Ethernet for Data Center: Reliable, Channelized and Robust

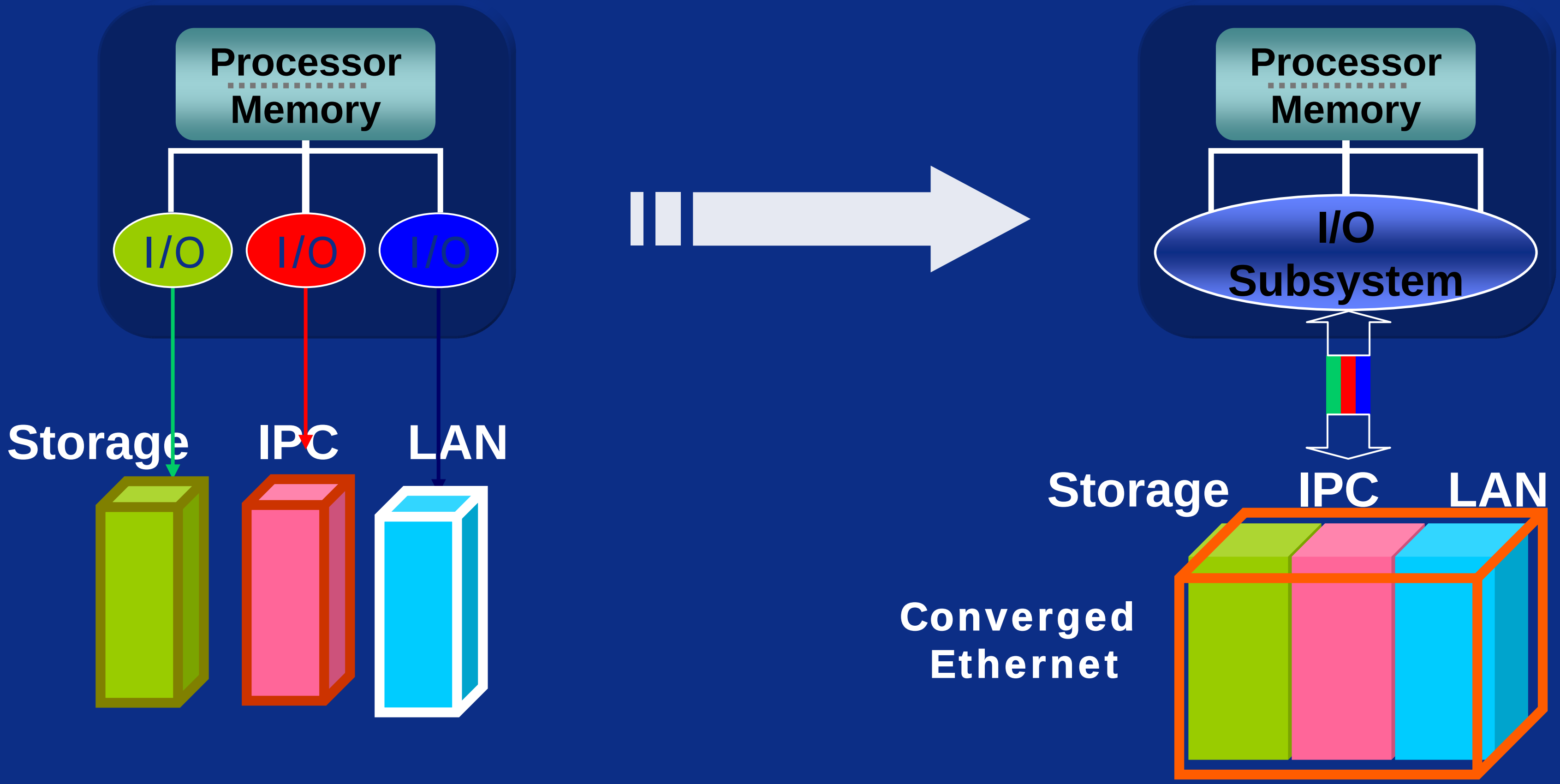
June 11, 2007

Manoj Wadekar

Agenda

- Why?
- What?
- How?
- What next?

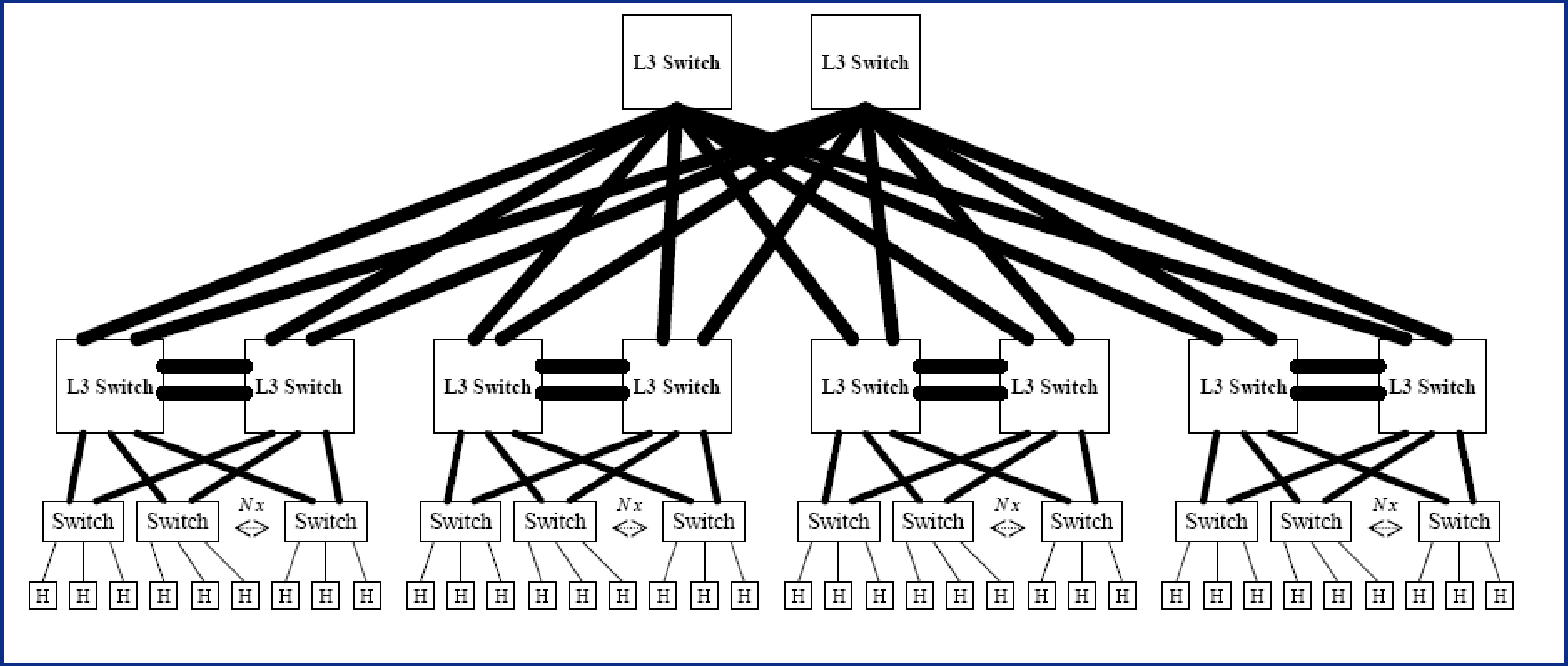
Need for IO Consolidation



Hardware and management complexity is growing

- Many fabrics (LAN, SAN, IPC, Management..)
- Cables, power, provisioning..

Increasing demand for bandwidth



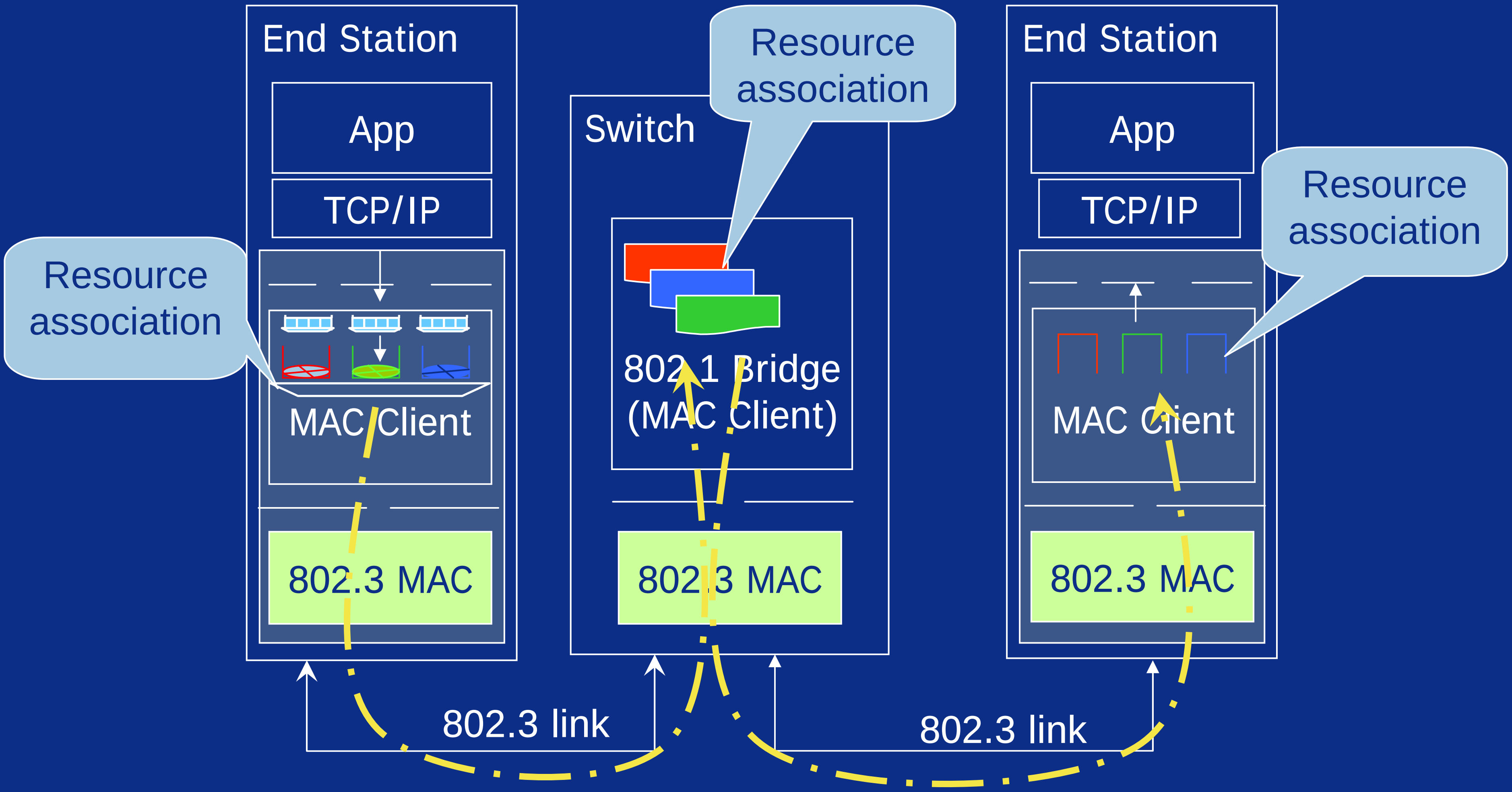
- Demand for bandwidth and connectivity is growing in the core
  - Scale out clustering, lots and lots of commodity networked machines

NOTE: http://www.ieee802.org/3/hssg/public/mar07/bechtel_01_0307.pdf

Ethernet needs to be enhanced

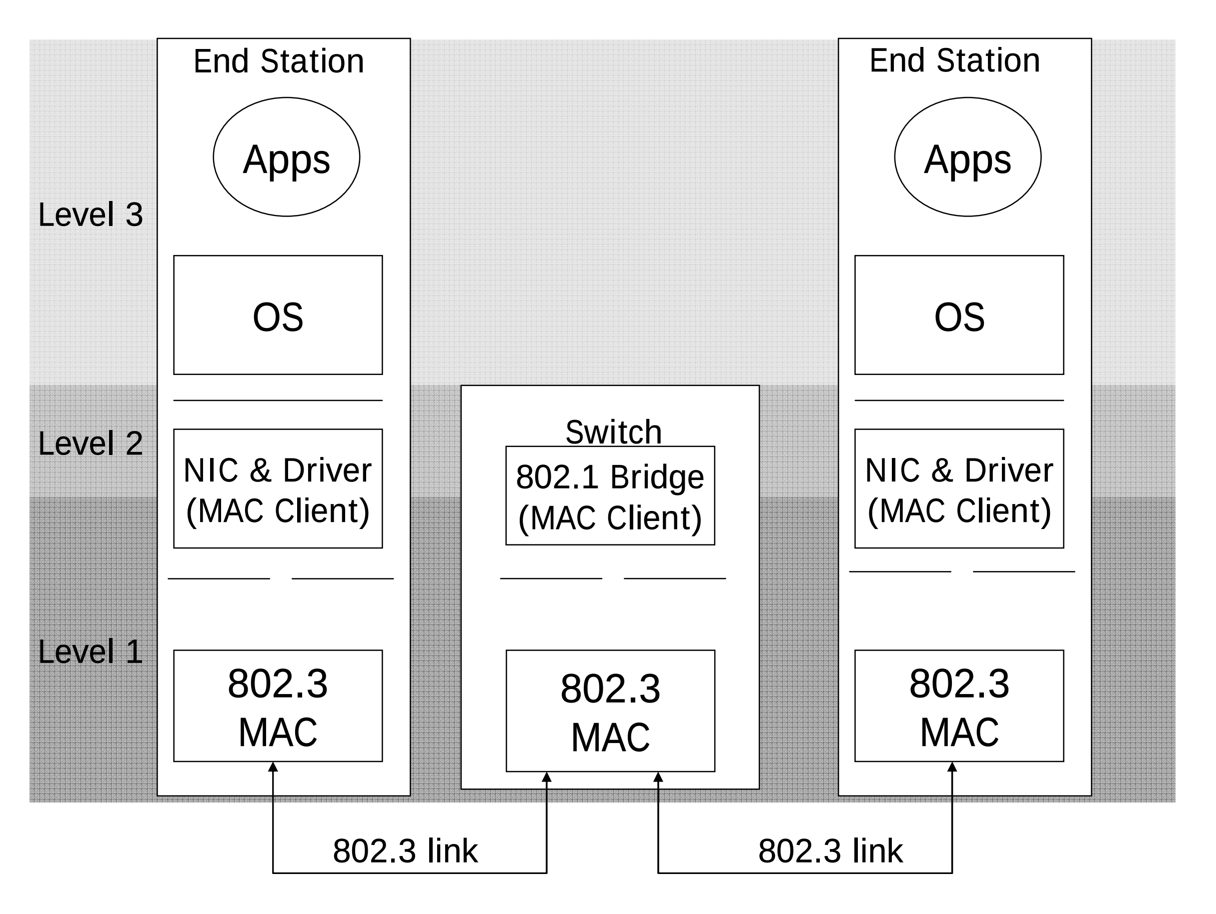
- Channelization: Consolidated traffic needs differentiation
  - Provide better than strict priority (provided by 802.1p)
  - BW/Resource sharing/Provisioning
- Reliability: Ethernet bridges drop packets when congested
  - Only solution available in 802.3X – but creates HOL blocking for whole link and also has challenges of congestion spreading
  - Need end-to-end congestion management
  - Need granular link level flow control to guarantee “no-drop” behavior
- STP reduces available bisectional bandwidth
  - Use available links in the network
  - Use Shortest Path First for forwarding

Channels: End to End Traffic Differentiation



Channels allow latency optimization for one application while allowing throughput optimization for other application

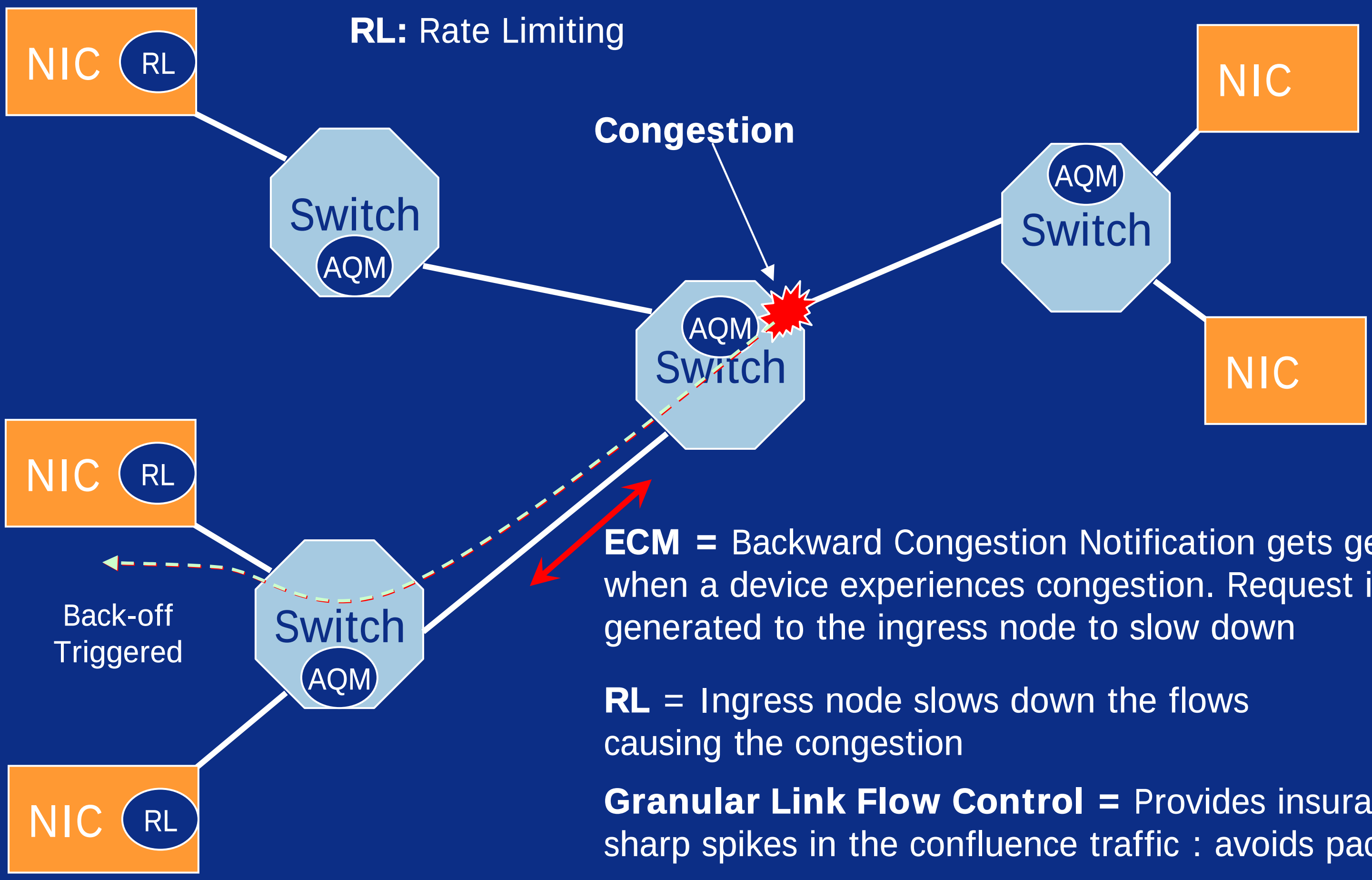
Congestion Control Hierarchy



End-to-End Congestion Management

AQM: Active Queue Management

RL: Rate Limiting



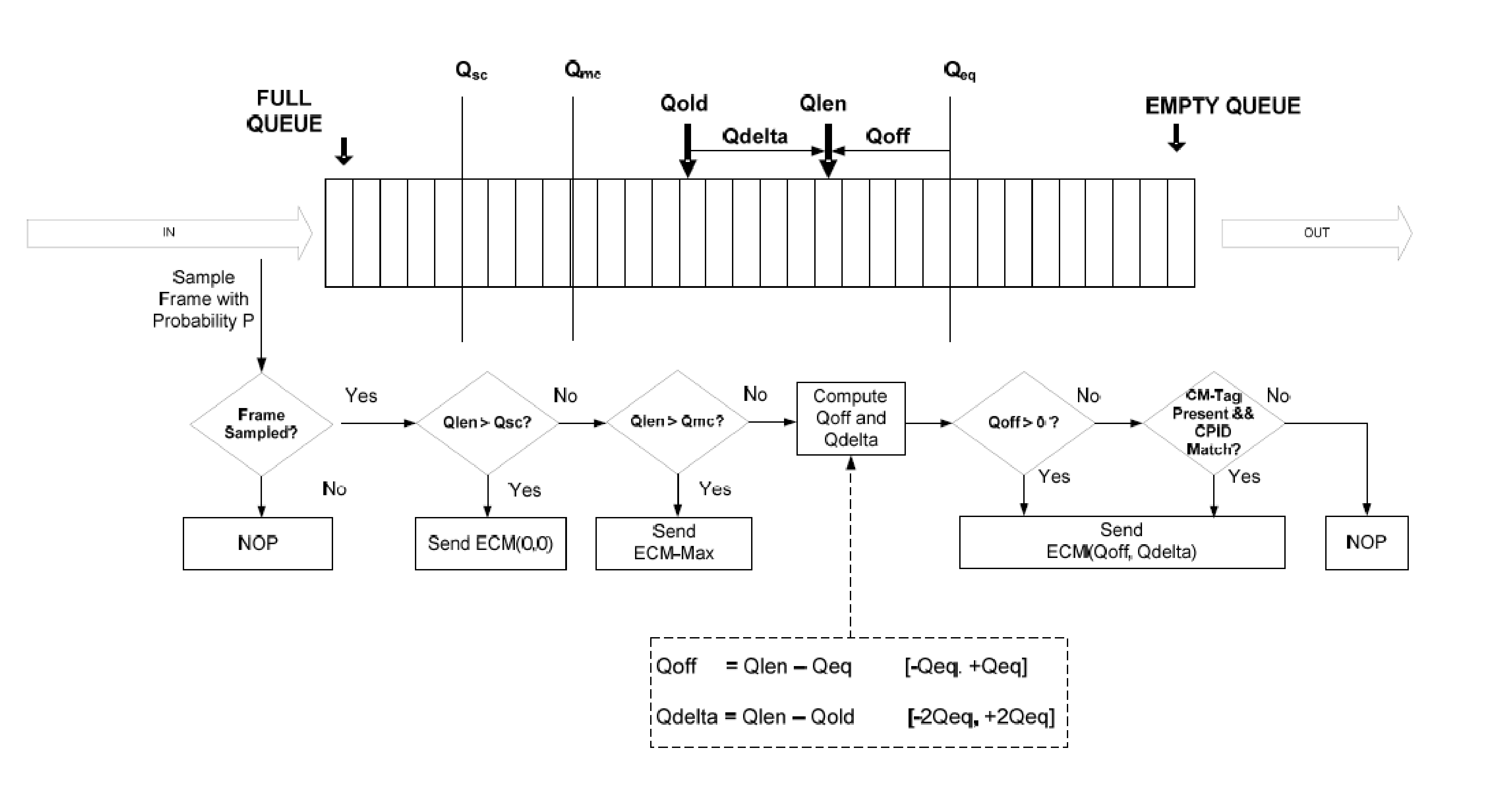
ECM = Backward Congestion Notification gets generated when a device experiences congestion. Request is generated to the ingress node to slow down

RL = Ingress node slows down the flows causing the congestion

Granular Link Flow Control = Provides insurance against sharp spikes in the confluence traffic : avoids packet drops

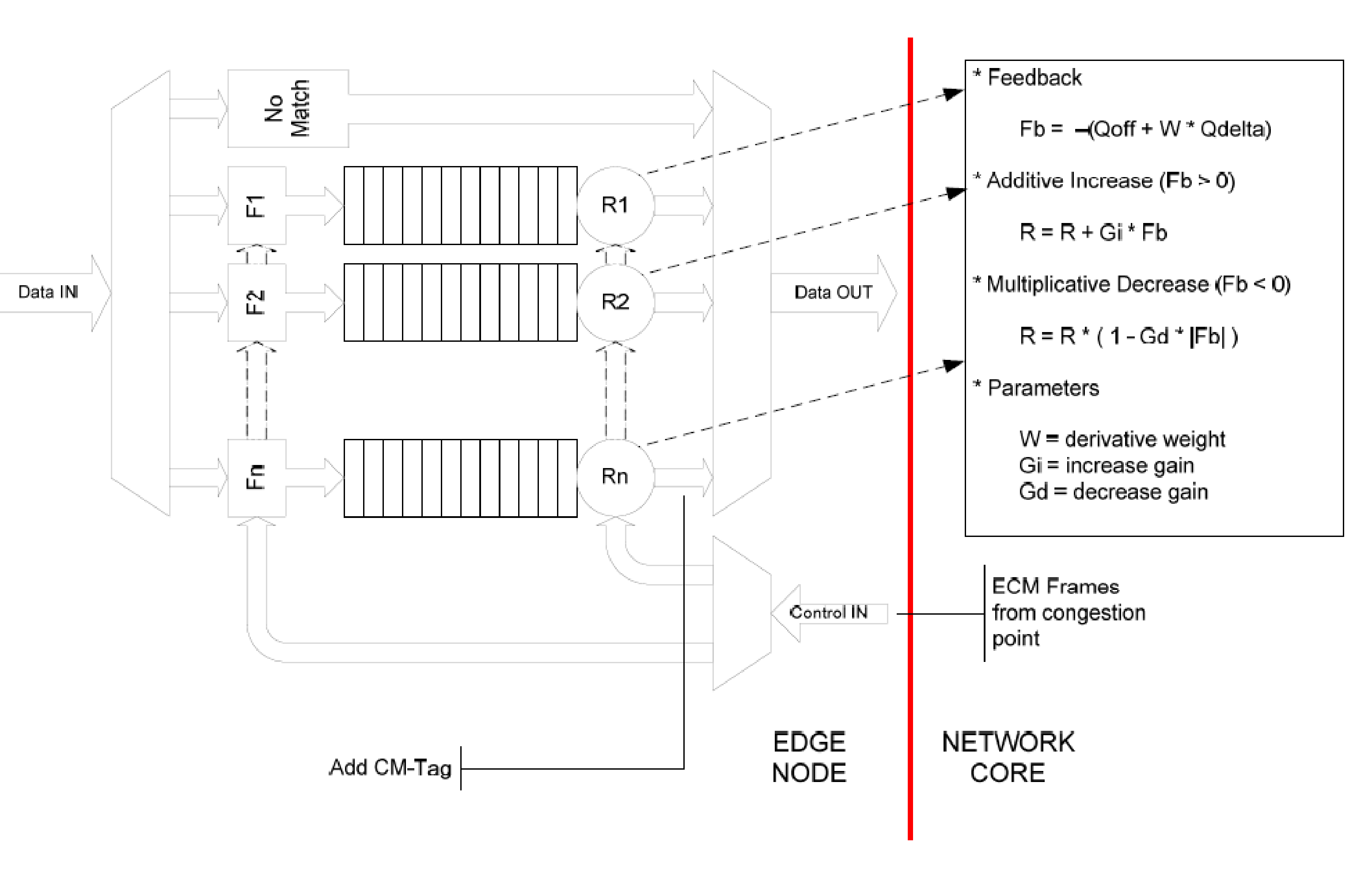
ECM = When congestion disappears, positive notification is generated to the ingress device allowing to grow the rate

Congestion Detection



NOTE:http://www.ieee802.org/1/files/public/docs2007/au-bergamasco-ethernet-congestion-manager-070313.pdf

Congestion Response



* Feedback

$Fb = -(Qoff + W * Qdelta)$

* Additive Increase ($Fb > 0$)

$R = R + Gi * Fb$

* Multiplicative Decrease ($Fb < 0$)

$R = R * (1 - Gd * |Fb|)$

* Parameters

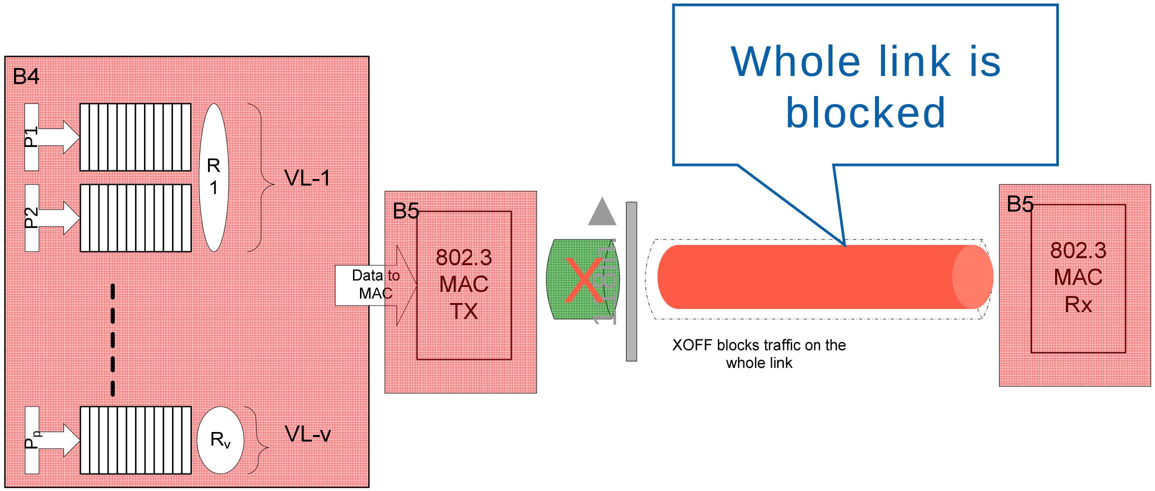
W = derivative weight
Gi = increase gain
Gd = decrease gain

ECM Frames from congestion point

More work is being done..

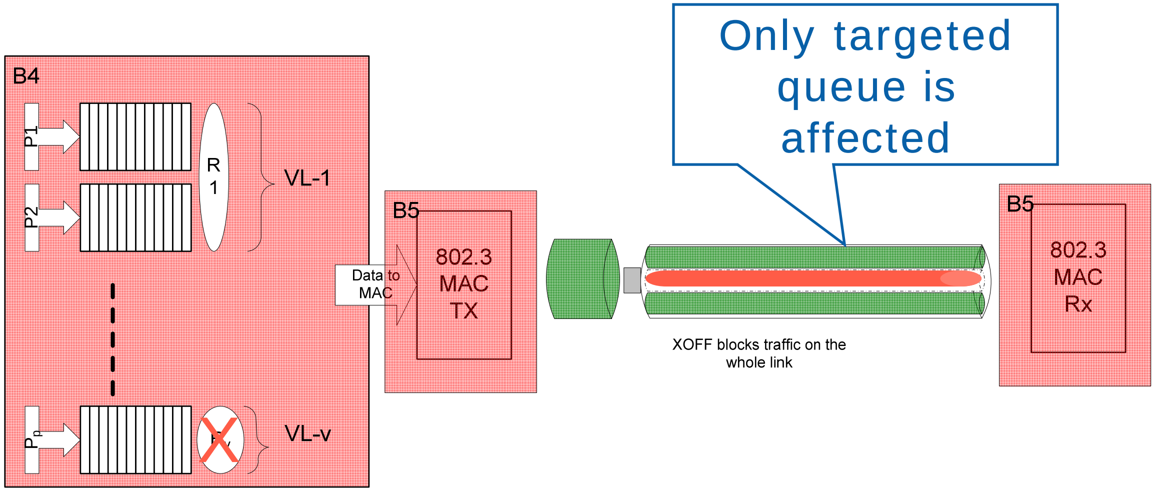
- Check IEEE 802.1Qau web site:
  - http://www.ieee802.org/1/pages/802.1au.html
- Enhanced version is being discussed: QCN (Quantized Congestion Notification)
  - Provides quantized feedback
  - Removes +ve feedback (allows disassociation of RP-CP)
  - Implements TCP-BIC type rate recovery at the RP
- There are other proposals as well
  - E2CM (Enhanced Ethernet Congestion Management): Adds probes on the top of ECM proposal – improves fairness
    - http://www.ieee802.org/1/files/public/docs2007/au-sim-IBM-ZRL-E2CM-proposal-r1.09b.ppt
  - FECN (Forward ECN): Rate Allocation mechanism
    - http://www.ieee802.org/1/files/public/docs2007/au-jain-fecn-enhanced-20070530.pdf.filepart

Priority based flow control



Whole link is blocked

Pause

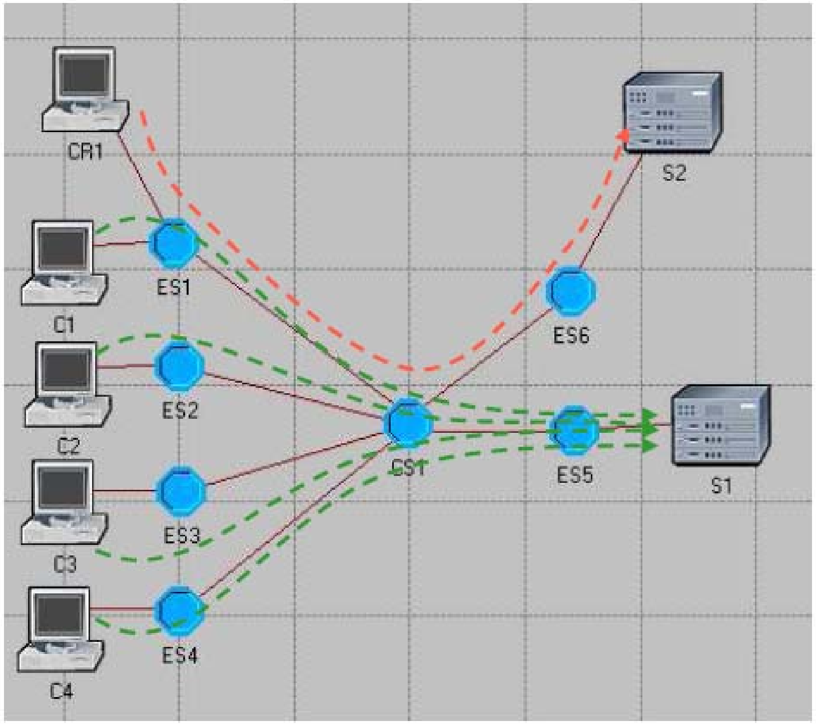


Only targeted queue is affected

Granular Pause

Simulation Results: Congestion Control

Topology and Workload

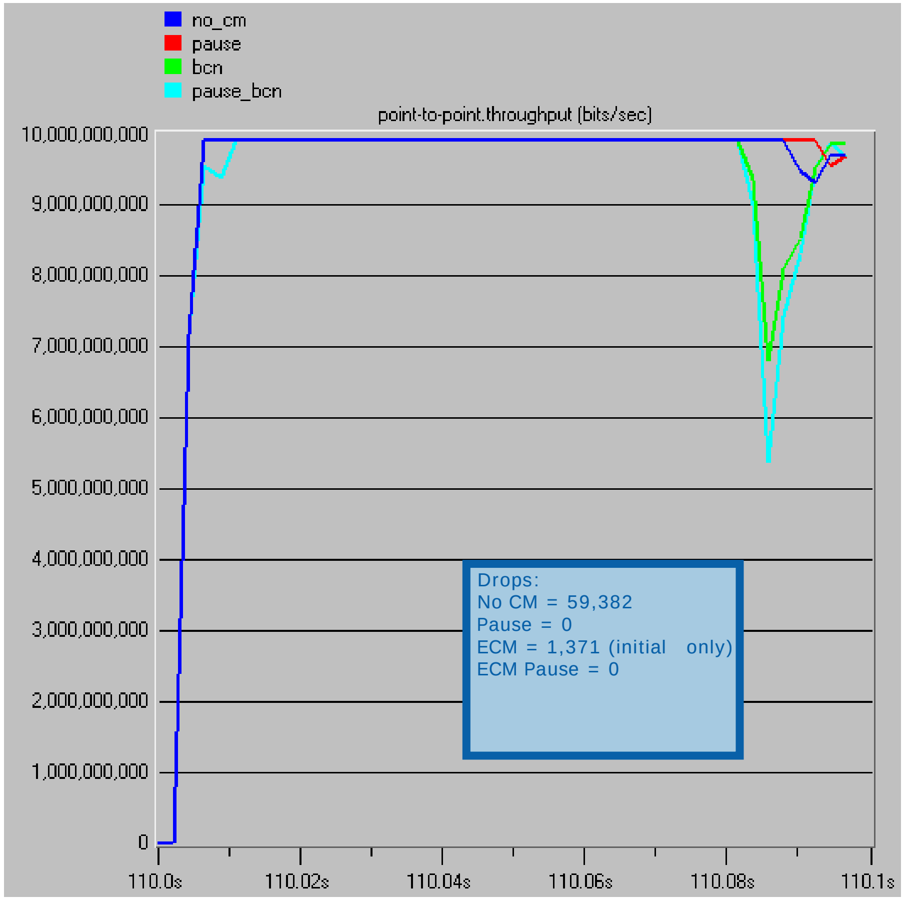


- All links 10Gbps
- Output buffered Switch with 150KB/port
- 150KB of buffering in Host, but traffic source stops after memory full (no drops)
- Latency
  - Switch = 1us
  - Each link = 0.5us
  - Host response time = 2uS
- Sources C1, C2, C3, C4 sending ~4.8Gbps of UDP data to S1
- Reference Source CR1 sending ~4Gbps of UDP data to S2
- 1500 byte fixed payload size
- Bernoulli temporal arrival distribution
- Total run time = 100ms
- All sources start at 5ms
- 2 sources stop at 85ms

Parameters

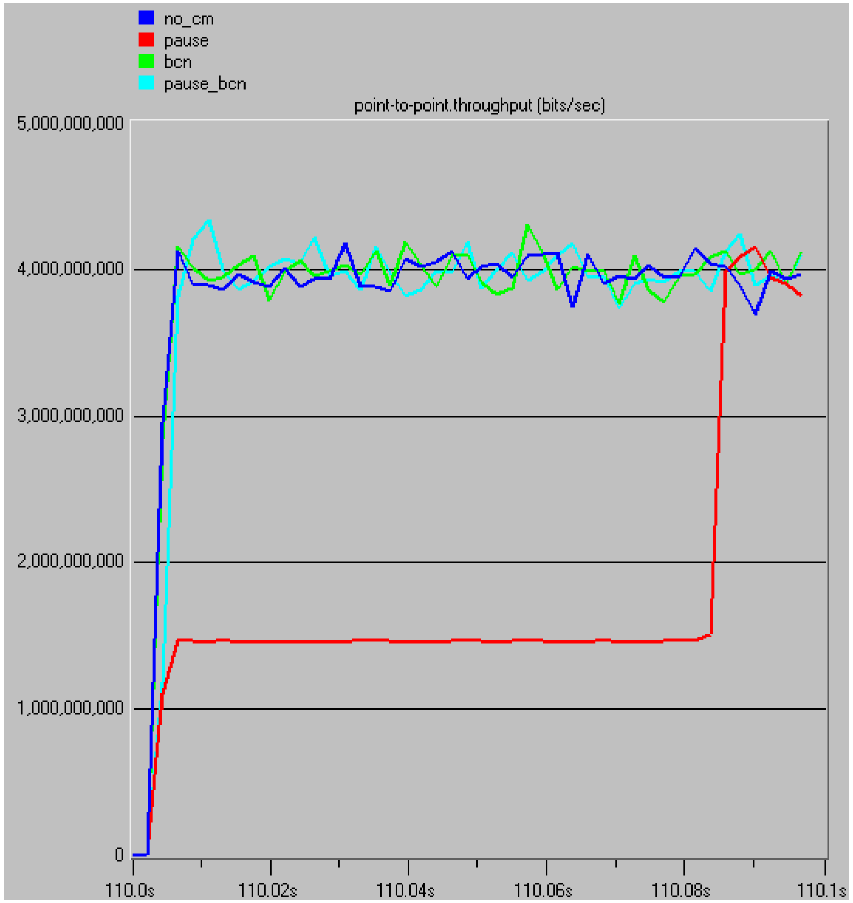
- BCN parameters:
  - Qeq = 375 64 byte pages
  - Sampling interval = 150KB +- 20KB
  - W = 2
  - Gd = 5.3 * 10 ^-1
  - Gi = 2.6 * 10^-4
- PAUSE parameters:
  - XON/XOFF threshold sets towards the top of the switch output port buffer
    - XOFF threshold = 136,192 bytes
    - XON threshold = 123,392 bytes
- Global PAUSE
  - XOFF is sent to all ports except the current port when buffer >= XOFF Threshold
  - XON is sent to all ports except the current port when buffer falls below XON threshold

Link Throughputs: Congested and Innocent



Drops:
No CM = 59,382
Pause = 0
ECM = 1,371 (initial only)
ECM Pause = 0

Congested Link



Innocent Link

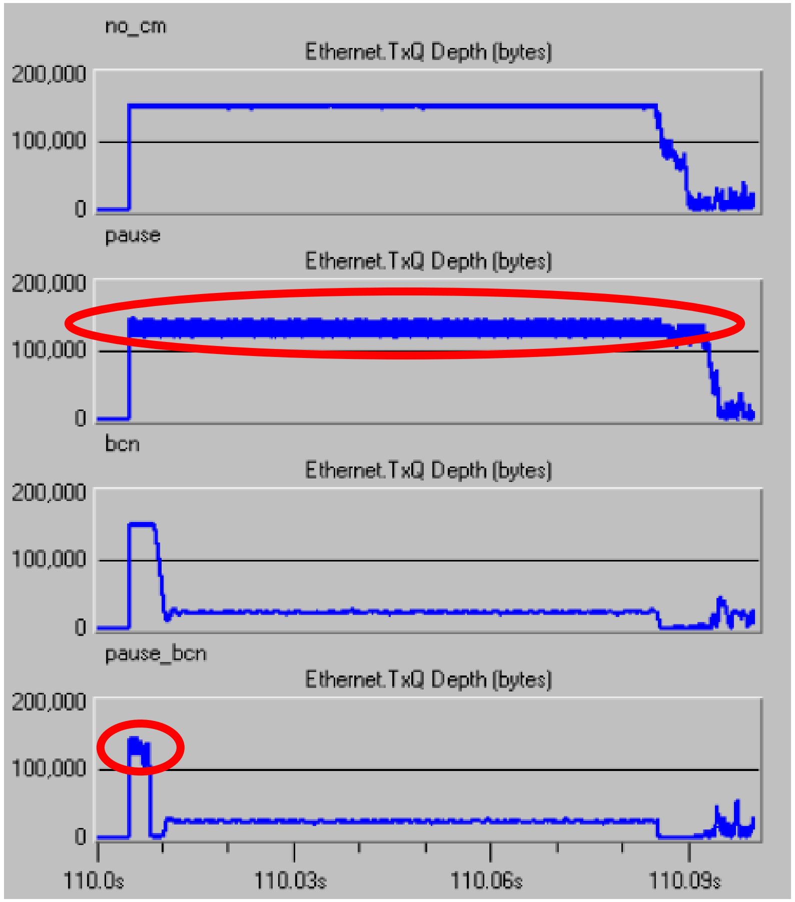
Congestion Spreading

No_cm

Pause

BCN

Pause+BCN



Pkt drops = 1371

Pkt drops = 0

Summary

- Channelization enables traffic differentiation for consolidated traffic types (LAN, SAN, IPC)
- Congestion Management enables Ethernet to allow “no-drop” service
  - Enables storage traffic over Ethernet (Refer to work in T11: FC over Ethernet)
- Shortest Path Bridging allows improvement in available bisectional bandwidth in data center networks
  - Also allows reduction in end-to-end latency
- Enhanced Ethernet for Data Centers will enable newer protocols
  - FCoE is the beginning
- What’s next?
  - Plug-n-play protocols?
  - IPC traffic improvements?
  - TCP for data centers?