BROCADE

Enabling Block Storage over Ethernet:
The Case for Per Priority Pause

A joint proposal from:
Pat Thaler: Broadcom
Joe Pelissier: Brocade
Claudio DeSanti: Cisco
Asif Hazarika: Fujitsu
Mike Ko, Mitch Gusat: IBM
Manoj Wadekar, Shelto Vandoorn: Intel
Diego Crupnicoff: Mellanox
Craig Carlson: QLogic
Guenter Roeck: Teak Technologies

Block Storage = Huge Opportunity for Ethernet

Enables a unified network connection for servers

End user benefits – consolidation of multiple networks

- Storage, IPC, traditional LAN / WAN

Helps to fill the 10G Ethernet pipe

- Drives 10G technology adoption

Improving iSCSI Performance

iSCSI is valuable for lossy networks (e.g. LANs Internet, etc.)

- Because its built on TCP
  - It is reliable even when run over unreliable LAN/WAN protocols
  - Contains congestion management protocols

However, frame loss impacts iSCSI performance:

- Increases processing requirements for scatter / gather
- Frame retransmission slows transaction completion time
- TCP back off algorithms can reduce utilization of available bandwidth

Eliminating frame loss significantly improves the performance of iSCSI

Introducing FCoE

FCoE= FC over Ethernet

- FC is a natural fit for block storage in lossless Ethernet environments
- Simply encapsulate FC frames in Ethernet frames
- Evolutionary – not revolutionary

Preserve investments and skill

- Industry invested heavily in developing and deploying FC

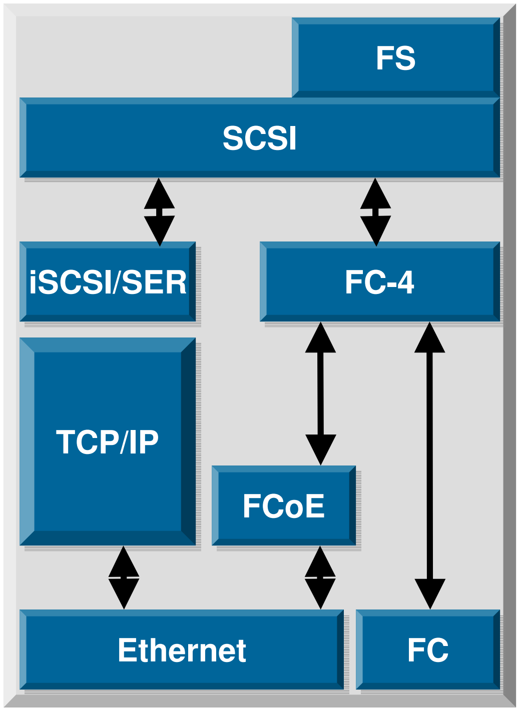
Reduced Risk

- FC has large installed base proven to work for storage

Shorter Time to Market

- Less complexity, fewer unknowns

Standardization underway in ANSI/t11



FCoE (like Fibre Channel) requires a lossless fabric

FCoE does not provide for recovery of lost frames

- Done by higher level protocol (e.g. FCP, class driver application)
- Recovery in certain cases requires 100’s to 1000’s of ms

Excessive loss may result in link resets, redundant fabric failovers, and severe application disruption

FCoE therefore requires PAUSE to be enabled

- With 802.1X PAUSE, this implies a separate fabric for FCoE
  - Since traditional LAN/WAN traffic is best served without PAUSE
- Which significantly reduces the value proposition of FCoE
- Which reduces the value proposition of higher speed Ethernet

iSCSI also enjoys significant benefits of a lossless fabric

Congestion Management and Frame Loss

Tremendous effort has been expended in developing congestion management schemes for Ethernet

- Simulation efforts indicate that these schemes are likely to dramatically reduce frame loss
- However, frame loss not sufficiently eliminated
  - Especially under transitory bursts and in topologies that one would reasonably expect for storage traffic
  - Congestion Management does reduce the congestion spreading side effect of flow control
- Therefore a supplemental flow control mechanism that prevents frame loss is viewed as a requirement for successful deployment of block storage over Ethernet
  - A simple method is sufficient.

Flow Control Requirement Overview

Work in conjunction with or without Congestion Management

Operates within confined region of the network (e.g. a congestion managed region of a data center)

- Necessary to prevent congestion spreading

Allow other types of traffic that are not well served by flow control to continue to operate in the traditional fashion (i.e. packet drop)

Does not attempt to solve larger scale flow management problems that are best solved by other means

- E.g. metro area

Keep it simple!

Flow Control Proposal Overview

Define a “per priority” flow control scheme

- Each priority may be independently flow controlled
- May be enabled / disabled for each priority
  - Disabled priorities discard frames during congestion

Base on existing pause mechanism

- A credit mechanism might provide certain technical advantages
- But our main objective is to keep it simple
- Pause is sufficient for the problem we are attempting to solve

See, for example, new-cm-barrass-pause-proposal.pdf

Some Thoughts on Deadlock

Flow control in conjunction with MSTP or SPB *in theory* could cause deadlock

- See “Requirements Discussion of Link Level-Flow Control for Next Generation Ethernet” by Gusat et al, January ’07 (au-ZRL-Ethernet-LL-FC-requirements-r03)

To create deadlock, the following conditions must occur:

- A cyclic flow control dependency must exist
- Traffic must flow across all corners of the cycle
- Sufficient congestion must occur on *all* links in the cycle *simultaneously* such that each switch is unable to permit more frames to flow

At this point, all traffic on the affected links halt until frames age out

- Generally after one second

Feeder links also experience severe congestion and probable halt to traffic flow

Some More Thoughts on Deadlock

However, the probability of deadlock becomes negligibly small:

- Data Center topologies (e.g. Core / Edge) typically do not contain the necessary cyclic flow control dependencies
- Even if such dependencies existed, the MSTP setup required to cause deadlock would be unusual
  - Edges would rarely be defined as the root of a MSTP
- Even if the topology and the MSTP setup was sufficient, it would be unlikely that the traffic flow would be such that deadlock would occur
  - Storage traffic generally travels between storage arrays and servers
  - Insufficient traffic passes through certain “corners” of the loop to create deadlock
- Even if the traffic flow was sufficient around all “corners” to create deadlock
  - Congestion Management would make sufficient congestion at all necessary points in the network highly unlikely

Even More Thoughts on Deadlock

The previous assumptions are not without demonstration

- Fibre Channel is widely deployed in mission critical environments:
  - Fibre channel is flow controlled
  - Topologies with loops are commonly deployed
  - FSPF, Fibre Channel's routing protocol, naturally creates the traffic routing necessary (but not sufficient) for deadlock
  - No Congestion Management

A deadlock in Fibre Channel, if it were to ever occur, would be highly noticeable

- Link resets, re-routing, fabric failover

Even under these circumstances, deadlock has been a complete non-issue in the installed base of Fibre Channel

- What is being proposed for Ethernet has additional "deadlock hardness"

Some Final Thoughts on Deadlock

In Summary:

- Deadlock has never been an issue in the huge installed base of mission critical Fibre Channel deployments
- In all respects, what is being proposed for Ethernet is less likely to result in deadlock than Fibre Channel

Therefore, the current proposal is sufficient for deadlock prevention in Ethernet

BROCADE

Questions?

BROCADE

™

Thank You!