INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN 2278 – 0211 (Online)

A Survey on Congestion Control Mechanism's Based on ECN

Yogesh A. Sale
Department of Information Technology, Walchand College of Engineering, Sangli, India
B. S. Shetty
Department of Information Technology, Walchand College of Engineering, Sangli, India

Abstract:
This paper is an exploratory survey of congestion management in multistage interconnection network (MIN) and techniques which are based on Explicit Congestion Notification. By studying congestion control techniques used in TCP implementation software and network hardware we can better comprehend the performance issues of packet switched networks and in particular, the multistage interconnection network.

Key words: *TCP, MIN, Congestion Management, Message Throttling, ECN*

1. Introduction

Network congestion appears when there is contention between several packets trying to use the same output link. If this situation remains for long, packets start to accumulate at the queues of the affected switches. As a consequence of the back pressure caused by the flow control mechanism, the packet advance in the previous switches is also delayed, generating the Head-Of-Line (HOL) blocking phenomenon, which prevents the advance of packets addressed to non congested links. Note that in high-performance interconnects for clusters, the communication model assumes a lossless network [8], so packets cannot be dropped to deal with congestion, as happens in other interconnect environments.

Congestion management has generated a lot of research and many mechanisms have been proposed over the years. Many of them have scalability problems, increasing the number of required resources as the network size rises. Others cannot manage congestion when it lasts for long. Recently, some mechanisms based on marking packets in transit have been proposed to detect and manage congestion in Infiniband. Unfortunately, these approaches do not guarantee, for all traffic distributions, that corrective actions are only carried out on those packet flows causing congestion. However, if switches with queues at both their input and output ports (CIOQ switches) [10] are used, as it is the case of many recent designs, a more selective packet marking mechanism can be applied. In this way, "cold" and "hot flows" could be distinguished, evenly distributing the available network resources among the devices that demand them, maximizing network throughput.

Explicit Congestion Notification (ECN) is a technique that just marks packets instead of dropping them as RED usually does. The idea behind implementing ECN instead of RED is to avoid packet drops, particularly where the delay involved caused by retransmission needs to be avoided.

Basically, the congestion management mechanisms based on the ECN [6] use a congestion detection strategy based on marking packets in transit, usually when a predefined threshold is exceeded. This packet marking action is carried out at the switches of the interconnection network. The marked packet will continue its travel toward the destination node, carrying out the congestion detection information. When this packet reaches its final destination, the ECN technique takes advantage of the Acknowledgment packets (ACK) sent back to the source to carry out the congestion detection information to the origins hosts. As a result of receiving marked ACK packets, those origin hosts will apply some corrective actions, normally based on limiting the injection rate into the network.

Infiniband switches [11][12] detect congestion on a Virtual Lane (VL) for a given port when a relative threshold set by the CCM has been exceeded. The threshold is specified per port between 0 and 15; 0 indicates that the switch is not to mark any packets on this port, 15 specifies a very aggressive threshold. Since the switch architecture affects how the level of congestion should be determined, the exact meaning of a particular threshold setting is left to the switch manufacturer.

Routers can mark two bits in the IP Type of Service (ToS) header field to signal whether or not congestion is occurring. TCP senders can then adjust their rate of transmission appropriately if they see that these bits are set to indicate a network congestion condition is occurring. The source response mechanism controls the injection of packets into the network in response to ECN information delivered to the source via ACKs.

Window-based congestion control is a common approach which adjusts the number of outstanding packets for a flow based on the congestion feedback. A window-based mechanism offers the benefit that packet injection is self-clocked and it limits the amount of buffer space that a flow can consume in the network.

2. Different Congestion Control Mechanisms Based on ECN

2.1. Input Packet Marking

To apply an IPM strategy [7] for packet marking and to warn about a congestion situation. Switches need to have buffers at the input links in order to apply this strategy. In particular, the proposed IPM strategy operates in three steps. First, a switch input buffer triggers packet marking each time it becomes full. Second, any output link that is requested for at least one packet in such a full buffer is classified as a congested link. Third, all packets stored at any input buffer at the switch that are destined to a congested output link will be marked. In response to the reception of a marked packet, the origin hosts apply injection limitation based on a window based technique combined with a waiting interval insertion technique.

Packets arriving at an input buffer are marked if the number of stored packets in the buffer exceeds a predefined threshold. This is performed by activating the Marking Bit in (MB=1) in the packet header.

2.2. Output Packet Marking

To apply OPM strategy for marking, Switches need to have output queues in order to apply this strategy. This proposal does not use any window-based technique to manage congestion. However, after receiving marked packets at origins host, sources reduce its injection rate by inserting WS (waiting slot).

This proposal does not use any window-based technique to manage congestion. However, after receiving marked packets at origins host, sources reduce its injection rate by inserting WS.

2.3. Marking and Validation Congestion Management Mechanism

The main goal of this new CMM is to properly identify the flows responsible for congestion, in order to apply packet injection limitation only at the source nodes that are actually causing congestion. Marking and Validation Packet Marking (MVPM)[2], that combines packet marking at input and output buffers in such a way that packets are marked at input buffers and validated at output buffers. To this end, 2 bits are dedicated in the packet header. To implement the mechanism in a standard interconnect, we can use any of the header bits usually reserved by the specs for vendor applications.

The MVPM strategy [9] works as follows: first, packets arriving to an input buffer are marked if the number of stored packets in the buffer exceeds a threshold. This is performed by activating the Marking Bit (MB) in the packet header. In the same way, when a marked packet is forwarded through a saturated output link, even in a different switch, we proceed to validate it by activating a second bit in the packet header, the Validation Bit (VB). We assume that an output link is saturated when the number of packets stored in its buffer exceeds certain threshold. Notice that a packet can be marked or validated several times, but never unmarked. Moreover, a packet cannot be validated if it has not been previously marked as shown in fig 1.

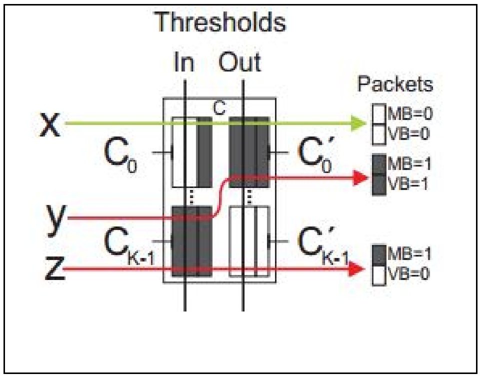

Figure 1: Marking packets by the MVPM strategy

MVCM proposes two phases of corrective actions as shown in table 1.The first phase is based on adjusting the packet injection rate by using a Dynamic Window (DW). It is based on the idea of limiting, for each flow, the number of outstanding packets into the network by using a window-based mechanism. In this case, the window size is dynamic, allowing to fluctuate between the maximum value (DWmax) and the minimum value of 1.If congestion persists after the window size is fixed to one, a second phase of actions will reduce even more the injection rate by introducing a waiting interval between the injections of two consecutive packets.

Ack bits		Types of flows	Actions
MB	VB		
0	0	Cold flow	No actions
0	1	Not possible	
1	0	Warm flow	Moderate(DW)
1	1	Hot flow	Imminent(DW+Ws)

Table 1: Corrective actions applied by MVPM

3. Congestion Correction Techniques

3.1. Window Based Technique

Basically, this technique defines a window size to limit the maximum number of outstanding packets per flow. The value of the window size depends on vendor criteria .Notice that the outstanding packets are those sent packets that have not been acknowledged yet. The window-based technique can be based either on a Static Window (SW) or a Dynamic Window (DW).If the window-based technique is based on a static window, the window size value will be kept fixed during all the time. To this end, the congestion management mechanism just controls the maximum number of outstanding packets per flow. Therefore, applying a window-based strategy based on a SW seems appropriate to palliate the congestion and simple to implement. However, the chosen value for the window size may not be the most appropriate value for any traffic condition in the network. Moreover, initializing the window with a value of one, as Renato’s proposal does, may negatively impact over the network throughput.

However, if a dynamic window is applied, the window size value can vary depending on network behavior. Initially, the window size will start with the maximum value defined for the target network configuration. When congestion appears, the window size will be progressively reduced until the minimum value of one. Later on, when congestion vanishes, the window size will gradually recover its initial value. Therefore, by applying a dynamic window strategy, the window size will vary between the values 1 and the maximum window size defined at configuration time.

3.2. Waiting Interval

The Waiting Interval Insertion technique allows to reduce the injection rate in a progressive way by injecting Waiting Slots (WS) between two consecutive packets. The size of a waiting slot depends on the vendor criteria, and it is defined at the network configuration time. Depending on the severity of congestion, the elapsed time between the injections of two consecutive packets will be increased or decreased. As long as the congestion ceases, the waiting interval between packet injections will be decreased until disappearing. Basically, the technique works as follows. When an origin host receives a marked ACK packet, it waits during a waiting slot before injecting a new packet into the network. If more marked ACK packets are received, then more waiting slots will be inserted, thus enlarging the waiting interval. It should be noted that the injection of new packets is forbidden along the waiting interval. Later, when unmarked ACK packets are received, the number of waiting slots between packet injections will be decreased.

4. Comparison between Different Congestion Control Mechanisms

4.1. IPM

When applying the IPM strategy [5], two drawbacks appear. First, a delay in detecting congestion, and second an incorrect identification of the flows truly responsible for congestion. This strategy produces a delay, because an output buffer has to be completely filled before any input buffer can detect congestion

4.2. OPM

The OPM strategy [3][4] will not be able to mark packets belonging to those flows till the congestion reaches the output buffer at the previous switch. So, that delay in marking packets could affect in its turn other flows or even other switches not involved in the initial congestion.

Both IPM and OPM strategies dedicate only 1 bit to mark packets in transit, they are not able to handle different levels of congestion and, as a consequence, to apply different corrective actions Depending on the severity of the congestion

4.3. MVCM

MVCM mechanism based on a more refined packet marking strategy combined with a fair set of corrective actions, that makes the mechanism able to effectively manage congestion regardless of the congestion degree. It dedicate 2 bits of packet header to mark and validate packets in transit

5. Conclusion

MVCM mechanism contributes with both a new packet marking strategy that efficiently detects the root of congestion and correctly classifies flows belonging to the tree congestion, and a fair set of corrective actions, that makes packets belonging to the flows responsible of congestion wait at their source hosts, instead of remaining blocked into the network.

6. References

1. M. Allman, V. Paxsm, and W. Stevens, "TCP Congestion Control "http://www.rfc-editor.org/rfc/rfc2581.txt, 1999.
2. J. Ferrer, E. Baydal, A. Robles, P. Lopez, and J. Duato, "Congestion Management in MINs through Marked & Validated Packets,"Proc. 15th Euromicro Int'l Conf. Parallel, Distributed and Network-Based Processing (PDP '07), 2007.
3. G.Pfister and V. Norton, "Hot Spot Contention and Combining Multistage Interconnection Networks," IEEE Trans. Computers,
4. vol. 34, no. 10, pp. 943-948, Oct. 1985.
5. G. Pfister et al., "Solving Hot Spot Contention Using Infiniband Architecture Congestion Control," Ion High Performance Interconnects for Distributed Computing, 2005.
6. J. Renato Santos, Y. Turner, and G. Janakiraman, "End-to-End Congestion Control for Infiniband," Proc. IEEE INFOCOM, 2003.
7. Sally Floyd, TCP and Explicit Congestion Notification, ACM Computer Communications Review. October 1994, p.10-23.
8. Jose Renato Santos, Yoshio Turner, and G.(john) Janakiraman.Evalutiation of congestion detection mechanisms for Infobahn switches. In Hewlett Packard Laboratories HPL-2002-224, 2002.
9. J. Duato, I. Johnson, J. Flich, F. Naven, P. Garcia, and T.Nachiondo, "A New Scalable and Cost-Effective Congestion
10. Management Strategy for Lossless Multistage Interconnection Networks," Proc. 11th Int'l Symp. High-Performance Computer Architecture, 2005.
11. E.Baydal and P.Lopez, "A Robust mechanism for congestion control: INC", in Proc.9th International Euro-Par conference, pp.958-968, Aug.2003.
12. Anderson, S. Owicki, J. Saxe, and C. Thacker, "High-Speed Switch Scheduling for Local-Area Networks," ACM Trans.Computer Systems, vol. 11, pp. 319-352, 1993.
13. W. Dally, P. Carvey, and L. Dennison, "The Avici Terabit Switch/ Router," Proc. Hot Interconnects, 1998.
14. http://www.infinibandta.org, 2011.