

A SYNTHESIS VLAN APPROACH TO CONGESTION MANAGEMENT IN DATA CENTRE ETHERNET NETWORKS

K. C. Okafor, T. A. Nwodoh

Department of Electronic Engineering, University of Nigeria, Nsukka.

Abstract

This paper discusses, develops and analyses technique for the control of congestion in data center Ethernet networks by priority tag assignment. In very large networks with heavy traffic, congestion control plays an important role in network resource management. There have been several proposals to address the challenge of traffic congestion particularly in the data center and internet applications. A scheme known as Backward Congestion Notification with Active Queue Management (AQM) was reviewed in relation to congestion management. The paper then develops a more reliable and stable candidate scheme for congestion control for Data centre Network (DCNs): Synthesis Virtual Logical Aggregation Network Segmentation (S-VLANS). The approach used is based on the measurement of QoS parameters with variation of load intensities alongside with the buffer sizes of the core switch. This scheme is able to control link utilisation by assigning tags to high priority end systems running different services, hence logically reducing the amount of queuing delay in the DCN networks as well as traffic overhead, thereby improving network performance.

Keywords: *Ethernet, Data Center networks, Synthesis VLAN Segmentation, Priority Tag Assignment*

1.0 Introduction

Contemporarily, data centre networks (DCNs) have continued to attract a lot of interest in the enterprise networking market segments. DCNs are used to provide data storage and file transfer where end stations are interconnected as clusters and bladed systems. In this case, FTP flows are very typical in the networks. Basically, DCN requires high throughput and low latency for efficient communications in relation to data transfers. Due to the large number of end stations configured in clusters, these networks are vulnerable to link congestions resulting in packet losses and timeouts that can severely jeopardize the overall system throughput.

This research develops an efficient congestion management methodology in switched DCNs based on traffic priority. This paper starts its investigation with these questions:

- Why do packets drops in switched DCNs and what are the effects in mission critical applications like:

ERPs, Application Servers, E-commerce servers in DCN networks, etc.?

- How do systems control traffic flow from access layer sources into switched data center networks without affecting network performance?
- What is the behaviour of switch buffers at the congestion points, and reaction points under network oversubscription considering the various performance indices?
- How can Agility in a congested switched links be achieved?

Basically, congestion is intrinsic to the operation of networks and is usually handled by a combination of algorithms at the link and network/transport layers. Link level algorithms alleviate "transient congestion" caused by the temporary oversubscription of a link due to a burst of packets arriving at a switch or router buffer. Transient congestion is effectively dealt with using buffers at the links and

dropping the excess packets, while sustained congestion is alleviated by transport protocols like TCP which reduce the transmission rate of a flow when its packets are dropped or marked. This combination of algorithms provides stable, scalable and fair bandwidth sharing to flows which use the Internet and data centres services. A representative sample of the literature may be found in the book [1], the survey papers [2], [3], and at the website [4].

Consider a different link level mechanism: when a buffer begins to fill up, a back pressure mechanism pauses the link(s) feeding the buffer, thereby preventing buffer overflow. This type of mechanism leads to no packet losses but, potentially, spreads congestion to upstream buffers and eventually to the sources. A detrimental effect of congestion spreading is that it is possible for a source, say S, which sends data at a high rate to cause a number of links to be paused and to adversely affect the throughput of other (non-congesting) sources which share these links with S. To mitigate this, it is desirable to have a mechanism for the network to signal to S that it should decrease its sending rate. Such a mechanism has been proposed: the Backward Congestion Notification (BCN) mechanism [5] which is currently being considered for deployment in switched Ethernets DCNs.

The purpose of this paper is to propose an approach to congestion management via Synthesis VLAN that will allocate resources to end systems based on priority tags. A conceptual model is developed and analyzed.

2.0 Related Work

2.1 The BCN Mechanism

The Backward Congestion Notification (BCN) mechanism is a way for mitigating the congestion spreading effects due to link-level pause by enabling the network switches to signal sources to reduce (or increase) their sending rates [6]. The

Figure 1, shows the Backward Congestion Notification with Active Queue Management (BCN-AQM) System Model proposed in [6]. It assumes random signal sources equipped with the rate regulators, which can be token-bucket traffic type. At the core switch, where congestion happens, a congestion detector, which is an up-down counter, is integrated in an ASIC FPGA switch to generate a feedback message.

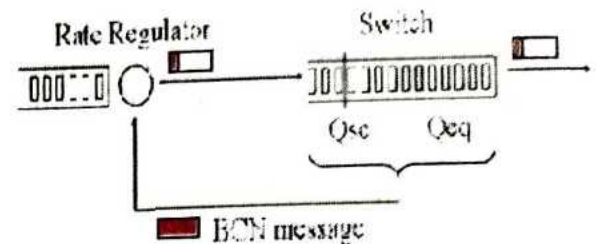


Figure 1: A BCN-AQM System Model

As shown in the Figure 1, the core switch monitors the length of its output queue, and decides whether there is congestion based on active queue management (AQM) scheme. If AQM sets its flag, hence congestion is detected; the switch now signals the sources with BCN message that contains traffic rate adjustment advertisements. The sources acknowledge and reacts to the received BCN/AQM feedback signals and updates the rate of its regulator (MAC address table).

BCN-AQM have a major limitation of increasing latency in a DCNs, owing to the presence of its feedback loop (Overhead of 0.311% due to BCN messaging, i.e. the traffic overhead in backward direction). Quality of service on the other hand, is highly influenced by the traffic pattern as resources are allocated based on the packets generated by remote sources in a DCN.

2.2 BCN Scheme Limitations

The structure and evolution of DCNs has been extensively studied [7], but little

work has been done on traffic priority in DCNs. The most closely related work to this paper was done by J. Jiang et al [8], [9]. The work extends the idea of Backward Congestion Notification by taking into account the characteristics of a distribution layer switch which rely heavily on high traffic density flow. However, the authors of [8], [9] did not explain how they will make switched DCNs logically assign priority tags to end sources as well as achieving bandwidth optimization without a feedback loop.

This research takes advantage of the limitation of BCN technique to achieve a high throughput DCN with low latency on priority based end systems. The observed limitations of the BCN scheme include:

- i. High Traffic overhead in reverse direction (0.311%) because of the feedback loop.
- ii. BCN ideal behaviour requires per-flow queuing
- iii. Network round trip time (RTT) is far greater than the flow duration.
- iv. Lack of Agility for server service provisioning.

3.0 Modeling Performance of DCN System

Since DCNs are meant for high speed traffic flow devoid of congestion, this work is aimed at designing a congestion control scheme: Synthesis Virtual Logical Aggregation Network Segmentation (S-VLANS) optimised for data center networks that scales gracefully with network capacity, providing high utilization, low queuing delay, high throughput, dynamic stability, and fairness among users.

The Synthesis Virtual Logical Aggregation Network Segmentation approach adopted in this work is quantified using both Process modeling with SimEvent and network modeling with Packet Tracer showing a typical industry

application. The modeling is handled collectively for each of the software tools. SimEvents [10] extends Simulink with a discrete event simulation model of computation. With SimEvents, process model of the proposed S-VLANS is developed to evaluate system parameters, such as throughput, resource utilization and average delays and latency. It is possible to configure entities with user-defined attributes and then aggregate entities and attributes to model data hierarchy and then perform the simulations of the DCN application. Also, in this work, Packet Tracer which is a simulation-based environment for developing network designs, making different configurations, and troubleshooting computer networks, is used for the S-VLANS priority tag assignments. Various types of configuration can be realized through the internetwork operating system (IOS) command line (CLI).

The network architecture employed in this paper is illustrated in Figure 2. This architecture had its origins based on the Cisco's design methodology [11]. The operation of network model can be explained briefly below.

Work stations and DCN servers are assigned to S-VLAN 10, S-VLAN 20, S-VLAN 30, S-VLAN 40, S-VLAN 50, S-VLAN 60 and S-VLAN 70. S-VLAN 10 to S-VLAN 60, constitute the access layer module which communicate via the transmission media (Ethernet Cable) to the attribute broadcast channel (length, source and destination), the FIFO queue links, the set attribute modules and the path combiner. From the output of the switch, the DCN servers go to the sink module. There are seven S-VLANS in this conceptual model.

For each S-VLAN simulation run, the buffer sizes of FIFO Queue S-VLAN switch is varied in consonance with the load intensities and the QoS parameters are computed via the command prompt of

MATLAB environment. Starting from a buffer size of 25 to 1000 in the SimEvent process model, this summed up to 10525 buffer sizes. Also, subsystems block set are created with the first S-VLAN run,

followed by the second S-VLAN, third S-VLAN, till the last S-VLAN (VLAN 70). Consequently, the trend pattern and responses are ascertained.

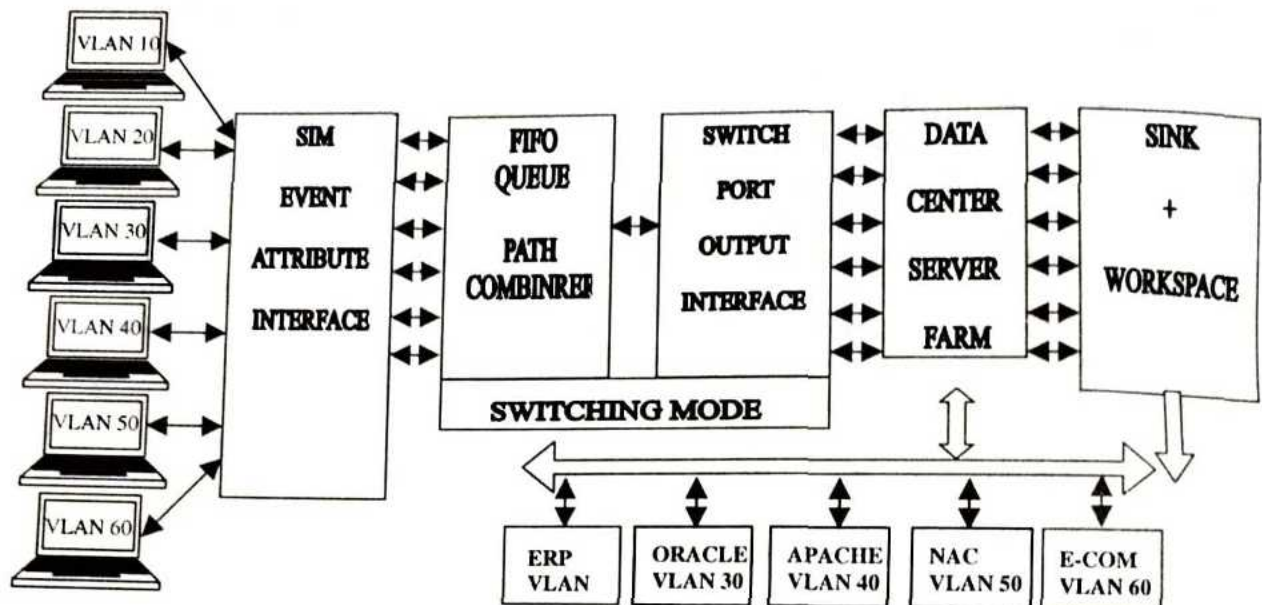


Figure 2: A DCN Network Model (Test Bed)

4.0 Simulation Parameters

The MATLAB SimEvent version 7.4.0 is used for the simulations while packet tracer is used to realise the priority tagging for the end systems and S-VLAN setup. The results obtained from the S-VLAN model are presented and the various plots are shown below.

Based on the buffer sizes, offered load intensity, throughput, queuing delay, latency and channel utilization, the following input parameters are utilized based on the workload of FTP applications:

- Traffic Distribution = Exponential
- Entity Type = Standard
- Generation Event Priority = 300 bytes
- Service Time = 1sec.
- S-VLAN Switch Password = "Cisco"
- Number of S-VLANs = 7
- Configuration Register = 0xF
- Number of Entity Output Port = 6
- Number of Entity Input Port = 6

- Packet generation rate for each node = 100 packet/s
- Maximum packet size = 1024 bytes
- Cable length between each nodes = 100m with Gigabit Ethernet Cabling.

A classless interdomain Routing (CIDR) IP allocation of a class C Address 192.168.10.14/29 is assumed. The IP design for the end systems is presented below:

- The Network Address = 192.168.10.0.
- The Subnet Mask = 255.255.255.248 by CIDR Notation.
- The Number of Subnets = $2^5 - 2 = 30$. (Where 5 = Number of Turned off Bits for Hosts).
- Number of Hosts Per Subnet = $2^3 - 2 = 6$ Hosts per Subnets.
- Block Size (Number of Valid Subnets) = $2^8 - 248 = 8$ Subnets.

4.1 Network Model IP Assignment (192.168.10.0/29)

There are seven workgroups considered in this paper. The corresponding valid host ranges are computed from the subnet table map, thus:

- Location A= S-VLAN 10 on Subnet 0. (Host Range = 192.168.10.1 --- 192.168.10.8).
- Location B= S-VLAN 20 on Subnet 8. (Host Range = 192.168.10.9 --- 192.168.10.14).
- Location C= S-VLAN 30 on Subnet 16. (Host Range = 192.168.10.17 --- 192.168.10.22).
- Location D= S-VLAN 40 on Subnet 24. (Host Range = 192.168.10.25 --- 192.168.10.30).
- Location E= S-VLAN 50 on Subnet 32. (Host Range = 192.168.10.33 --- 192.168.10.38).
- Location F= S-VLAN 60 on Subnet 40. (Host Range = 192.168.10.41 --- 192.168.10.46).
- Location G= S-VLAN 60 on Subnet 48. (Host Range = 192.168.10.49 --- 192.168.10.54).

5.0 S-VLANS Performance Metrics

Synthesis VLANS offers high bandwidth optimization with low latency and consequent high throughput in a congested link. A seven source star topology was used in the model. The seven S-VLANS are assigned to end system/workgroups locations based on priority tags as shown above. The simulation deals with seven different cases; the first one is devoted to the case in which all the workstations are connected to S-VLAN 10, the second case defines S-VLAN 20 membership (by port and MAC), the third case defines S-VLAN 30 membership by MAC address, the fourth case defines S-VLAN 40 membership (by both port and MAC), the fifth case is S-VLAN 50 membership (by port and MAC address simultaneously to get VLAN identification), the sixth case is S-VLAN 60 (by port and MAC) and finally S-VLAN 70 (by port and MAC). Starting from S-VLAN

10, the priority tags were assigned for resource allocation and congestion management. In the S-VLAN process model, the subsystems creation is used for the load intensities. (thus: S-VLAN 1, S-VLAN 2, S-VLAN 3, S-VLAN 4, S-VLAN 5, S-VLAN 6, S-VLAN 7, S-VLAN N). The figures in the next section are plots from the simulation process models.

6.0 Analysis of Results

The simulation was run for several times while varying the buffer sizes of the S-VLAN core switch from 25 to 1000 bytes for different packet generation rates (arrival rates) of nodes while varying the load intensities of the various S-VLANs as well. The Synthesis VLAN was seen to significantly regulate broadcasts and improve throughput particularly as the buffer sizes are varied for the congested core switch. With increase in load intensities, bandwidth optimization on the core switch is enhanced by S-VLAN. From the process model, the FIFO Queues holds packets based on service policy of S-VLAN; hence the switch is capable of determining the number of bits in each of the incoming buffers at the beginning of service of each packet arrival.

Figure 3 shows S-VLAN buffer size throughput plot. At the core switch, the output buffer can hold a minimum of 25 packets of 1024 bytes each. Essentially, the buffer sizes were varied starting from 25 packets to 1000 packets for various S-VLAN loads. It was observed that the location independent S-VLANs fairly assumed a gradual exponential rise in throughput due to load delay effects as the buffer sizes were increased. The S-VLAN curves are on top of each other with gradual transitions giving way to higher and efficient throughputs. From a buffer size of 750 packets and above, throughput of bulk traffic S-VLANS was significantly high. This guarantees fair allocation in the congested link. Thus, an increase in buffer size leads to a better performance network

considering the various S-VLAN load intensities.

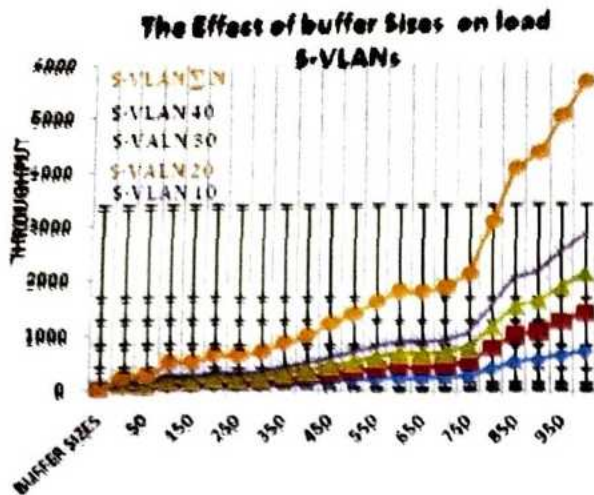


Figure 3: The Effect of Buffer Sizes on Load Intensity.

Figure 4 shows S-VLAN buffer size distribution in a pie chart. The variation is based on an increment of 50 packets after a previous buffer size selection. This distribution optimally reflects their respective traffic carrying capacities with respect to the priority tag distributions in the S-VLAN. It ranges from 25 packets to 1000 packets for the core switch in the DCN setup.

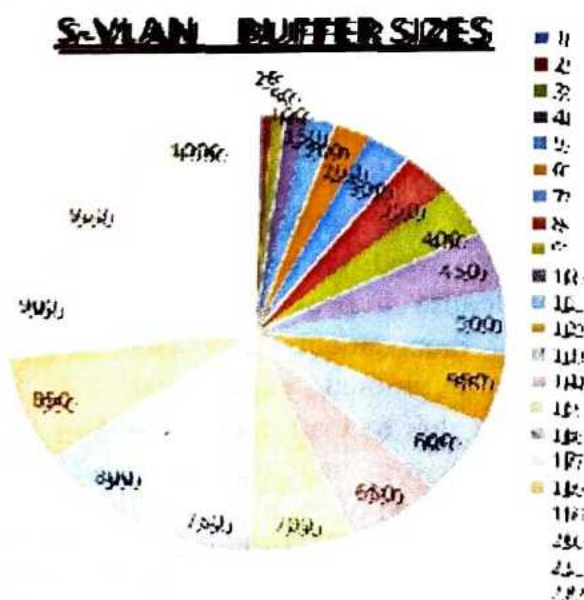


Figure 4: S-VLAN Buffer Size Distribution.

Figure 5 shows the S-VLAN priority tag distribution used in this research. For the different S-VLANs, priority tags are assigned so as to logically identify and reduce queues of the end systems. The core S-VLAN switch assigns resources based on the end system priorities. Terminals with high priority traffic are processed before others with lesser priority. Notice that since all S-VLANs were assigned priority tags, this scheme will guarantee high security, traffic control and bandwidth optimization at large thereby enhancing congestion management in switched DCNs.

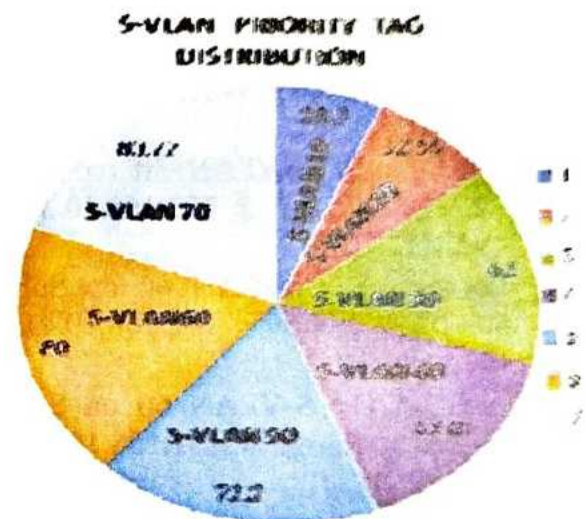


Figure 5: S-VLAN Priority Tag Distribution (PTD)

Figure 6 shows S-VLAN resource utilization in the DCN design. In the S-VLAN DCN environment, latency, throughput, utilization, bandwidth mismatch and usage are vital resources considered in congestion management. With the traffic load sources in the seven star topology used in this research, an initial gradient was established before resource allocations were fairly distributed. With a connection request, feasible regions of resource allocation are first established. From the test bed, enterprise servers are dumped in S-VLAN 50, 60 and S-VLAN 70. This corresponds with the load sources 5, 6 and 7 from the plot. Resource utilization in these regions is

quite high as shown by the peaks from 120 to 140 load times. These are the regions with high priority tags and as such the computational power of the DCN servers are high. Thus, in DCNs, the regions of high priority tags have the highest resource utilization cycles. Also, at zero load time from the plot, the connection requests have been established in the S-VLAN model. This literally depicts a high performance network since low latency networks have negative or zero initial load times.

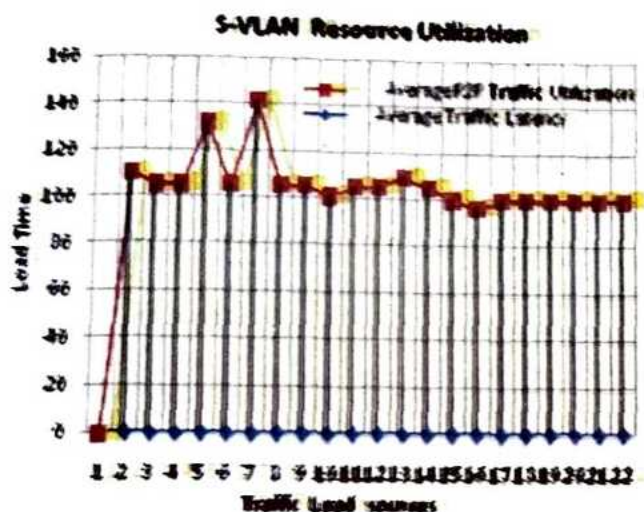


Figure 6: S-VLAN Resource Utilization.

Figure 7 shows the percentage S-VLANS throughput responses against the Buffer sizes. In this context, after each simulation run, the throughput was processed and expressed in percentage distribution. From the plots, efficient throughput for the various S-VLANS has a considerable transients and oscillations before fairness and relative stability is maintained. Defining fairness and stability as maximum queue state of 450 packets, it is observed from the trace file (generated script) that the system (DCN) reaches the stable state within a 0.45s. This is acceptable for scalable data centre network. Initially, the throughput increases with the offered buffer sizes and reaches a peak value of 25 packets. After the maximum peak value is reached, a further increase in buffer sizes only leads to a gradual stable and efficient throughput. Therefore, in order to maximize the

throughput in a congested network, the buffer sizes must be carefully chosen to optimize performance.

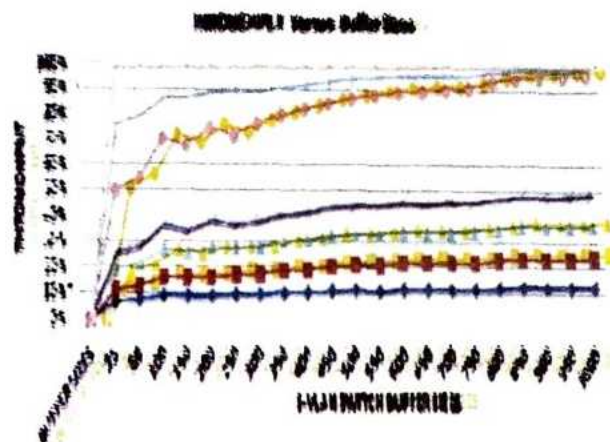


Figure 7: Percentage S-VLAN Throughput Responses against the Buffer Sizes

7.0 Conclusions

In this research, an analysis and simulation of Synthesis Virtual Logical Aggregation Network Segmentation is presented as an approach to congestion management. It was observed that S-VLANS improves throughput and latency while maintaining stability and fairness in a congested network link thereby enhancing performance in DCNs. Unlike BCN that have feedback control with high and unpredictable overhead, S-VLANS does not require any feedback loop to regulate congestion effects, but logically identifies sources based on priority tags. S-VLANS logically sets up a service policy that guarantees no loss for the FIFO Queues regardless of the source transmission rates. Consequently, it is obvious that by introducing the S-VLANS priority tagging in core switches in Data centre domains, a function demarcation between the different services is possible, and the possibility of minimizing the congestion problem is increased with little overhead.

References

- [1] D. Bergamasco, R. Pan, "Backward Congestion Notification Version

- 2.0", IEEE 802.1 Interim Meeting, Garden Grove (CA), Sep 2005. [Online]. Available: <http://www.ieee802.org/1/files/public/docs2005/newbergamasco-bcn-september-interim-rev-final-0905.ppt>
- [2] S. Floyd. Papers on congestion control research. [Online]. Available: <http://www.icir.org/floyd/papers.htm>
- [3] G.Mcalpine.M. Wadeker et., " An Architecture for congestion management in Ethernet clusters, "IEEE International parallel and Distributed processing Symposium, 2005.
- [4] [Http://www.cse.wustl.edu/jain/ieee.htm](http://www.cse.wustl.edu/jain/ieee.htm)
- [5] D.Bergamasco, "Data Center Ethernet congestion Management: Backward Congestion Notification," IEEE 802.1 meeting, May 2005.
- [6] Jinjing Jiang, Raj Jain, and Manoj K.Wadekar, "Analysis of Backward Congestion Notification (BCN) for Ethernet in Datacenter Applications," IEEE INFOCOM 2007, pp 2456-2460, May 2007.
- [7] Jinjing Jiang, Raj Jain, and Chakchai So-In, "Congestion Management for Ethernet in Datacenter Application Using Forward Explicit Rate Notification," WUSTL technical report, 2007.
- [8] Davide Bergamasco and Rong Pan, "Backward Congestion Notification Version 2.0," IEEE 802.1 Meeting, September 2005.
- [9] M.Seaman, "Congestion Notification," Notes for IEEE 802.1 Congestion management Interim Meeting, January 2006.
- [10] "SimEvents," URL: <http://www.mathworks.com> Cisco Systems,
- [11] "Packet tracer". [Online]. Available, <http://www.cisco.com>
- [12] Ethernet Technologies, Cisco Systems, Inc. [Online]. Available: <http://www.cisco.com/univercd/cc/td/doc/ethernet.pdf>
- [13] D. M. Nicol, S. W. Smith, M. Zhao, "Evaluation of Efficient Security for BGP Route Announcements using Parallel Simulation", Simulation Modelling Practice and Theory, Vol. 12 (2004) 187-216.