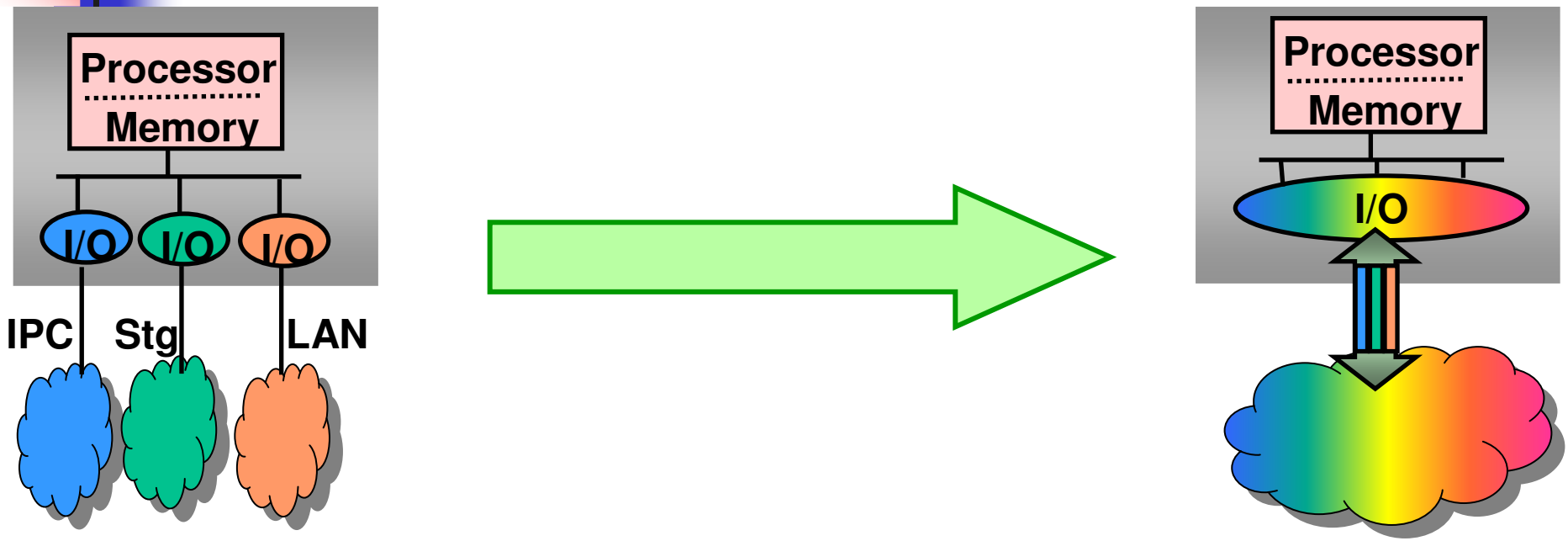
Ethernet Enhancements for Storage in a Datacenter

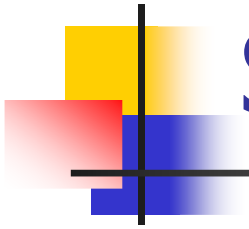
Mike Ko, IBM
Renato Recio, IBM
Manoj Wadekar, Intel
Joe Pelissier, Brocade
Davide Bergamasco, Cisco

July 16, 2007

I/O Consolidation in the Datacenter



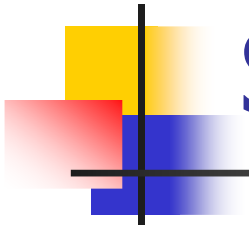
- Enhancing Ethernet to enable I/O consolidation in the datacenter has been discussed in 802 meetings since 2004
- Proposals on congestion management are currently being debated in 802.1Qau working group



Storage and I/O Consolidation

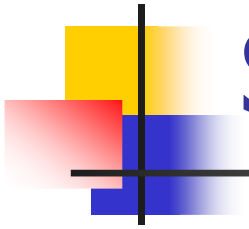
- Fibre Channel is still the dominant storage technology for the enterprise market
- Can Ethernet hardware deliver an enterprise storage solution?
  - New storage protocol currently being considered for standardization at T11
  - Layers Fibre Channel frames directly over Ethernet
  - Provides a lighter weight implementation by eliminating TCP/IP
  - Known as Fibre Channel over Ethernet (FCoE)
  - Leverages existing FC management infrastructure
- But FCoE alone is insufficient for I/O consolidation
  - Uses PAUSE mechanism to prevent frame loss
    - Causes head-of-line blocking problems for other traffic
  - Ethernet enhancements will be needed in order for storage to share the link with other classes of applications such as IPC and LAN

Support in Ethernet for Storage in the Datacenter



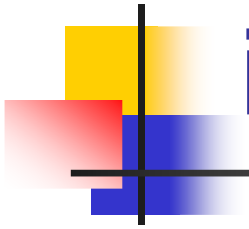
- Ethernet needs to be enhanced in the following areas:
  - Enhanced transmission selection
  - Priority-based flow control
  - Discovery and capability exchange protocol
- These Ethernet enhancements:
  - Provide the support needed by enterprise storage solutions
  - Enable storage, IPC, and LAN traffic to share the same I/O fabric
  - Critical for future enterprise storage solutions such as FCoE

Enhanced Transmission Selection



- Provides priority processing and packet scheduling
  - Queuing requirements for different traffic classes are needed to allow for different resource allocation
    - To enable each class of applications to use the same consolidated layer 2 transport
- Different traffic classes need to be managed separately
  - LAN
    - Large number of flows, not very sensitive to latency
    - E.g. dominant traffic type in Front End Servers
  - SAN
    - Large packet sizes, sensitive to packet drops
    - E.g. Middle Tier and Back End Servers
  - IPC:
    - Mix of large and small messages
      - Small messages are latency sensitive
    - E.g. Back End Servers, HPC Applications

Use of Queuing Requirements in Storage



- Priority groups allow storage traffic to be managed as a group with configurable QOS guarantees
  - Ensures that storage traffic will get its fair share of resources
  - Allows the scheduling mechanism to apply different disciplines
    - Provide minimal latency for delay sensitive traffic in other bandwidth groups
- If necessary, different queues can be set up within the storage traffic class group with different QOS allocation

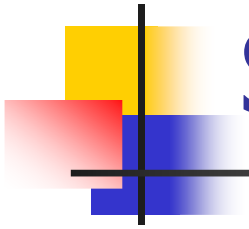
PAR for Priority Processing and Packet Scheduling

- Consensus in the 802.1Qau working group on congestion management to support the following position:
  - "The CM task group should draft a PAR, 5 criteria and objectives for transmission selection for 802.1Q bridges and end nodes to provide priority grouping and per-group traffic class allocation, for review by IEEE 802.1 at the July plenary"
  - Straw poll was taken in the interim meeting in May '07
- Draft of proposed PAR now in document area:
  - "new-cn-thaler-trans-select-par-070716"

Priority-based Flow Control and Storage

- No packet drop behavior is required by storage protocol such as FCoE
  - Priority-based flow control will be needed
  - E.g. per priority PAUSE
- Per Priority PAUSE extends the granularity of 802.3x PAUSE mechanism to accommodate different priority classes
  - Selective pausing avoids impacts to high priority and delay sensitive traffic
  - For storage protocols layered over TCP/IP, priority-based flow control enables service differentiation at the link layer (vs at the IP layer)
- Current proposals on congestion notification in 802.1Qau can reduce frame loss
  - But frame loss is still possible under transient conditions
  - Priority-based flow control is necessary to prevent frame drops

Impact of Dropped Packets in Storage



- For storage traffic that uses TCP/IP as the transport such as iSCSI and iSER
  - Besides retransmission delay, TCP/IP also exhibits additive-increase-multiplicative-decrease (AIMD) behavior in response to packet drops
    - Hurts throughput and latency
- For storage traffic that does not use a transport layer such as FCoE
  - Detection at the SCSI level is in the order of 10s of seconds
    - Detection time is in the order of seconds if Read Exchange Concise (REC) extended link service is supported
  - Recovery is at the SCSI command level
    - Severely hurts throughput and latency
    - May cause severe system malfunction (e.g., unexpected server reboots)

Priority-based Flow Control Considerations

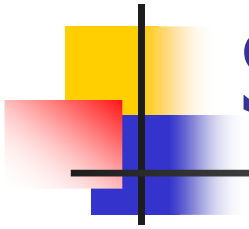
- Concern about Priority-based Flow Control causing deadlocks
  - But deadlock is rarely an issue in Fibre Channel in a datacenter environment
  - Ongoing discussion on potential deadlock issues
    - "au-ZRL-Ethernet-LL-FC-requirements-r03"
    - "new-cm-pelissier-enabling-block-storage-0705-v01"
  - Will continue to explore refinements to alleviate any potential deadlock problems
- Concern about Priority-based Flow Control concept being extended beyond the datacenter
  - Can limit the scope of Priority-based Flow Control to datacenter deployment only
  - Other alternatives can be explored as well

PAR for Priority-based Flow Control

- Consensus in the 802.1Qau working group on congestion management to support the following position:
  - "The CM task group should draft a PAR, 5 criteria and objectives for granular (priority-based) link level flow control for 802.1Q bridges for review by IEEE 802.1 at the July plenary"
  - Straw poll was taken in the interim meeting in May '07
- Draft of proposed PAR to be uploaded in document area soon
- Proposal on Per Priority PAUSE now in document area
  - "new-cm-barrass-pause-proposal"

Discovery and Capability Exchange Protocol

- For the enhanced Ethernet, a mechanism is needed to discover the boundary of the enhanced Ethernet components and exchange capabilities
  - Support for priority classes (such as bandwidth allocation)
  - Support for congestion management (optional)
  - Support for priority-based flow control
  - Etc.
- Current plan is to participate in 802.1AB-REV project to incorporate Discovery and Capability Exchange Protocol for Ethernet enhancement
  - Can the 802.1AB-REV schedule accommodate additional input?
  - If not, should a new PAR be submitted?

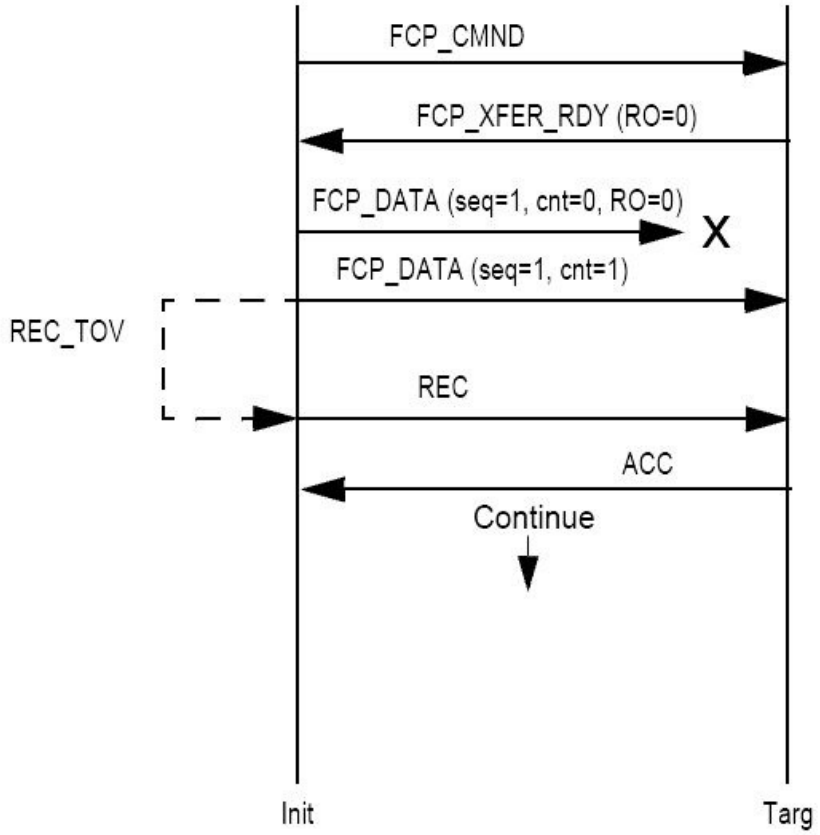


Summary

- Work on congestion management in 802.1Qau is a good first step
  - But not enough for Ethernet to become the converged fabric in the datacenter
- We intend to request the IEEE 802.1 community to approve the request to circulate the PAR, 5 criteria, and objectives for the following areas in this plenary meeting:
  - Enhanced transmission selection
  - Priority-based flow control
- We intend to participate in the 802.1AB-REV project to incorporate Discovery and Capability Exchange Protocol for Ethernet enhancement

Backup

FCP Error Detection with REC for Lost Write Data in Class 3 Service



FCP Error Detection with REC for Lost Read Data in Class 3 Service