The Case for An Alternative Link-Level Flow Control Mechanism at High Speeds

Anoop Ghanwani, Dell
Pat Thaler, Broadcom
Jeff Lynch, IBM
Mitch Gusat, IBM
Mark Gravel, HP

Acknowledgement

- For their review
  - Robert Winter
  - Joseph White
  - Shivakumar Sundaram

Overview

- Motivation
- Buffer requirements for priority-based flow control (PFC)
- Buffer requirements and the bandwidth delay product
- Other drawbacks of PFC
- Can credits offer a better solution?

Motivation

- DCB networks offer lossless operation for certain traffic types
- Several protocols either require, or benefit from, a lossless network
  - FCoE, RoCE, iSCSI
- PFC has inherent inefficiencies in buffer utilization
- Can we do better?

Buffer Requirements for PFC

- PFC requires a certain amount of buffer space to be reserved for each lossless class
- The computation of this space is quite involved
  - See Annex N of 802.1Q-2012
  - Depends on cable length, PHY type, use of MACsec, etc.
- The reserved buffers are never used until *after* a PFC event
  - They remain unused during normal operation, e.g. absorbing bursts
- This results inefficient buffer utilization

Buffer Requirements for PFC (2)

- Higher layer/MAC/PHY delays
  - Typically negligible, amounting to << 1 MTU
  - Use of MACsec increases this significantly and is proportional to link speed
- Boundary conditions
  - Delay in putting a PFC frame on the wire because the link is busy
  - Delay in effecting PFC once a message is received because a transmission has already begun
  - This is 2 MTU regardless of link speed or cable length
- RTT
  - For a given cable length, the number of bytes required depends on link speed

Buffer Requirements and the Bandwidth Delay Product

- Consider the following example
  - Cable length = 50 m
  - Speed of light in optical fiber ~= 2x10^8 m/s
  - RTT = (2 x 50) / (2x10^8) = 0.5 usec
  - MTU = 2000 bytes (802.3as) [Ignoring preamble and IFG]

Link Speed	# Bytes in 1 RTT	# MTU in 1 RTT
10 Gbps	~0.61 KB	< 1
40 Gbps	~2.44 KB	~1.25
100 Gbps	~6.1 KB	~3.12
400 Gbps	~24.4 KB	~12.5
1 Tbps	~61.0 KB	~31.2

Buffer Requirements and the Bandwidth Delay Product (2)

- For a given cable length, as link speeds are increased, the amount of buffer that must be set aside for PFC goes up
  - Again, this is not used during normal operation
- The boundary conditions are relatively fixed
  - ~3 MTU regardless of link speed (ignoring MACsec)
- With increasing link speed, RTT is the dominant contributor

PFC Buffer Requirement for Some PHYs

- As mentioned earlier, in addition to the RTT, the buffer required depends on the type of PHY, use of FEC, MACsec, etc.
- Consider the following example
  - Cable length = 50 m
  - Speed of light in optical fiber ~= 2x10^8 m/s
  - RTT = (2 x 50) / (2x10^8) = 0.5 usec
- More calculations on next slide

PHY	2x Max Frame + PFC frame + RTT + 2 x ID + HD	+ FEC	+ MACsec
10GBASE-T	14.64 KB	NA	10 KB
10GBASE-SR	8.51 KB	NA	10 KB
40GBASE-SR4	17.47 KB	NA	10 KB
100GBASE-SR4	35 KB	NA	10 KB

Calculations for PFC for PHYs

DV (in Bit Times)

= 2 x max frame size + PFC frame size + 2 x Cable Delay + 2 x ID + HD

= 2 x 16160 + 672 + 500 nsec x Link Speed + 2 x ID + 614.4 ns x Link Speed

= 32992 + 1114.4 x Link Speed in Gbps + 2 x ID

ID (Bit Times)

- 10GBASE-T = 10GBASE-T Delay (25600) + XGMII MAC/RS and XAUI (12288) = 37888
- 10GBASE-SR = RS (8192) + 10GBASE-R PCS (3584) + 10GBASE-R PMA (512) + 10GBASE-SR PMD (512) = 12800
- 40GBASE-SR4 = RS (16384) + 40GBASE-R PCS (11264) + 40GBASE-R PMA (4096) + 40GBASE-SR4 PMD (1024) = 32768
- 100GBASE-SR4 = RS (24576) + 100GBASE-R PCS (35328) + 100GBASE-R PMA (9216) + 100GBASE-SR4 PMD (2048) = 71168

MACsec = (2000+20) + 4 x (64+20) bytes, times 4 because it adds to transmit and receive at each end affecting RTT, not affected by link speed

Other Drawbacks of PFC

- PFC is reactive
- Incorrect calculations for buffer space are expensive
  - On the conservative side, they lead to added waste
  - On the aggressive side, they lead to loss
- For a given buffer size and link speed, at some cable length, it becomes impossible to provide lossless operation

The Case for Credit-based Flow Control

- With credits, there are no buffers set aside for the congestion event
- Instead, the credits must be sized to cover the RTT
  - If sufficient buffers are not available, the link will be underutilized
- Lossless operation is always guaranteed

Credits Are Not Perfect

- Credit size
  - If too big, could lead to underutilization by fragmentation
- Lost credits
  - If credits are not reliably returned, leads to lower available credits and consequently lower utilization
- Solutions exist for these problems
  - We can debate and refine them during the process of standardization

Summary

- This presentation discusses some of the drawbacks of PFC as a link level flow control mechanism for high speed links
- It may be worthwhile to look at credit-based flow control
- If developed, there should be a mechanism for negotiating the behavior so as to be backwards compatible

THANK YOU