Lossless Traffic Over Long Distance Links in DCB Networks

Anoop Ghanwani, Dell

Joseph White, Dell

Overview

- Motivation
- Lossless applications
- Cable lengths in the data center
- Cable length and buffering
- Possible solutions
- Summary

Motivation

- PFC was developed for supporting lossless service in DCB networks
- PFC requires the provisioning of buffers for each lossless class
- Among other parameters, the amount of buffer that must be provision depends on the cable length and interface speed between the two devices
- Some products may not have adequate buffering for the number of lossless classes required
- Is there something we can do about this?

Lossless Applications

- iSCSI
  - Used for block storage
  - Lossless transport not required, but often recommended
- FCoE (Fibre Channel over Ethernet)
  - Storage protocol
  - Requires lossless transport
- RoCE (RDMA over Converged Ethernet)
  - Requires lossless transport
  - Gaining popularity because of applications such as SMB Direct
- Could have more than one of these, or multiple classes of these in any deployment

Cable Lengths in the Data Center

Location	Cable Length
Server to ToR	<= 3 m
ToR to Leaf	<= 20 m
Leaf to Spine	<= 500 m
Spine to Central Colocation	<= 1000 m
Between Central Colocation in the Metro	<= 10 – 80 km

See booth_400_01a_1113.pdf

Cable Lengths and Buffering

- Consider the following example
  - Link Speed = 40 Gbps
  - Speed of light in optical fiber ~= 2x10^8 m/s
  - MTU = 2000 bytes (802.3as) [Ignoring preamble and IFG]
- Buffering required per lossless class per port

Cable Length	# Bytes in 1 RTT	# MTU in 1 RTT
50 m	~2.44 KB	~1.25
500 m	~24.4 KB	~12.5
1000 m	~48.8 KB	~25
10 km	~488 KB	~250

Possible Solutions

- Credit-based flow control
  - Always lossless
  - Discussed in new-ghanwani-llfc-01-14-v01.pdf
  - Not enough consensus due to complexity with buffer sharing across ports and priorities
- Use PFC with enhancements
  - Requires knowledge of RTT at the sender and precise shaping
  - Lower utilization may be acceptable since bandwidth can still be allocated to lossless classes
  - Shaping is discussed in new-ghanwani-enhanced-sched-dcbx-0714-v01.pdf
    - Restrict number of bytes transmitted in an RTT
    - Minimum BW guarantees are not needed for this problem

Summary

- Use cases for support of lossless traffic over long distance links are emerging
- Using PFC as is would require provisioning large buffers for each traffic class
- In bridges will smaller buffers, it may be possible to provide a solution by enhancing PFC whereby the amount of traffic is restricted

THANK YOU