among competing flows during congestion. The congestion is modeled in terms of buffer occupancy. A switch is said to be congested if buffer occupancy exceeds an upper threshold. A flow is said to be a congesting flow if flow rate exceeds fair distribution of bandwidth, and non-congesting flow if the flow rate is below the fair distribution of bandwidth.

3. Related Research

Backpressure can be classified as selective and non-selective. During congestion, selective scheme blocks a subset of flows, whereas non-selective scheme blocks all the incoming flows. Rate-based congestion control

0-7803-7600-5/02/$17.00©2002 IEEE

[2] is an example of selective scheme where it provides dynamically adjusted transmission rates for each flow passing through the switch by using feedback information coming periodically from the neighboring switches. The transmission rate of each flow is decided based on the buffer occupancy information in the feedback signals. The drawback of this algorithm is the computational overhead and huge buffer size required to keep buffer occupancy information of each output port of the neighboring switches, and an extra periodic messaging burden in the network.

Studies cited in [8], [9], [10] are restricted to non-selective scheme. When applied to homogeneous networks with matched link speed, non-selective flow control helps to improve the performance reducing packet losses. However, when link utilization is high, bandwidth is not distributed fairly among the active stations. In addition, mismatched input link rate restricts the throughput because of the head of line blocking at the slower link.

Selective backpressure is studied in [5] as destination MAC address backpressure. Destination MAC address backpressure prevents the degradation in the performance of flows going to another destination by selectively blocking packets for the destination of the most aggressive flow. Although this scheme prevents unnecessary control of flows going to other destinations, it does not discriminate between the sources, which results in punishing non-congesting source as well as congesting sources.

Several hybrid schemes use both hop-by-hop control and end-to-end control in order to satisfy QoS requirements of real time applications. One of the schemes presented in [3] modifies 802.3x hop-by-hop flow control and the ECN mechanism for TCP end-to-end responsiveness. This scheme provides virtual pipes for the real time data flow and allow TCP traffic to flow in the remaining bandwidth by marking the ECN bit of their packets in case of congestion. In spite of the benefits as shown in [3], LAN switches need to be aware of TCP. This requires increased data processing. Furthermore, it is a violation of abstraction for layer styled communication model.

4. FLORAX

As mentioned in previous sections, the main goal of XON/XOFF messages is to avoid the packet drop. When congested, a switch blocks upstream by sending XOFF messages. When buffer occupancy falls below a certain level, a switch turns on the upstream by sending XON message.

The main difference of FLORAX from the original XON/XOFF scheme is the selectiveness of throttling based on flow rate. A flow identified by destination and source MAC address pair rather than IP address pair. The primary aim of FLORAX is to distribute bandwidth evenly among flows during congestion. We assumed that the sender informs upper layer in order to decrease its window size upon receiving XOFF message in TCP.

4.1 Motivation

4.1.1 Identification of Congesting Flows

As briefly mentioned above in Section [3], a congesting flow can trigger non-selective or destination-based XON/XOFF scheme to throttle non-congesting flows. When the buffer occupancy reaches high-threshold level, the congested switch sends XOFF to all flows regardless of their transmission rate.

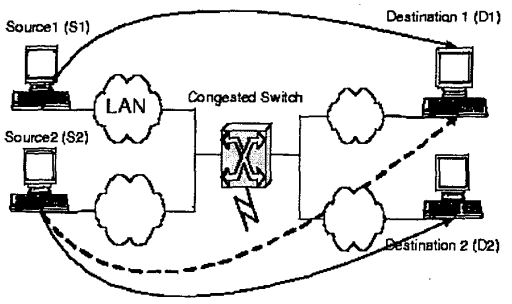

Figure 1: Scenario for Upstream Resource Sharing and Vulnerability to Non-conforming Devices

To see this, consider the situation in Figure1. Both Source1 (S1) and Source2 (S2) are sending packets to Destination1 (D1). At the same time, S2 sends packets to Destination2 (D2). Suppose also that S1-D1 transmits at very high rate causing congestion. With the simple XOFF control, all flows will be stopped due to the S1-D1 flow. With destination based flow control, the unnecessary backpressure to S2-D2 flow can be prevented. However, the performance of S2-D1 is still degraded. The FLORAX finds the most aggressive flows and only stops these flows without affecting the non-congesting flows.

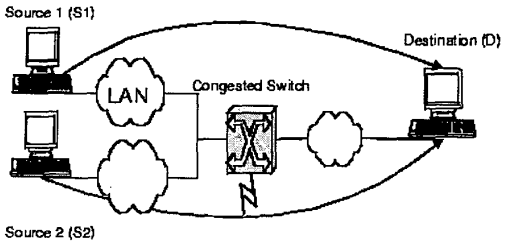

Figure 2: SLA

4.1.2 Enable SLA Provisions for Flows

Simply stopping the most aggressive flows is not always the best strategy. This is because different flows in a LAN may have different QoS specifications. Using SLA is a good way to optimize the aggregate satisfaction of various applications running on top of the network by providing differentiated service. For instance, user of exported file system may be willing to pay more for smaller latency than email users. To illustrate the use of SLA, let's consider the scenario in

Figure2. Both S1 and S2 send to D. S1-D and S2-D have agreed bandwidth of 100 Mbps and 10 Mbps, respectively. These two flows merge on a link that has capacity of 110 Mbps. Without SLA, when S1 sends at rate 90 Mbps and S2 sends at rate 30 Mbps, S1-D will be chosen to be XOFFed. With SLA, S2-D is XOFFed since it does not conform to agreed bandwidth. Use of SLA gives preference to flows conforming to their agreed bandwidth. In case of over subscription, SLA does not reserve bandwidth for a flow but gives better service to flows with higher agreed bandwidth.

4.1.3 Vulnerability to Non-Conforming LAN Device

The simple congestion control scheme that does not discriminate among flows are more vulnerable to non-conforming LAN devices in the way that non-conforming devices can degrade the performance of flows from conforming devices. Consider again the scenario in Figure1. Both S1 and S2 send files to D1 with rate 100 Mbps and 10 Mbps respectively. If S1 does not respond to XOFF control messages and keeps sending packets at rate 100 Mbps, S2 will decrease its rate. The FLORAX can be used in building resilience to such non-conforming behaviors. A flow that does not respond to XOFF messages can be singled out with the flow based XOFF control. This prevents the unfair decrease in throughput of other flows sharing the same output buffer with this flow. According to FLORAX, if the packets of this flow are seen in the buffer despite the XOFF message, they are dropped without affecting the performance of other flows.

4.2 *Protocol Description*

4.2.1 Elements

The following are the elements required in each outgoing buffer:

Flow Table/List: Flow Table/List keeps the information for each flow. Entries for each flow include the estimated rate, start time of current burst of traffic, the time when last packet is received, total number of bytes received since the start if burst and other parameters for SLA such as reserved bandwidth. We expect that the flow table size is manageable since the size of flow table is proportional to number of active flows.

The reason to keep the estimated rate for each flow is that the packets in the buffer are not a reliable representative of traffic characteristics of flows. According to [4], average latency of most of the gigabit LAN switches is around 16μ seconds. This means that the average buffer size at the switch is $\frac{16*10^{-6}*10^{9}}{8}=2*10^{3} bytes$, which implies that the number of packets in the buffer does not contain enough history for rate estimation.

XOFF Table/List: XOFF Table/List keeps track of the XOFFed flows. This table is used to restore XOFFed flows.

XOFF Control Messages: XOFF messages are used to reduce or block flows. Each message identifies the flow by its source-destination address pair. XOFF control messages specify the fair bandwidth for the flow, the rate at which congested switch will accept packets from the flow. XOFF control messages contain expiration time.

XON Control Messages: XON messages are used to restore XOFFed flows. Each XON message identifies the flow by its source address-destination address pair.

Upper Threshold: If the buffer occupancy exceeds the upper threshold, the all congesting flows are notified to stop or reduce its rate via XOFF message.

Max Upper Threshold: If the buffer occupancy exceeds the max upper threshold, all flows are blocked via XOFF message.

Lower Threshold: If the buffer occupancy is reduced below the lower threshold level, all XOFFed flows, if any, will be restored via XON messages.

4.2.2 Outline of Algorithm

For each received packet,

```
1  estimate transmission rates
2  if (nonconforming)
3     drop packet
4  if (buffer_occupancy > max_upper_threshold)
5     send XOFF to all flows
6  else if (buffer_occupancy > upper_threshold)
7     for each active flows
8        estimate transmission rates
9     send XOFF to all congesting flows
```

For each transmitted packet,

```
10 if (buffer_occupancy < lower_threshold)
11    restore all XOFFed flows
12    reset upper_threshold
```

At each switch running FLORAX, corresponding output buffer occupancy is checked against the upper and the lower threshold levels at reception and transmission of packets respectively. When the buffer occupancy exceeds the upper threshold, the congested switch calculates what fair rate should be for each congesting flow and sends XOFF messages. After XOFF notification, the congested switch enforces the fair rate by simply discarding packets from nonconforming flow.

Once XOFF messages are sent out, the total input rate is less than the available bandwidth given that our estimation of rates are correct. Thus the buffer occupancy is expected to decrease. In case the estimation of rate is incorrect or new flow arrives

during the congestion, the buffer occupancy can actually exceed the maximum upper threshold. In this case, all flows are stopped to prevent buffer overflow as shown in line 4-5 of the outline. Despite that fact that the total input rate maybe forced to be lower than the link output rate during the congestion, the link utilization does not decrease since output buffer is not empty during the entire congestion period.

To hedge against potential packet loss for XON packets, each XOFF packet also holds expiration time. The flows that have been XOFFed will either be restored when this time expires or XON is received.

The following formula are used for the rate estimation:

ER=TB/(CT-SB)

where,
ER=Estimated rate
TB=Total Burst Received Bytes
CT=Current Time
SB=Start of Burst

In the implementation, a burst is considered to be a group of transmissions connected temporarily. More formally, the burst A is defined to be a group of transmissions where transmission x belongs to group A if and only if there exist transmission y in A such that x is not equal to y and temporal distance between x and y is less than a some period **T**. We simply reset the start of burst for a flow whenever the gaps between transmissions are separated by **T**. The reason why we find the rate based on burst is to determine the flows that are bursty at the time of congestion.

5. Performance Evaluation

We evaluate the performance of FLORAX by using the network simulation tool, NS-2 [11].

The topology of network used for the performance analysis, which is depicted in Figure 3, consists of 5 switches. The assumptions made for the simulation are as follows:

1. Switches buffer packets when the incoming traffic rate is higher than the outgoing traffic rate.
2. The traffic sources can trigger TCP to reduce the packet generation rate upon reception of backpressure signals.

In our simulation scenario, the traffic patterns are chosen to demonstrate advantages of FLORAX scheme over the FIFO/Droptail and simple XON/XOFF scheme [1]. There are four TCP flows and one UDP flow that are concurrently active. Flows 1,2,3,4 are TCP flows and flow 5 is UDP flow as shown in Figure 3.

Flows 1,2, and 3 are destined to Destination B from Source A, Source B, and Source C respectively. We examine the bottleneck between switch D and switch E due to traffic merging and rate mismatch. Flows 4 and 5 are destined to Destination A to create a congestion between switch E and destination A due to traffic merging. In addition, we have chosen flow 5 to be congesting UDP stream to see the effect of nonconforming source (see figure3).

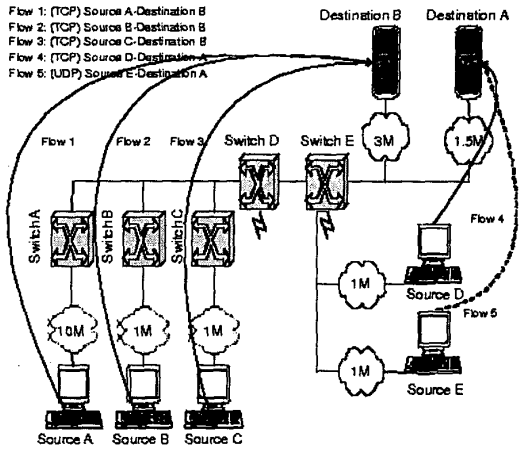

Figure 3: Scenario

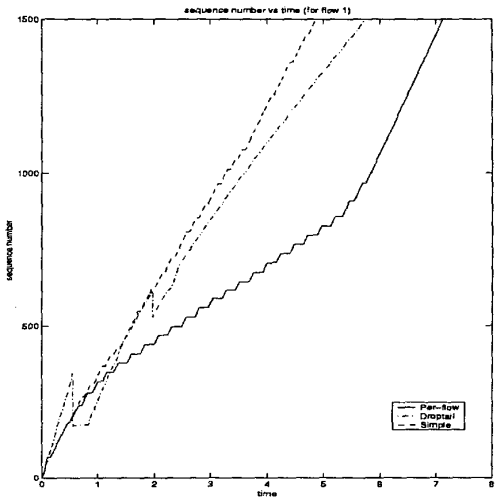

Figure 4: Sequence Number vs. Time for Flow 1

The available buffers in the switch LANs play a vital role in the backpressure-based hop-by-hop flow control. In the backpressure scheme, the buffers in the bottleneck links are augmented by sharing the available buffers of the upstream switches. The buffers are used to hold the packets due to temporary congestion in the bottleneck links. In FLORAX's algorithm, we assume a separate buffer for each output link.

Our aim in the simulation is to understand that how each flow control scheme affects packet drop, end-to-end transmission delay, and throughput for each flow. We also observe how each scheme reacts to congesting flows.

In Figures 4,5, and 6, the sequence numbers of the flows are plotted against time. The sequence number is measured at the destination. According to the figures,

per-flow scheme gives the highest throughput for the flows on the low speed link while giving the lowest throughput for the flow on the high-speed link. This is observed by the fastest increase of the sequence number for the low speed link and the slowest increase for the high-speed link.

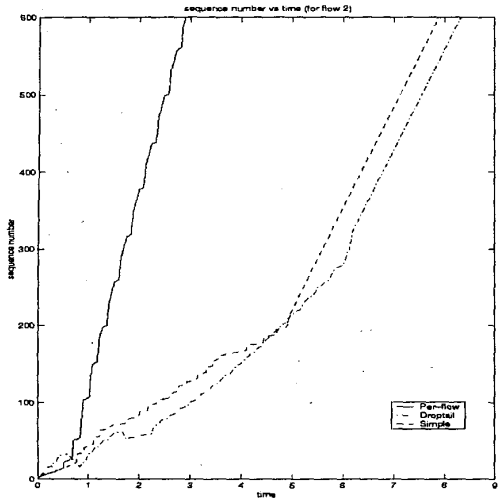

Figure 5: Sequence Number vs. Time for Flow 2

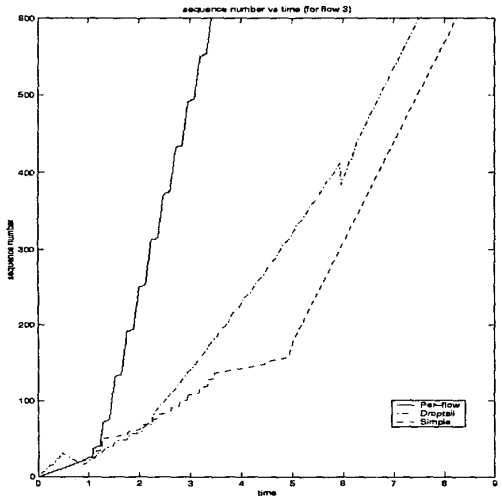

Figure 6: Sequence Number vs. Time for Flow 3

This is a desirable characteristic of the per-flow based scheme since the scheme does not reward congesting source. Instead, it attempts to distribute the bandwidth equally. One can see that the simple hop-by-hop flow control actually rewards congesting flows.

Another aspect to notice from the figures is that packet loss is extremely rare for backpressure schemes. The per-flow scheme didn't drop any of the packets from flow 1,2 and 3 for the duration of the simulation. The packet loss are represented by the dip in sequence number, as shown in Droptail plot in Figure 6.

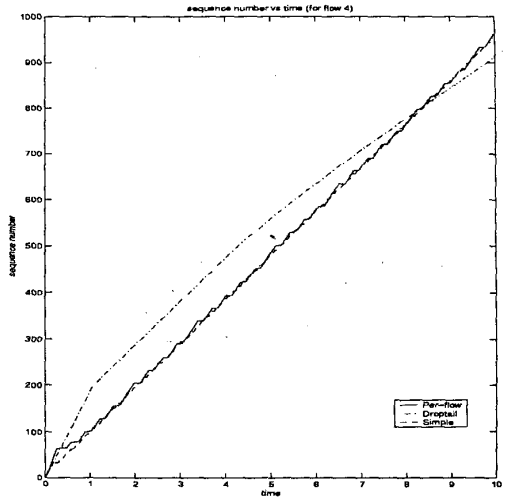

Figure 7: Sequence Number vs. Time for Flow 4

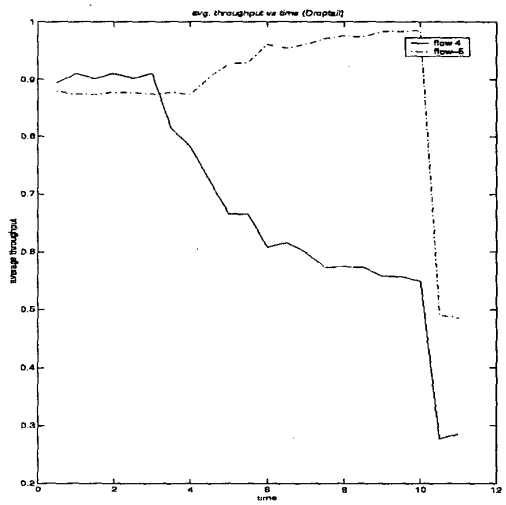

Figure 8: Average throughput vs. Time (Drop-tail)

Figure 7 displays a constant slope for sequence number vs. time plot for flow 4 for simple and per-flow case because of fair distribution of bandwidth between the two competing flows, flow 4 and flow 5. For droptail case, one can observe piecewise discontinuity of transmission rate at 1 sec when the stream starts, and at 4 sec when TCP decrease its window size.

The average throughput plots give an understanding of how a congesting flow dominates the throughput and how the simple and per-flow scheme improve the utilization. Figure 8 displays the congesting UDP stream (flow 5) dominating the bandwidth. The TCP flow (flow 4) decreases its window size due to packet loss whereas UDP flow does not employ any such congestion control. As a result, TCP rate is decreased allowing UDP to aggressively take the remaining bandwidth.

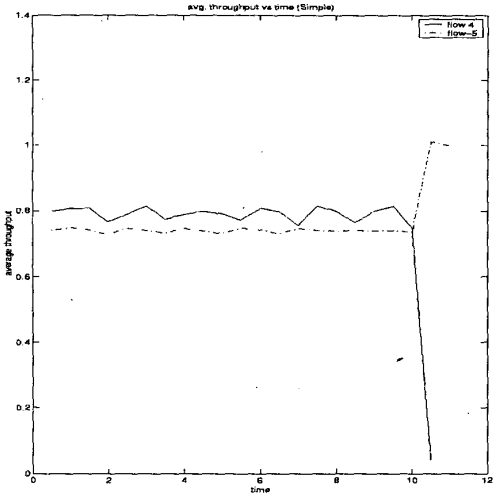

Figure 9: Average Throughput vs. Time (Simple)

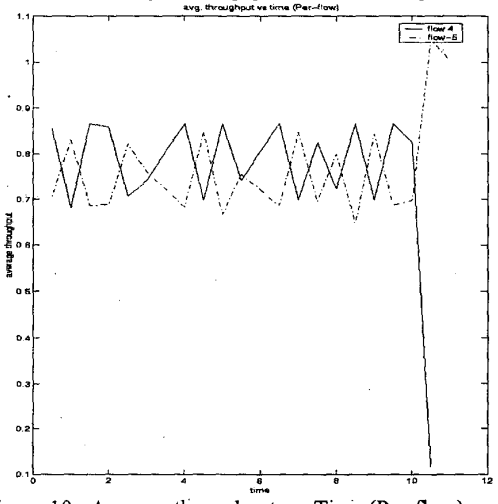

Figure 10: Average throughput vs. Time (Per-flow)

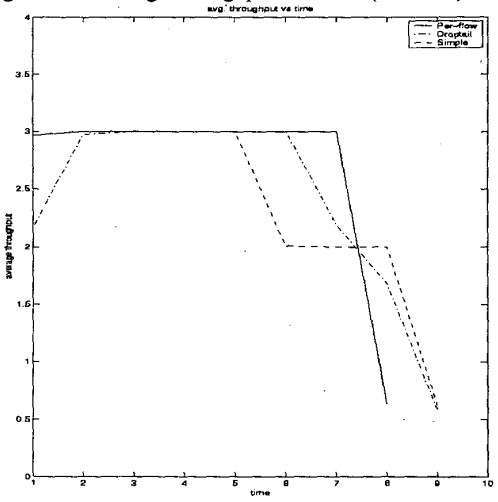

Figure 11: Average Throughput vs. Time

This undesirable dominance of UDP is addressed in the backpressure-based schemes. Both simple and per-flow backpressure scheme divide the throughput equally between flow4 and flow5 as shown in Figure 8 and 10. The fluctuation of the throughput of flow4 and flow5 in Figure 10 in per-flow case is due to the selective nature of the scheme which sends XOFF and XON signal alternatively to input flows.

From the throughput of the link between switch D and switch E in Figure 11, one observes less amount of time to finish transmission which indicates that per-flow scheme have better link utilization. Simple and drop-tail scheme finishes approximately at the same times.

5. Conclusion

In this paper, we compared different proposals for hop-by-hop flow control in MAC layer with a novel per-flow backpressure algorithm, FLORAX. We showed that previous backpressure schemes (including the destination based control) results in unnecessary blocking of non-congesting flows. In addition, those schemes are vulnerable to flows that do not conform to backpressure signals.

FLORAX is shown to have the following properties. No unnecessary congestion control is sent to flows that are not causing the congestion. Link utilization is high. And it is robust against flows that do not conform to backpressure signal. As shown through our simulation, under temporary congestion, per-flow scheme achieves fair bandwidth distribution, efficient use of link bandwidth, and reduction in packet loss.

Recognizing possible asymmetry between flows in terms of quality of service, we have implemented quality of service using SLA. SLA is used to stop the flows that are not conforming to their agreement in case of congestion. One of the SLA specifications, agreed rate, is the maximum expected rate of the flow.

Acknowledgments

We would like to thank Prof. Pravin Varaiya for his valuable comments and guidance throughout the project.

6. References

[1] IEEE802.3x *Specification for 802.3 Full Duplex Operation* IEEE Standard 802.3, 1998.

[2] P. P. Mishra, H. Kanakia. *A Hop by Hop Rate-based Congestion Control Scheme.* COMM ,1992.

[3] J.Wechta, M. Fricker, F. Halsall *Hop-by-Hop flow Control as a Method to Improve QoS in 802.3 LANs.* IEEE,1999.

[4] http: //www.bcr.com/bcrmag/08/98p25.htm

[5] W. Noureddine, F. Tobagi. *Selective Backpressure in Switched Ethernet LANs.* IEEE,1999.

[6] C. Ozveren, R. Simcoe, G. Varghese *Reliable and Efficient Hop-by-Hop Flow Control.* SIGCOMM, 1994.

[7] J. Walrand, P. Varaiya. *High-Performance Communication Networks.* Morgan Kaufmann Publishers, 2000.

[8] O. Feuser, A. Wenzel. *On the Effects of the IEEE 802.3x Flow Control in Full-Duplex Ethernet LANs.* IEEE, 1999.

[9] J. Wechta, A. Eberlein, F. Halsall and M. Spratt. *Simulation-based Analysis of the Interaction of End-to-End and Hop-by-Hop Flow Control Schemes in Packet Switching LANs.* Proc. of the Fifteenth UK Teletraffic Symposium on Performance Engineering in Information Systems, 1998.

[10] J-F Ren, R. Landry *Flow Control and Congestion Avoidance in Switched Ethernet LANs.* Proc. of ICC, 1997.

[11] http: //www.isi.edu/nsnam