Performance Enhancement of Flow Control in 10GbE WANs

ROHAN DE SILVA
School of Information Systems, Technology & Management
University of New South Wales
Sydney 2052
AUSTRALIA

Abstract: This paper presents a new method of improving the flow control mechanism in 10 Gigabit Ethernet (10GbE) WANs. The ON/OFF control method prescribed by the standard leads to fluctuations of flow rates and the queue levels of the switches. Apart from the ON/OFF control, there are a few recent flow control mechanisms developed for the 10GbE networks in the literature. However, they do not comply with the present 10GbE standard as they require other quantities such as queue level to be included in the control frame. Our approach converts the fair share calculated by the switch to an equivalent pause time. As a result, our algorithm complies with the standard. Performance of this scheme has been studied via simulations and compared with the performance of the ON/OFF control method.

Key-Words:- 10GbE WANs, Elastic traffic, Flow control, Pause Frame

1 Introduction
10GbE uses the frame format and minimum and maximum frame sizes of the traditional Ethernet and has a maximum transmission range of 40 km. 10GbE has been specifically designed for the full-duplex operation only. As a result, 10GbE does not need to employ Carrier Sense Multiple Access with Collision Detection (CSMA/CD) because there cannot be any collisions.

Apparently, 10GbE offers numerous advantages over Asynchronous Transfer Mode (ATM). It is simpler, works well with Transmission Control Protocol/Internet Protocol (TCP/IP) and gives greater flexibility to the network administrators in adding new customers via standard Ethernet connections [1, 2]. As the network layer protocol used in 10GbE networks is Internet Protocol (IP), it is believed that 10GbE does not offer the reliability and the Quality of Service (QoS) that ATM and SONET provide [3].

However, an issue that has not yet been resolved properly by the research community is the effective flow control of the 10GbE WANs. As the speed increases, the delay-bandwidth product of the WAN links also increases proportionately. Thus, 10GbE WAN links will have much higher delay-bandwidth products than the existing ATM WAN links. If a proper congestion control scheme is not in place, the 10GbE WAN switches will need to have larger buffers than their ATM counterparts. Given that the frames can leave the buffer at a rate of 10 Gbit/s over the outgoing link, the queuing delay may not be higher than that of ATM.

Although there are several congestion and flow control mechanisms developed for the networks of IEEE 802.3x family [4-13], except [4, 13], they are all based on ON/OFF control. The method described in [4, 13] does not confirm to the standard as it requires special parameter fields in the PAUSE frame. Approach described in [14] to prevent deadlocks can also be considered as an indirect flow control mechanism. As we know from the experience of congestion control in ATM networks, ON/OFF type of mild flow control using PAUSE frames will not be sufficient for high speed WANs.

As such, we have investigated the possibility of improving the flow control mechanism in the 10GbE WANs and developed a new flow control mechanism by converting the fair share calculated by the switch to an equivalent pause time. As we will see in Section 4, the ON/OFF control method prescribed by the standard leads to fluctuations of flow rates and queue levels of the switches. Our algorithm does not suffer from these undesirable properties and complies with the standard. Performance of this scheme has been studied via simulations and compared with the performance of the ON/OFF control method.

The rest of the paper is organised as follows. After discussing the related background work and the research problem in Section 2, we introduce our flow controller in Section 3. Simulations and the analysis
of the results are found in section 4. The paper is concluded in section 5.

2 Background

Similar to the other members of the IEEE 802.3x family, 10GbE allows a weak form of hop by hop flow control using a type of control frames known as PAUSE frames. Hop by hop flow and congestion control enjoys benefits such as better accuracy and easily controllable queues compared to its end to end counterparts. A general discussion on comparison of hop by hop and end to end congestion control, can be found in [15].

PAUSE frames are released by the switch when the switch experiences congestion and they travel in the opposite (backward) direction to the directly connected upstream switches. This happens in most cases [4] when the switch buffer exceeds a certain threshold. This flow control occurs in the Medium Access Control (MAC) layer and the idea is to stop the upstream switch sending any more frames for a duration indicated in the pause_time field of the PAUSE frame [16]. If the queue level drops below a second threshold during the period specified by the pause_time, the switch can send a subsequent PAUSE frame with zero pause_time to the upstream switch. This will be treated by the upstream switch as a request to restart the data frame transmission. To the best of our knowledge, the IEEE 802.3x standard [17], does not specify that the PAUSE frames should only be released when the queue level exceeds or drops below a threshold.

Unlike in the Resource Management (RM) cells in ATM networks, there are no fields in the PAUSE frame to carry the queue level or explicit rate information to the upstream switch. There is also no provision for the PAUSE frames to travel all the way to the source or to the Source End System (SES). However, this may change in the future should the Ethernets that operate in full-duplex mode, whether they are Gigabit Ethernets or 10GbEs, extend right up to the desktop.

The fields of PAUSE frame [16, 18] are shown in Fig 1. In the figure, we can see that apart from the well known other fields in Ethernet frame format, we have two fields that are used in the flow control. The first is the opcode field which identifies the MAC control frame as a PAUSE frame. This field contains the value 00-01. Since all other values are reserved for future use, this is the only valid opcode at present.

Opcode is used to interpret the values in the parameter field. The second field is the parameter field which contains an integral number of octets. Only one parameter is currently used in PAUSE frame though many parameters could be included in the future should the standard evolves. This field in the PAUSE frame presently contains the pause_time. The pause_time indicates the length of the time the receiving station is requested to inhibit transmitting data after receiving a PAUSE frame. The amount of time that the receiving station inhibits its transmission of frames is equal to pause_time multiplied by 512 bit times.

<table>
<thead>
<tr>
<th>Length of field (Octets)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>PREAMBLE</td>
</tr>
<tr>
<td>1</td>
<td>SFD</td>
</tr>
<tr>
<td>6</td>
<td>DESTINATION ADDRESS</td>
</tr>
<tr>
<td>6</td>
<td>SOURCE ADDRESS</td>
</tr>
<tr>
<td>2</td>
<td>LENGTH/TYPY (88-08)</td>
</tr>
<tr>
<td>2</td>
<td>OPCODE (00-01)</td>
</tr>
<tr>
<td>2</td>
<td>PARAMETERS (pause_time)</td>
</tr>
<tr>
<td>42</td>
<td>RESERVED</td>
</tr>
</tbody>
</table>

Fig 1: Fields of PAUSE frame

The methods described in [5-12] are based on ON/OFF control. ON/OFF type of mild flow control using PAUSE frames will lead to oscillations in switch queues and link rates, and will not be sufficient if the links are underutilised. The method described in [14] has been designed to prevent deadlocks in the Internet. But it can also be considered as an indirect flow control mechanism. In that situation, the TCP connections will not need to time out and there will be a steady end-to-end data flow. Because of the growth of bandwidth hungry applications such as mobile agents and the increasing number of users connected to the LANs, we can expect that the 10GbE WANs will be full of traffic when they will be widely in place in 3-5 years as predicted by the proponents. In such an environment, poor flow control will lead to data loss and increased delay due to retransmissions and frequent timeout of TCP connections.

Some researchers have already observed this problem and have proposed to use priority for real-time traffic and similar algorithms that have been developed for ABR traffic in ATM networks for data traffic [4-13] in 10 GbE networks. However, they need the PAUSE frame to carry the queue level which is not compatible with the current standard.

An approach for the flow control in full-duplex Ethernet using the leaky bucket method has been
suggested in [9]. Leaky bucket approach does not take into account the round trip delay. Besides, this approach assumes that the upstream switch knows the bandwidth allocation for its aggregate flows reaching the downstream switch. This value is usually not known because there is no bandwidth reservation. A further problem is that it does not use priority for real-time traffic or separate buffers for real-time and data traffic. It has been reported in [10, 11] that under the UDP traffic carrying real-time data, the TCP traffic carrying non-real time data will suffer from bandwidth starvation if a priority mechanism or separate buffers are not used.

3 Development of the Flow Controller

Our approach here is to develop a flow control algorithm for 10GbE WANs without modifying or adding parameters to the PAUSE frame. Thus, the resulting algorithm will be conforming to the present 10GbE standard. We assume a fluid flow model and use a predicted value in the pause_time field. We incorporate prediction in the pause time data by using a target utilisation factor \( f(k) \) derived from a Predictive Dynamic Queue Control Function (PDQCF) [19]:

\[
f(k) = f(k-n-1) - \beta_f [z(k)-1] \tag{2}
\]

where \( \beta_f \) is a constant.

Target utilisation factor is then used with the available link capacity for data traffic of the outgoing link to calculate the target rate which is a modified value of the available link capacity. Since the value of pause_time is predicted, response will be faster and the oscillations of the queue level will be reduced and it is also expected that the PAUSE frames will be sent only very rarely. The feedback is only applied to non-real time traffic. We present this approach in the next paragraph.

We assume that the PAUSE frames are generated periodically by the switch with a period \( T \) and multicast to the upstream switches.

We define the variables:
Available data traffic rate of non-real time traffic in the outgoing link = \( C \) Gbit/s
Number of upstream switches that send data traffic through this link = \( N \)
Capacity of all the links in the network = 10 Gbit/s
Fair share calculated at the downstream switch by the controller at time \( t = FS(t) \) Gbit/s
Load factor \( z = \) Input rate of aggregate data traffic at the switch/target rate

Since we assume a fluid flow model, we can use the SAFARI controller [20] for the congestion control. The fair share calculated by the switch in this case is given by

\[
FS(t+1) = \max \{ [FS(t) + kAV(1-z(t))], AV \}
\]

with

\[
FS(0) = AV
\]

where \( AV = C/N \)

The target rate is given by

\[ TR = Cf, \]

where \( f \) is the target utilisation factor and \( C \) is the capacity of the outgoing link.

If the growth of queue level is \( \Delta Q \) and it is read each \( T \) sec, then assuming that the queue level is not less than \( CT \), we can write

\[
z-1 = \frac{\Delta Q}{CT}
\]

Thus, the pause_time written into the PAUSE frame is given by

\[
(10/FS(t+1)) - 1
\]

The downstream switch writes this quantity in the pause_time field of the PAUSE frame and multicasts it to the upstream switches each \( T \) seconds.

When an upstream switch receives this PAUSE frame, it completes the transmission of the frame that it has been delivering and leaves an inter-frame gap equal to the pause_time multiplied by the frame time of the last transmitted frame. It continues to leave gaps between each frame that it transmits according to this rule until it receives the next PAUSE frame. Then it continues the same process but uses the new pause_time. As the target utilisation factor \( f \) of the PDQCF converges to 1 [19], we have the upstream switch throughput at steady state as \( C/N \) which is the desired rate.

4 Simulations

We have investigated and compared the performance of this method with the simple pause time approach (ON/OFF) via simulations on a network consisting of seven switches (Fig 2). The parameters of the network and the flow rates of the upstream switches are shown in Table 1.

<table>
<thead>
<tr>
<th>Link</th>
<th>Propagation delay (time slots)</th>
</tr>
</thead>
<tbody>
<tr>
<td>switch 0 – switch 5</td>
<td>200</td>
</tr>
<tr>
<td>switch 1 – switch 5</td>
<td>300</td>
</tr>
<tr>
<td>switch 2 – switch 5</td>
<td>400</td>
</tr>
<tr>
<td>switch 3 – switch 5</td>
<td>200</td>
</tr>
<tr>
<td>switch 4 – switch 5</td>
<td>200</td>
</tr>
</tbody>
</table>

Table 1: Link delays of the network in Fig 2
Rates of incoming packets at the upstream switches (switch 0 to switch 4) were poisson distributed with a mean frame size of 576 octets and frame inter arrival times of 1.6, 1.6, 1.4, 2, 2.5 time slots respectively. A time slot is equal to the transmission time of a frame of 576 octets. When selecting parameters for the controller in eq (1), certain rules specified in [20] are used as follows: since the minimum round trip delay is 400 time slots, $T$ was selected as 400 time slots. Since the maximum round trip delay is 800 time slots, the parameter $k$ in eq (1) was selected as 0.5 (minimum round trip delay/maximum round trip delay). Parameter $\beta_f$ of the PDQCF of eq (2) was selected as $1/T$.

Simulation results for the two cases with the ON/OFF control and with proposed scheme are shown in Fig 3-4 and Fig 5-7 respectively. The traces in these figures clearly show the advantage of using proper flow control. As the queue level drops close to zero, the average queuing delay also becomes very small at steady state.
Fig. 6: Throughputs of the upstream switches with pause_time calculated using the proposed method.

Fig. 7: Utilisation of the outgoing link of switch 5 with pause_time calculated using the proposed method.

Just like with the ABR service in ATM networks, there may be bottlenecks in the upstream switches in a DiffServ domain as well. As a result, if we use an approach to provide an equal share bandwidth, it would not exploit the link capacity completely. Furthermore, achieving weighted fairness is extremely essential if we were to obtain QoS guarantees in the DiffServ networks. It could also be anticipated that the Ethernet technology will extend right through from one desktop at one location to another desktop on the other side of the globe. In such a scenario, should it happen that the end-to-end connection be full-duplex, then it may be possible that the standard evolves to include a PAUSE frame that can convey the pause-time right up to the source. In this case, we can easily adapt the SAFARI algorithm to achieve any general weighted fairness by allowing the pause_time in the PAUSE frame to be updated at each upstream switch by the new pause_time if it is greater than that is already in the frame.

5 Conclusion
The flow control method specified by the standard for the 10GbE WANs is the ON/OFF control which has its inherent disadvantages such as oscillations of rates and queue levels. The methods that have been put forward by the research community have the disadvantage of not complying with the standard. In this paper, we have developed and presented an improved flow control mechanism for the 10GbE WANs. As the simulation results suggest, this method is more efficient and results in eliminating oscillations in queue levels and rates.

References:


[20] de Silva, R., A Simple Approach to Congestion Control of ABR Traffic with Weighted Max-