Performance of a Switched Ethernet: A Case Study

M. Aboelaze
Dept. of Computer Science
York University
Toronto Ontario
Canada M3J 1P3
aboelaze@cs.yorku.ca

A Elnaggar
Dept of Electrical Engineering
Sultan Qaboos University
Alkhod 123
Oman
ayman@squ.edu.om

Abstract

In this paper, we study the performance of Switched full-duplex Ethernet. We assume a full-duplex, repeater that uses broadcast and two different kinds of traffic, regular data packets, and MPEG coded video frames. We investigate the network performance in terms of delay for both regular data and video frames (packets), and number of missed frames, We also investigate two methods for flow control, the first is basically stop and wait in which each node sends one packet only and waits until it has been already transmitted. In the other method, nodes can use frame bursting techniques and can send many packets (up to a full input port buffer length). Also we investigate the effect of the input port buffer length on the performance

Keywords¾ Fast LANs, 100base-T, Ethernet, CSMA/CD, Switched LANs, Full duplex repeater

I Introduction

No Network has achieved the popularity and success enjoyed by the Ethernet. Since its inception in Xerox Corporation in 1973, Ethernet grow up to become one of the most successful and widely used networks. Although Ethernet has a huge practical success, the mechanism of the channel sharing represents the worst possible scheduling discipline [5]. That resulted in usually a lower useful utilization and some very long delay for few packets (even at moderate utilization). Some packets will starve after the 16 collisions limit by the exponential backoff algorithm and never be transmitted [9].

That led to the idea of switched Ethernets. In switched Ethernets, the stations (nodes) are connected to a switch using point-to-point connections. The switches could be cut-through or store and forward. Cut-through switches start forwarding the message as soon as the destination port is known, while store and forward switches store the message first, and then forward it. Switches may be blocking or non-blocking. A non-blocking switch is capable of forwarding a message to the destination port as long as that port is free, while blocking switches may be not able to forward the message to a port although that port is free due to internal conflict in the switching fabric.

One of the most important factors in determining the switch performance is its collision domain. At one extreme is the *Port Switch*, which has one collision domain, and only one node connected to the switch can transmit at any time (that is not a real switch, it is rather a hub). In a half-duplex switched LAN, each switch port is a collision domain by itself. If only one station is connected to the port, a collision only occurs if both the node and the switch (port) decided to transmit to each other simultaneously (or within the one-way propagation delay of the link between the switch and the node).

In a full-duplex switching LAN, each station is connected to the port via 2 wires, one in each direction. In this case, there is no collision, not even carrier sense, since each node is the only one that uses that wire.

One major problem in switched networks (especially Ethernet) is the flow control mechanism. In shared media Ethernets, the CSMA/CD mechanism was used as a sort of flow control. If a frame is transmitted without a collision that frame is delivered to the destination. If the load increases and many stations attempting transmission simultaneously, collisions result and nodes have to slow down according to the binary exponential backoff algorithm.

In Switched Ethernets, although we might have decreased (or eliminated in case of full-duplex switches) collisions, we face another problem, namely the buffer overflow problem. This problem could be the result of a the switching speed of the switch fabric is being less than the aggregate speeds of all the ports, or when many ports are sending to the same destination port simultaneously. An overflow means that the frame is lost. While it is always possible to leave the upper layers to recover from that error, it is much slower than dealing with it at the switch level. On the other extreme implementing a link flow control mechanism between the port and the node, however that would complicates the design and increases the cost. [6], and [7]. IEEE [1] has proposed implementing a PAUSE function in its MAC control for 1Gbps, where the switch sends a PAUSE message with a field containing how many time units to wait before starting transmission, the time units is defined as the time to send 512 bits.

Another problem with switched Ethernets is that the speed of the switching fabric should equal to the aggregate speeds of all the input ports. That improves the performance but it is very expensive to implement (especially with high-speed Ethernets such as 100Mbps and 1Gbps).

A compromise solution is the Full-Duplex Repeater. The full duplex repeater is similar to the switch except that it uses a high-speed bus to broadcast the incoming messages to all the nodes. That saves on both the switching fabric and on the time and hardware required for look-up tables to forward the message to the required port only. Every input port has a buffer to store the incoming packets before they are broadcast by the switch. With input buffer and a full-duplex connection between the repeater and the node, a node can send a message to the switch while receiving another message from the switch (notice that the incoming message may or may not be destined to this node).

In our study, we assumed a full-duplex repeater with up to a max. of 20 nodes connected to it. We assumed that the nodes are sending either data packets or real-time MPEG coded video signal. For the regular data we assumed exponentially distributed packet length and a Poisson arrival, the For the video sources, we assumed MPEG coded frames with a distribution that is introduced in [2].

For video sources, we assumed that every source sends 30 frames per second. These frames could be I,B, or P frames [4], if a frame is kept in the buffer waiting its turn to be transmitted for more than 333.3 msec. The next frame will arrive and the older frame is discarded. In our study, every frame is either delivered or discarded, we did not consider the macroblock level where parts of the frame (macroblocks) could be delivered while other parts are discarded. However we assume that the frame is discarded by the receiving node That means the discarded frames will consume part of the repeater bandwidth. The other alternative is for the sending node to discard them but that could be very difficult to implement especially with the fact that the node has a little, if any, control over the frame after sending it to the repeater.

First, we considered only data sources and compared the average delay with a stop and wait-like protocol where every node send only one Ethernet frame and waits for it until it is transmitted. Then we assumed frame bursting [3] where a node can send more than one frame (up to a maximum limit) and measured the delay of such a repeater.

Second we considered mixed data sources and MPEG encoded video sources. We calculated the average delay per packet for both types of traffic, we also calculated the percentage of missed video frames as a function of utilization. Note that since we are not using the exponential binary backoff algorithm, data frames are never lost or discarded which is one of the main advantages of switched Ethernet.

The rest of the paper is organized as follows, in section II we briefly describe the full duplex repeater. In section 3 we presents our results for the case without frame bursting]. In section IV we present our results for the frame bursting case. Section V is a conclusion and future research.

II Full Duplex Repeater

In a full duplex repeater, every node is connected to the switch port using a full duplex connection. If a node wants to transmit, it sends the packet to the switch port, where it is stored in the port input buffer. A central arbitrator schedules the messages at the different ports for transmission on

the bus. The scheduling mechanism could be either a round robin or any other scheduling technique, in this paper we considered a round robin scheduling for simplicity. PAUSE mechanism is implemented to prevent input buffer overflow. The output buffer overflow is less of a concern since we assume that all the buffers operates at the bus speed and thus with a proper output buffer design, overflow can only occurs if the application program is too slow to drain the buffer. Which is not considered here since we only consider the repeater performance.

The difference between the full duplex repeater and a switch is that although in the case of a full duplex repeater every port is a collision domain by itself, which means there are no collisions. However, the switching fabric (a high speed bus) is shared between the different ports. And thus the repeater data rate is divided between the different ports. In a switch, usually there is a much more complicated switching fabric that supports operation at the full switch rate by all the ports simultaneously. Of course switches are more expensive than repeater due to the cost of the switching fabric and the hardware supporting the routing decision.

III Stop and wait protocol

First, we investigated a full-Duplex repeater with the input links and the bus speed at 100Mbps. We used a very primitive flow control mechanism in which any node can send only one message and then waits until the message is transmitted. A PAUSE or busy signal is used to block sending new packets. We assume that the bus controller scans the input lines in a round robin fashion. If there is a packet ready for transmission, the controller broadcasts it on the bus. Since the bus and the I/O are 100Mbps, we did not account for buffer overflow at the output as we explained it earlier. As we mentioned before, we assumed 20 nodes connected to the switch..

For regular data sources, we assumed an exponentially distributed packet size with an average length of 2000 bits. We also assumed a Poisson arrival with a variable arrival rate to control the load on the switch. For the video sources, we assumed a 30 frames per second, MPEG encoded video with a frame length distribution for the I,P, and B frames as mentioned in [KrT97]. The average data rate for MPEG encoded video sources is in the range of 1 to 1.5 Mbps. We also assumed that every node is either a data source or video source. We used CSIM simulation package [5] in our simulation.

Figure 1 shows the delay vs. utilization for data packets, Notice that although the bit rate for video sources is not veyry high, but because of a much larger packet size for video the delay for data packets sharply increases when there are 4 or more video sources.

Figure 2 shows the same for video packets, while Figure 3 shows the percentage of the lost video packets. In Figure 2 one notice that the average delay drops at utilization more than 60%, this is misleading since by looking at Figure 3, the percentage of lost packet sharply increases for utilization more than 50%, thus the delay drops mainly because there are a lot of packets that have been waiting for a very long time and eventually discarded by the destination node and not counted in the delay figure.

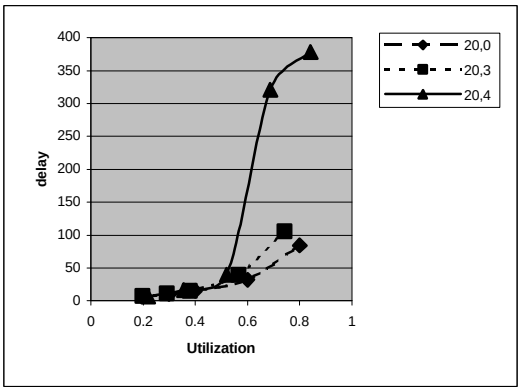
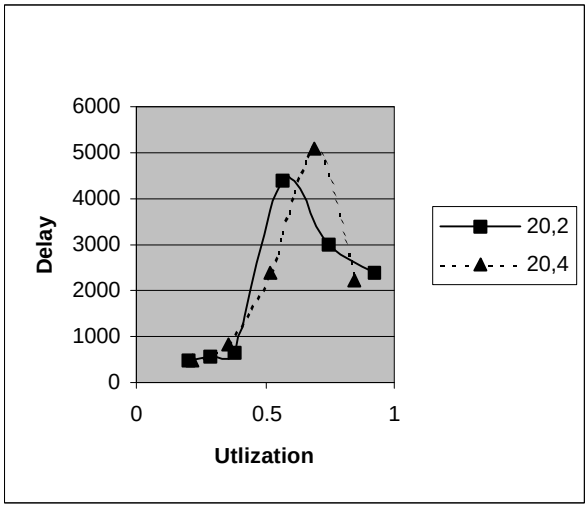


Figure 1: Delay vs. utilization for data packets

Figure 3 show also that full duplex repeater could be used to carry MPEG traffic as long as the delay is less than 40%. If the repeater is more heavily loaded than 40%, a large number of packets will be lost and the quality of the picture will degrade.



.

Figure 2: Delay vs. utilization for video packets

For pure data traffic, full-duplex repeater can support traffic up to 80% and 90% of the switch capacity with a reasonable delay, this is far beyond regular Ethernet using CSMA/CD We also simulated the case where the bus speed is more than 100Mbps, although the results were not reported here, that did not lead to much improvement, it just scaled back the utilization by a factor of 2.

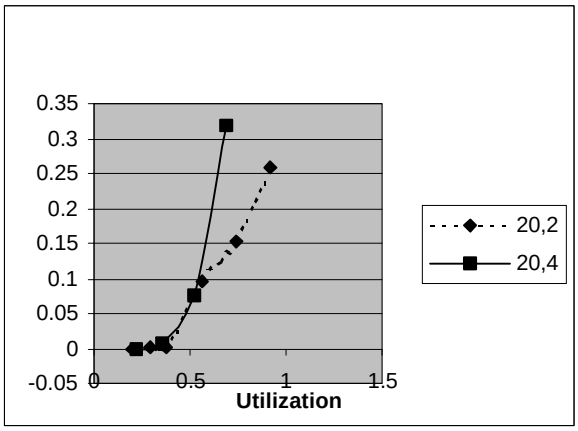


Figure 3: Percentage of lost video packets

IV Frame bursting

The previous flow control favors video packets over Regular data packet. Since the average size of the video packets are more than the average size of a regular data packets, and the nodes can transmit in a round robin fashion with every node sending one packet. Figure 1 shows that the average delay for a regular data packet when 4 nodes are transmitting video is more than doubled the delay at the case of 2 or zero nodes transmitting video for the same utilization level.

Frame bursting was proposed in [1] as a standard for Gigabit Ethernet LANs. In frame bursting a node that capture the media can transmit more than one frame without the need to wait and capture the media for every frame. In this paper we used a similar technique, where every input port has a buffer. The nodes can fill that buffer with one or more packets (but never exceed the buffer capacity). When that node turn comes, the switch will transmit all the packets in that node port buffer before moving to the next node.

The size of the buffer is a very important factor in determining the repeater performance. We consider two cases, the first where the buffer size is the same as the maximum Ethernet frame size, and when the buffer size is twice the maximum Ethernet frame size.

Figure 4 shows the utilization vs. the average data packet delay for a total of 20 nodes, 4 of them are video sources. Where Buffer=0 is the case where only one packet will be transmitted, Buffer=1 when we have an input port buffer equal to the maximum Ethernet frame size, and Buffer=2 is assuming an input port buffer that is twice the maximum Ethernet maximum frame size.

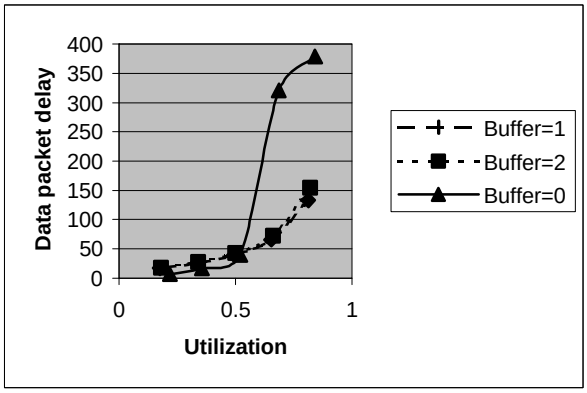


Figure 4 Utilization vs. Delay for 4 video nodes

At low utilization, where there is a little demand for buffer, one packet per transmission is the best policy. After the utilization approaches 50% and beyond, the performance for one packet per transmission severely deteriorate and Buffer 1 and 2 is a much better choice, with Buffer=1 has a slightly less delay

Figure 5 shows the video packet delay vs. utilization for different input port buffer size. Again, we notice that for low to moderate utilization a larger buffer could be advantageous, but at high utilization a smaller buffer is much better. However, if we compare Figure 5 to Figure 2 it is easy t notice that frame bursting is much better than one frame per transmission as in Figure 2.

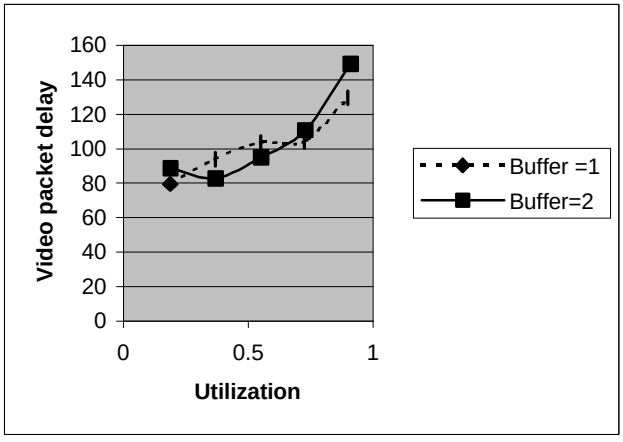


Figure 5 Utilization vs. Video packet delay using frame bursting

Figure 6 shows the average video packet delay as a function of the number of video sources. In this case there was no data sources, only video sources. Again we notice that there is more delay for larger input buffer size. Although the percentage of missed frames is , although it is not reported.

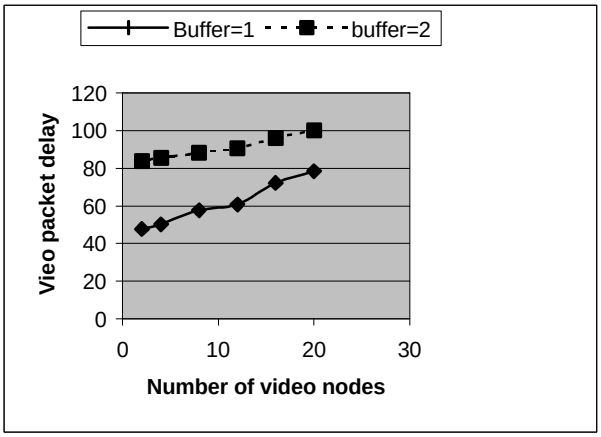


Figure 6 The average video packet delay vs. number of video sources (no data sources).

V Conclusion and Future Work

In this paper we presented our results regarding the performance of full-duplex repeater under a combined data/real-time MPEG encoded video traffic. We also considered the effect of the input buffer size on the repeater performance.

For future work, we are considering the effect of dividing the Ethernet packet into a smaller fixed size cells before transmission. This could be done independently from the higher layers, the switch/router will divide the Ethernet packet into fixed size cells and reassemble it back before it leaves the switch.

References

[1] IEEE Standard 802.3x, Specification for 02.3 Full Duplex Operation, 1997

[2] M. Krunz, and S. K. Tripathi "On the Characterization of VBR MPEG streams" *in ACM SIGMETRICS'97*, Seattle Washington, May 1997

[3] M. Molle, M Kalkunte, and J. Kadambi, “Scaling CSMA/CD to 1 Hb/s with Frame Bursting”, *Proceedings of the 22nd IEEE Conference on Local Computer Networks, pp 211-219, November, 1997*

[4] Networked Multimedia Systems, Prentice Hall, Upper Saddle River, N.J. 1998

[5] H. Schwetman, “Introduction to Process-Oriented Simulation and CSIM”, *Proceedings of the 1990 Winter Simulation Conference*, pp 154-157, December 1990

[6] R. Seifert, The use of Backpressure for Congestion Control in half duplex CSMA/CD LANS, Networks and Communication Consulting, 1996, available by anonymous ftp at ftp://ftp.netcom.com/pub/se/seifert/TechRept15.pdf

[7] R. Seifert, Gigabit Ethernet, Addison Wesley, 1998

[8] S. Shenker, “Some Conjectures on the Behavior of Acknowledgement-Based Transmission Control of Random Access Communication Channels” ACM SIGMETRICS’87 Conference on Measurements and Modeling of Computer Systems, pp 245-255 May 1987

[9] B. Whetten, S. Steinberg, and D. Ferrari . The Packet Starvation Effect in CSMA/CD LANs and a Solution, Proceedings on the 19th conference on Local Area Networks, pp206-217 Minneapolis, MN Oct. 1994