Congestion Management Capabilities of Various Fabrics

Tanmay Gupta, Manoj Wadekar

Intel Corporation

Jeff Wise

Motorola ECC

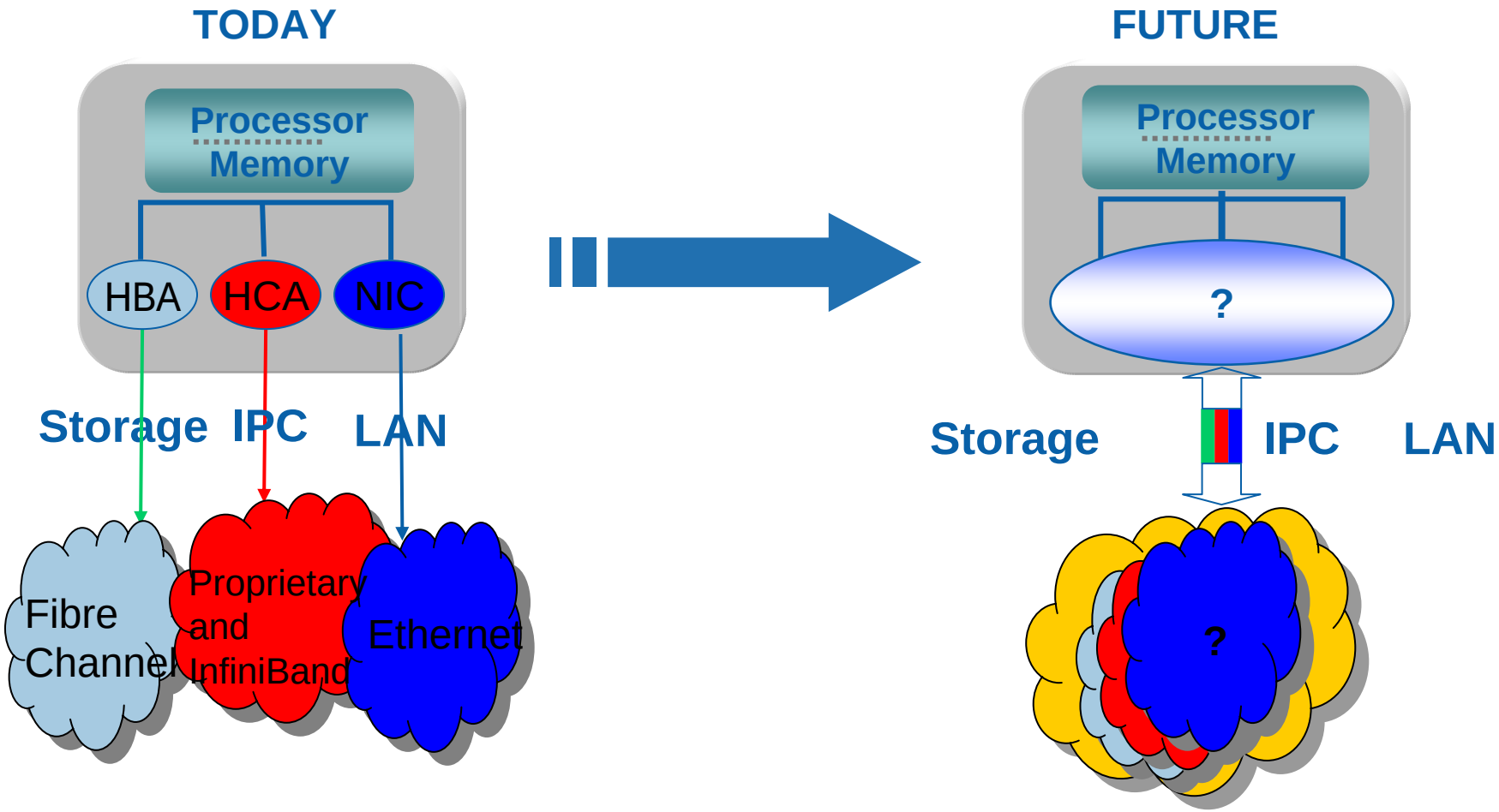
IEEE 802.1 Congestion Management Interim

January 2006

Some Fabric Interconnects in Today's Data Centers

- Ethernet
  - Default fabric for networking traffic
- InfiniBand
  - Emerging fabric for inter-process communication (IPC) traffic
- Fiber Channel
  - Preferred fabric for storage traffic
- Advanced Switching Interconnect (ASI)
  - Aimed at inter-board, backplane, and multi-rack systems typically used in Telecom environments
- Proprietary (Myrinet, Quadrics, ...)
  - Preferred fabric for low latency traffic

Consolidation



Trend towards consolidation due to higher cost and complexity of multiple fabrics – Can Ethernet be the consolidated fabric?

InfiniBand

Congestion management capabilities

InfiniBand

Congestion management capabilities

- Link level flow control
- End-to-End flow control
- Static Rate Control
- End-to-End injection rate control
- VL arbitration

InfiniBand

Link level flow control

- Granular link level credit based flow control
  - Per Virtual Link (VL) flow control (upto 15 data VLs)
    - Per VL flow control eliminates HOL blocking between VLs
    - Latency sensitive traffic using one set of VLs can be unaffected by congestion of best effort traffic in other VLs
- Effectively deals with transient congestion without dropping packets
  - Feedback-based mechanisms cannot deal with time constants of transient congestion

InfiniBand

End-to-End flow control

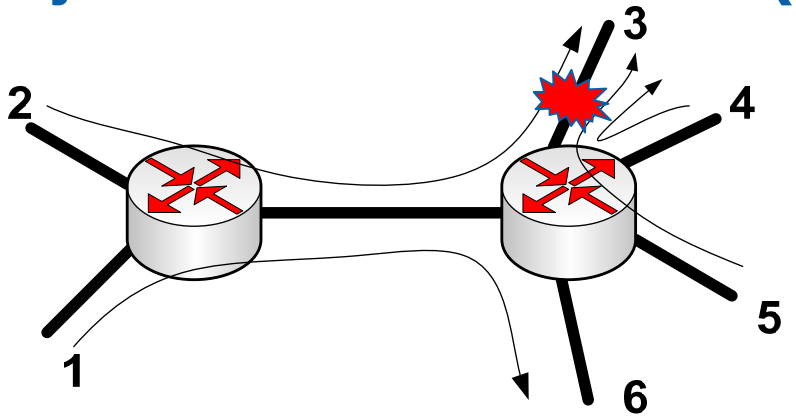
- End-to-End (message level) credit based flow control
- Used by reliable connections to prevent receive queue overflow
- End-to-End credits are generated by a responder's receive queue and consumed by a requester's send queue
- Encoded credits are transported from the responder to the requester in an ACK message

InfiniBand
Static Rate Control

- Enables Fabric Manager to configure the network to avoid oversubscribed links in the fabric
- Effectively handle link speed mismatches
  - A path with 4X link feeding a 1X link can be constrained to a 1X rate

InfiniBand

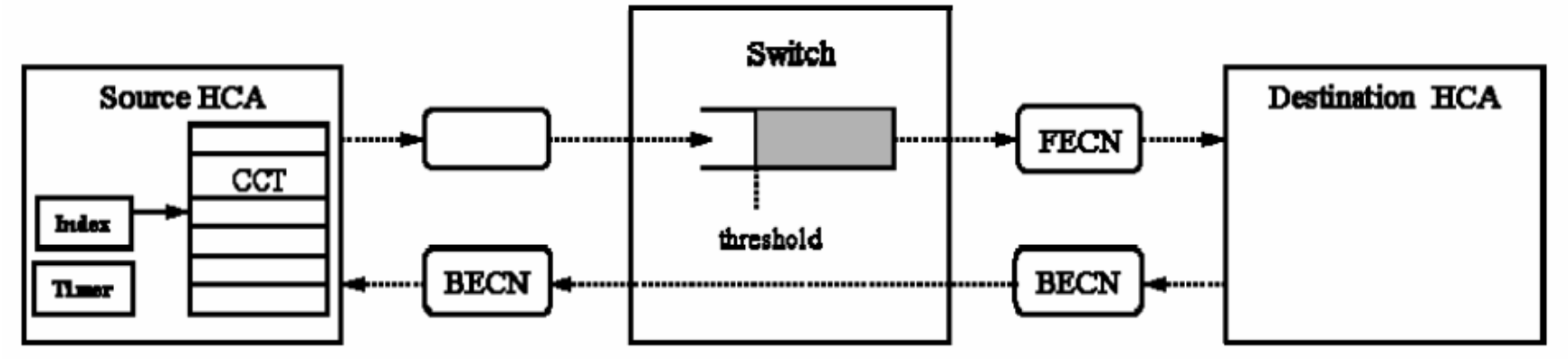
End-to-End injection rate control (1)



- Added in 2004 (Release 1.2) to address congestion spreading problem
  - Oversubscription to destination 3, blocks traffic on the inter-switch link
  - This blocks flow to destination 6 which is uncongested
- Uses Forward Explicit Congestion Notification (FECN) and Backward Explicit Congestion Notification (BECN)
- Pushes congestion to the source to avoid head-of-line blocking and congestion spreading

InfiniBand

End-to-End injection rate control (2)



- Switches detect congestion based on threshold and mark packets with FECN bit
- Destination responds to source queue pair with BECN, either piggybacked on ACK or special Congestion Notification packet
- Source reduces the injection rate of flow by increasing index into Congestion Control Table (CCT) which stores inter-packet delay values
- Source recovers rate based on timer by decreasing index into CCT on timer expiry

InfiniBand
VL arbitration

- Dual priority WRR arbitration scheme between VLs
- Allows configurable link sharing between VLs

Fibre Channel

Congestion management capabilities

Fibre Channel

Flow Control

- Buffer-to-Buffer credit based flow control
  - Destination port signals by sending a Receiver_Ready primitive signal to the transmitting port when it has free receive buffers
- End-to-End credit based flow control
  - Happens between initiator (source) and responder (destination) ports
  - ACK frames used to replenish credits

Fibre Channel

Class of Service

- Classes define communication strategy to use depending on the type of data to be transmitted
  - Determines the types of flow control used
- Class 1: Dedicated connection (reserves entire link on the path)
  - Only End-to-End flow control
- Class 2: Connectionless communication with end-to-end acknowledgements
  - Both buffer-to-buffer and end-to-end flow control
- Class 3: Connectionless with no end-to-end acknowledgements
  - Only buffer-to-buffer flow control
- Class 4: Similar to Class 1 except that only part of link bandwidth reserved for a VC
  - Only End-to-End flow control
- Class 5: Undefined
- Class 6: Multicast service
  - Only End-to-End flow control
- Intermix class: Allows Class 2 or Class 3 frames to be transmitted at times Class 1 frames are not being transmitted

Advanced Switching Interconnect (ASI)

Congestion management capabilities

ASI

Congestion management capabilities

- Link level flow control
- Endpoint source injection rate limiting
- Link bandwidth arbitration
- Reliable data transport
- Centrally managed fabric

ASI
Link level flow control

- Multiple “virtual channels” per link
  - Up to 16 unicast VCs: 8 are bypass-capable
  - Up to 4 multicast VCs
  - Separate flow control for each VC.
- Credit-based link-level flow control
  - Link VC receiver gives the VC transmitter a limit on how much data it can forward based on the currently free buffering in Rx.
    - The limit is updated as Rx buffering empties.
  - Credit update messages conveyed in a special, 4-byte packet.
- Optional Status-based flow control
  - Switches send output port status messages to their immediate upstream neighbors.
    - Provides one stage look-ahead view on congestion situation.
  - Upstream nodes temporarily stops data to congested downstream output ports.

ASI
Endpoint source injection rate limiting

- Optional
- Match bandwidth of
  - Slowest link along the path
  - Data rate of traffic's destination
- Source node maintains a connection queue (CQ) for each flow.
  - Per traffic class
  - Per path through the fabric
- Token bucket controls the flow of traffic through a CQ.
  - Shapes the transmission of the packets in time.
  - Allows limited burstiness
- Higher layer protocols determine the token bucket parameter values.
  - Protocols not specified in ASI spec., as far as I can see.

ASI
Link bandwidth arbitration

- Output port uses a "Minimum BW" scheduler to allocate link bandwidth to the active VCs.
  - An active VC has both waiting traffic and sufficient credits.
- Allocates available BW based on per-VC weighting parameter.
  - Allocates link BW on a data basis, not on a packet basis.
- Work conserving
  - If there is at least one packet ready to send, it will be sent.
  - Traffic is not held back for rate limiting/shaping at the switches.
- Specific algorithm not specified by ASI.

ASI
Reliable data transport

- Retransmission of lost or corrupted packets.
  - Minimum delay to detect the problem.
- Avoids the need to resend a lost packet end-to-end.
  - Reduces network BW needed to recover a lost packet.
  - Congestive collapse is almost impossible (?).

ASI
Centrally managed fabric

- Agents responsible for various aspects of fabric operation.
- Fabric-wide optimization and management
- Some fabrics have multiple paths between end nodes.
  - Important redundancy for failovers
  - For more BW or QoS than a single path can provide.
- Source routing of each packet allows traffic source to pick the route.
  - Potentially different for each traffic class * end-point.
  - Fabric manager typically discovers and defines the routes.
- Fabric management software may regulate access to the ASI fabric.
  - No new flows except when sufficient resources are available.

Ethernet

Congestion management capabilities

Ethernet
Priority and Flow Control

- Traffic Classes
  - Support for up to 8 priorities for traffic differentiation
- 802.3X flow control with link level granularity
  - Causes congestion spreading
  - Can not differentiate between latency sensitive and best effort traffic

Can Ethernet be the consolidated fabric?

- For Ethernet to be the consolidated fabric, it needs to have better congestion control capabilities including
  - End-to-End congestion control to push congestion to the sources
  - More granular link level flow control
  - Standard enhanced scheduling mechanisms

Backup