IEEE 802 CMSG Tutorial

San Antonio, TX

November 16, 2004

Contributors and Supporters

Agenda

- Kick off
  - Presenter: Ben Brown; Chair, Congestion Management Study Group
- Market Requirement / Business Case
  - Presenters: Gopal Hegde; Intel, Shashank Merchant; Nokia
- Distinct Identity & Joint work between 802.3 and 802.1
  - Presenter: Hugh Barrass; Cisco Systems
- Technical Feasibility / Modeling Data
  - Presenter: Manoj Wadekar; Intel
- Wrap-Up and Q&A

Introduction and Overview (taken from PAR and 5 Criteria)

- Ethernet networks are being used in an increasing number of application spaces (clustering, backplanes, storage, data centers, etc.) that are sensitive to frame delay, delay variation and loss

- Congestion management, when used, may reduce the offered load at the congestion points without spreading congestion. This specification will define a means of decreasing frame loss while permitting increased efficiency in the Ethernet network

- Mechanisms for congestion management using congestion indication are known in the industry for some protocols and standards. Simulations of similar protocols show there are alternatives that can be feasibly implemented to accomplish the objectives within IEEE 802.

History

- Nov, 2003: Backplane Ethernet CFI
- March, 2004: Congestion Management study group spawned from Backplane
- May, 2004: First meeting, decided not yet ready for PAR – still trying to understand the issues
- July, 2004: First objectives
- Sept, 2004: Refine objectives, PAR and 5 criteria. Split problem into 2 areas, solve one of them in 802.1

Participation

- March, 2004: 23 people, 16 companies
- May, 2004: ~25 people
- July, 2004: 22 people, 16 companies
- Sept, 2004: 30 people, 16 companies

Objectives

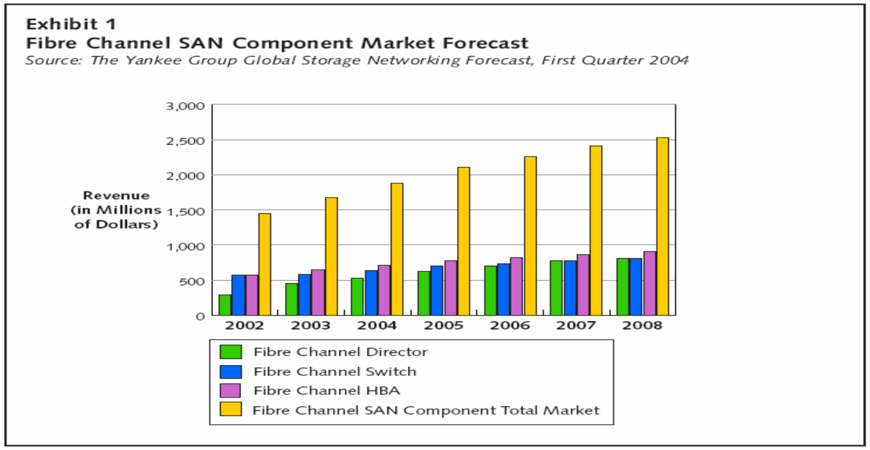
- Specify a mechanism to support the communication of congestion information
- Specify a mechanism to limit the rate of transmitted data on an Ethernet link
- Preserve the MAC/PLS service interfaces
- Minimize throughput reduction in non-congested flows

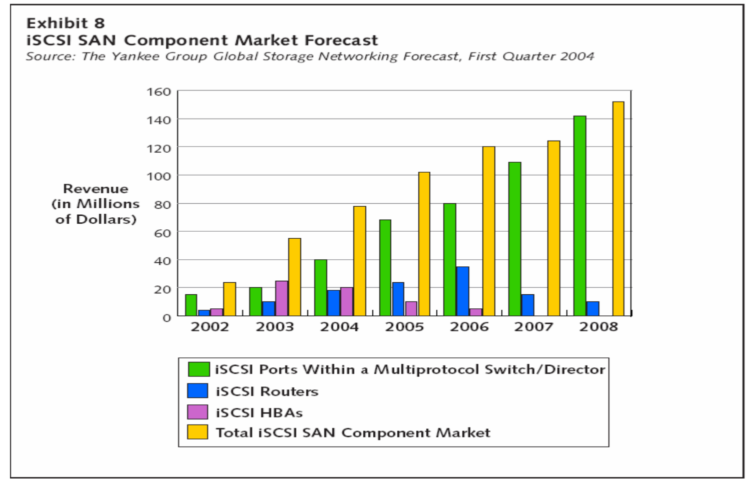
Market Requirements for Congestion Management

Gopal Hegde

Intel Corp.

Storage Components Market

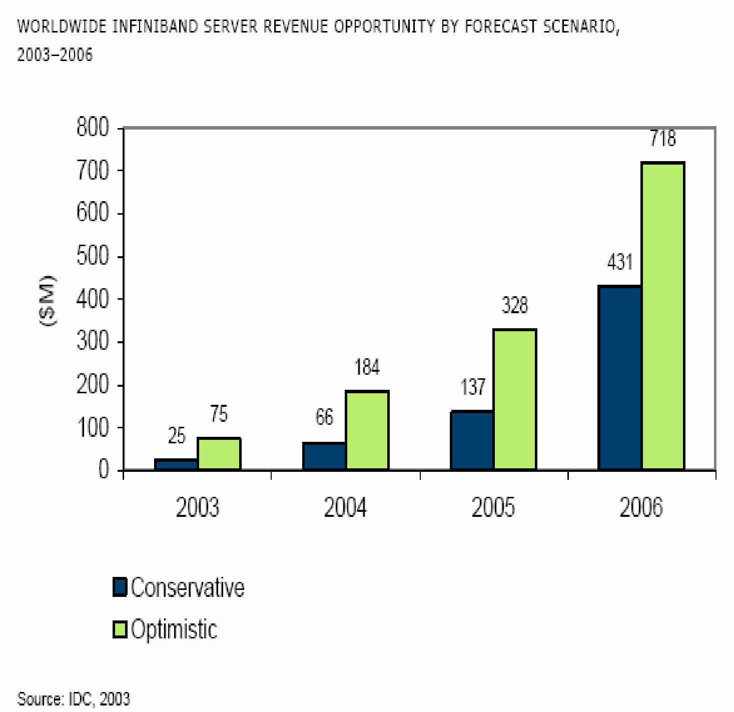




- FC continues to be the dominant SAN technology, ~70% MSS into ‘07
- iSCSI adoption has been slow despite being more cost effective
- F500 IT concerns include
  - Security
  - Performance -- Ethernet behaves poorly in congested environments, packet drops significant, adversely affects storage traffic

Improving Ethernet congestion management can accelerate iSCSI adoption – addresses IT perception & reality

Ethernet Opportunity for Clustering and IPC



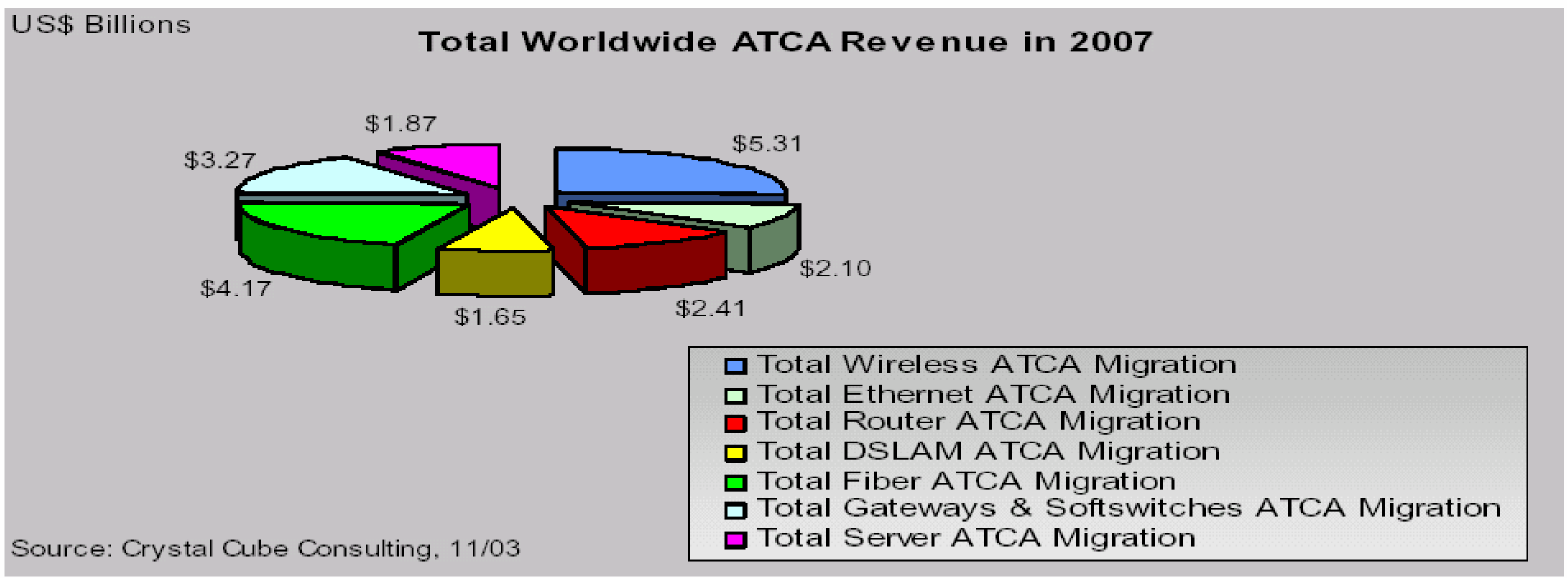
- Clustering –
  - Growth Opportunities include
    - "Technical Capacity" Servers ~ 20% of High Performance Computing (HPC) market by 2007
    - Database clusters
  - Clusters built using low cost servers connected by a high performance, low latency fabric
- Users like the cost structure and availability of Ethernet
  - However latency and congestion management are key issues
- Myrinet and Quadrics based fabrics are being deployed to address this need
- Infiniband ® emerging as fabric of choice for clustering

Addressing latency and packet loss opens up the cluster market for Ethernet

Telco Backplane Opportunity for Ethernet

- **Blades cut into Telco pie ~ 26% of Telco servers by '07 – In-Stat/MDR**
- **Advanced Telecom Computing Architecture (ATCA) is a PICMG based standard for Telecom blades**
- **ATCA specifications include Ethernet backplanes (1 GbE and 10 GbE)**
- **A number of major Telecom equipment vendors are adopting ATCA**

Figure 15. Worldwide ATCA Projection of Revenue in 2007 by Market Segment



Datacenter Requirements

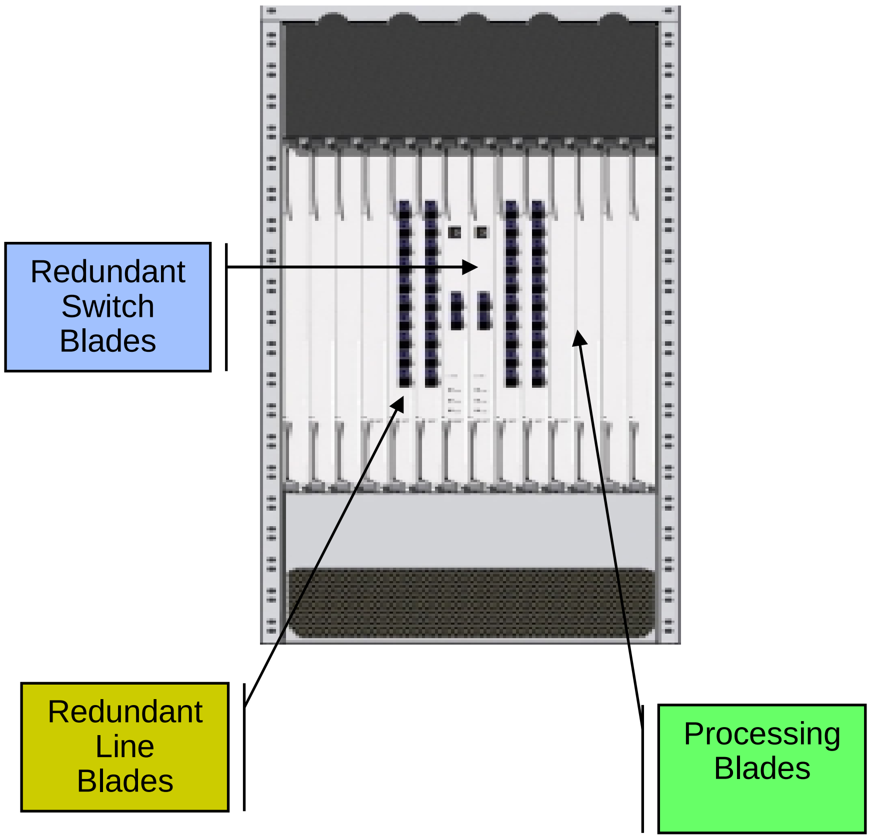
- Address IT perceptions:
  - "Ethernet not adequate for low latency apps"
  - "Ethernet frame loss is inefficient for storage"
- 802.3x does not help
  - Reduces throughput
  - Congestion spreading
  - Increases latency jitter
- Improve Ethernet Congestion Management capabilities that will:
  - Reduce frame loss significantly
  - Reduce end-to-end latency and latency jitter
  - Achieve above without compromising throughput

Congestion Management in a Bladed System

Shashank Merchant

Nokia

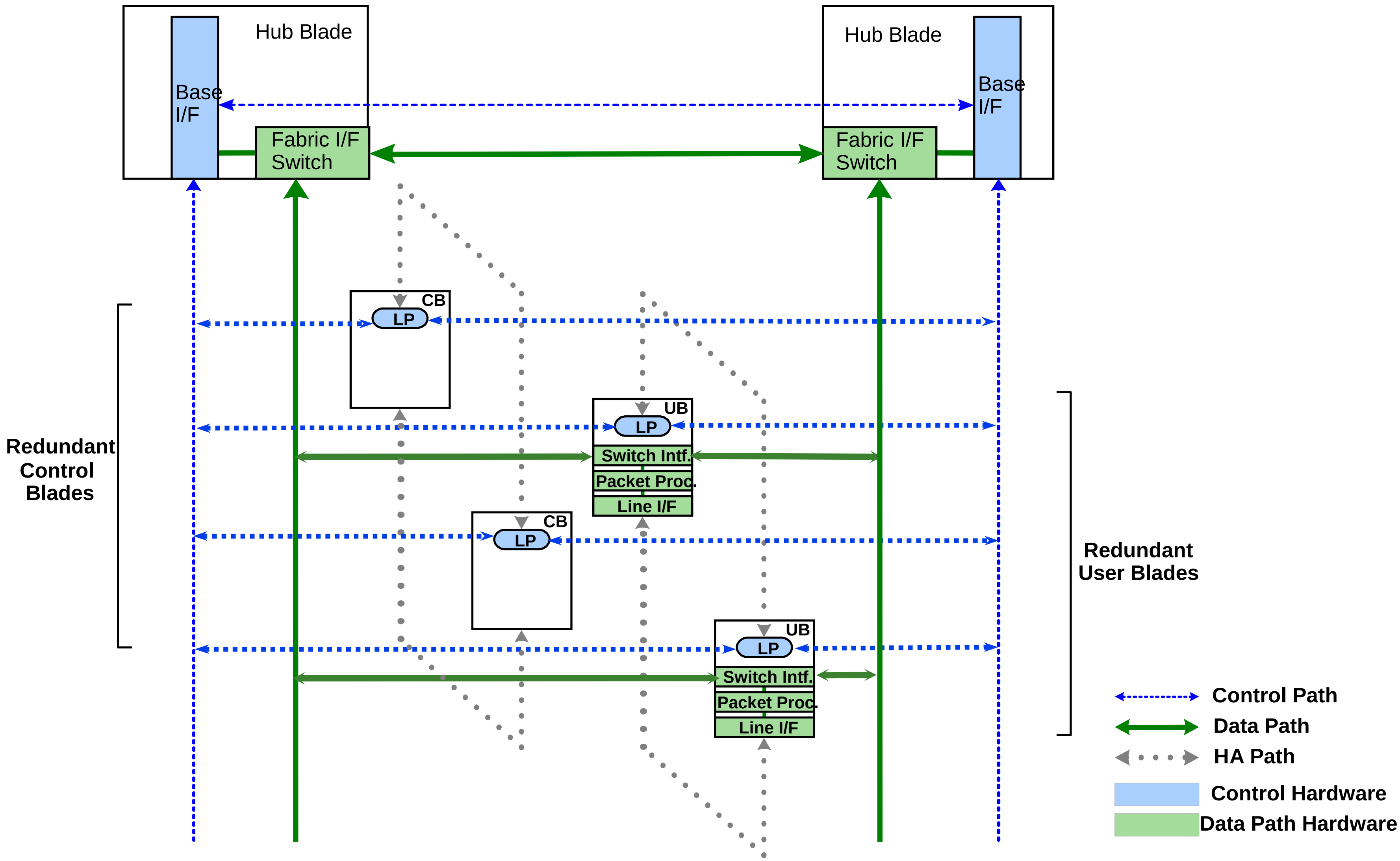
Example System



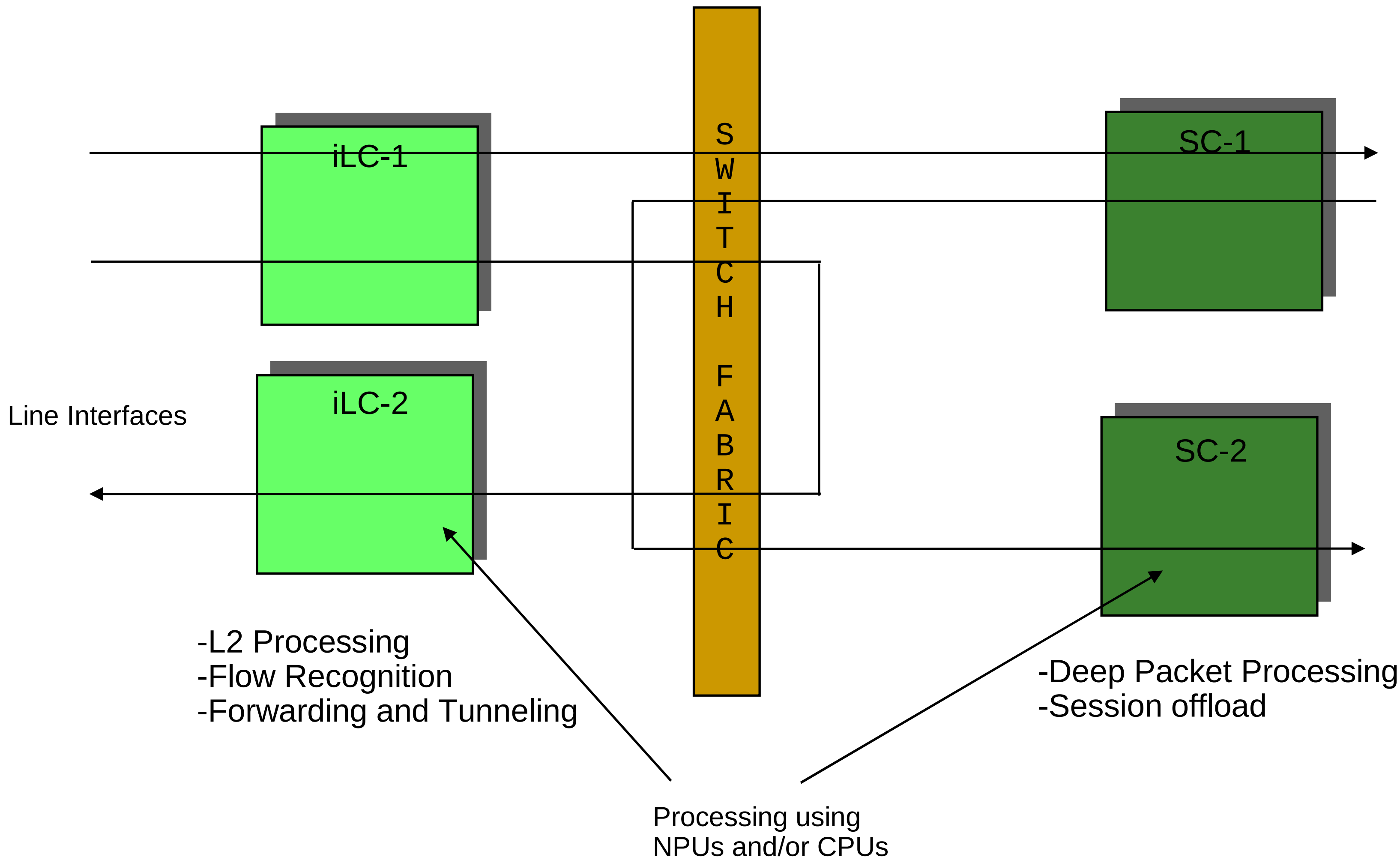
- Bladed System
  - Redundant Switch Blades
  - Multiple Line & Processing Blades
    - 1:1 or n:1 redundant
- Highly available (99.999% +)
  - Fast switch-over, minimum packet loss
- Line Blades provides I/O interfaces, and some processing
- Protocol and service processing in the processing blades
- Asymmetric bandwidth/ performance, and bursty traffic among blades

  Traffic aggregation and segregation is a natural consequence
- Latency/jitter for certain traffic classes is an absolute must

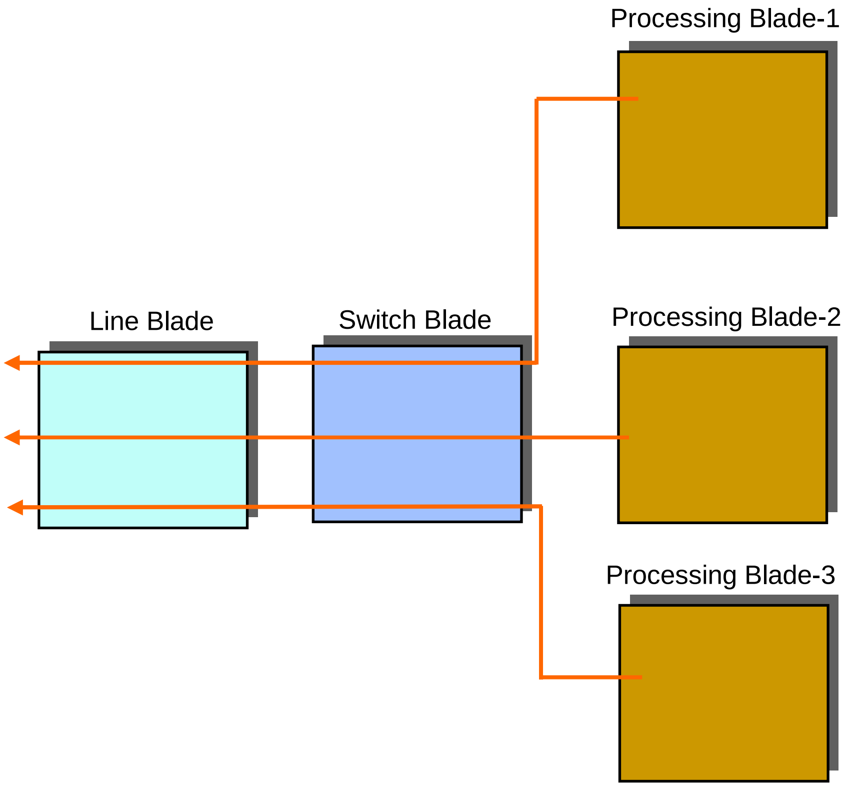
Separate User and Control Paths



Basic User-Data Path



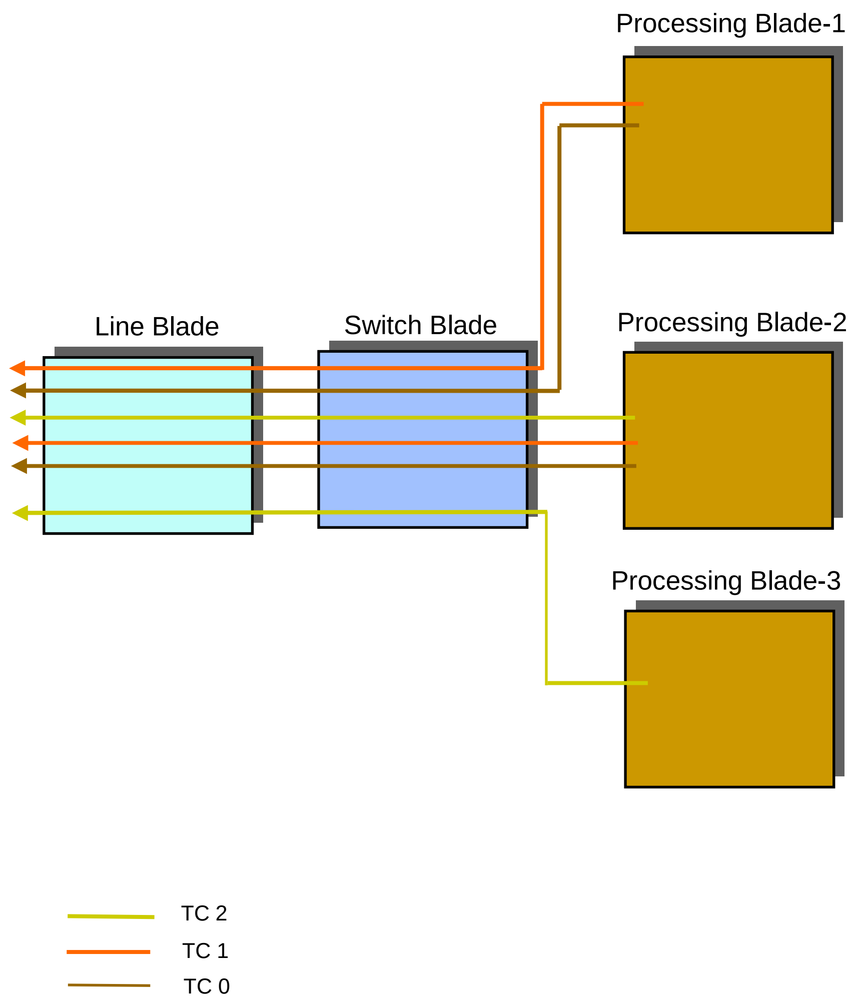
Scenario 1



- Traffic flowing from multiple processing blades to single line card
  - Single priority class

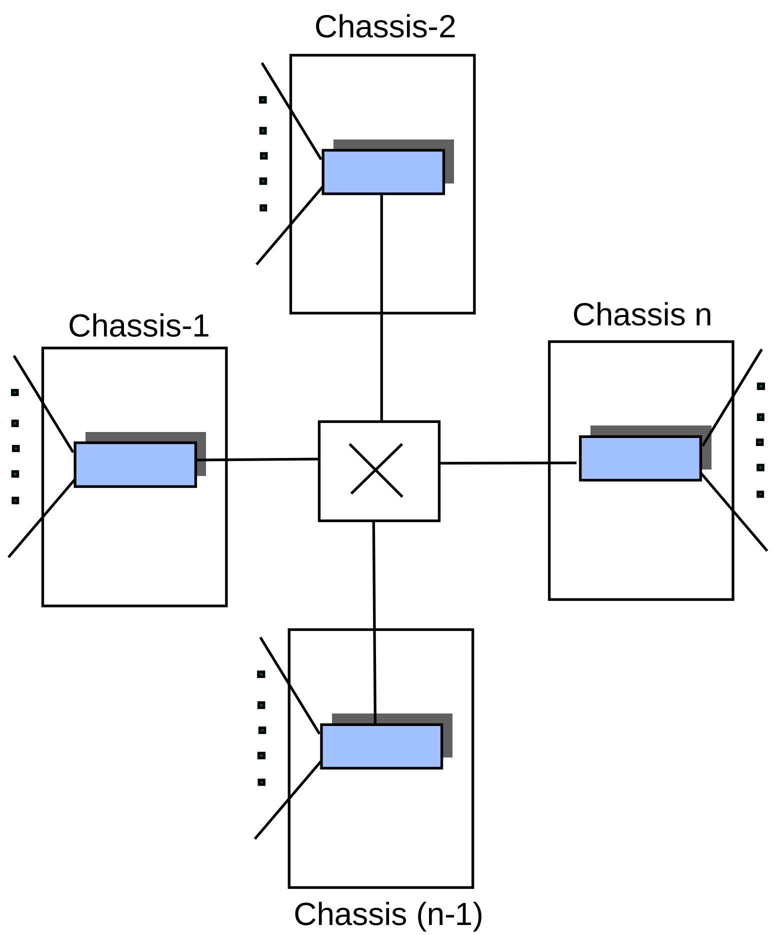
  (each one is independent, and not aware of other traffics)
- Packets should not be discarded in the switching sub-system
  - Discard else where based on service/traffic type

Scenario 2



- Traffic flowing from multiple processing blades to single line card
  - Multiple traffic classes
- Congestion information per traffic class
- Different latency/jitter requirements per traffic class
- Packets should not be discarded in the switching sub-system
  - Discard else where based on service/traffic type

Scenario 3



- Connection between Chassis may be blocking
- Multiple traffic classes and potentially mix of control and user traffic
- Need for congestion management scheme that doesn't drop packets in the switching sub-system
- Cabling requirements within 15-20m

Observations

- Effective congestion management is an absolute must for the carrier-grade systems
- Congestion Management implementations should be in Hardware.
  - Software involvement for configuration and monitoring purpose only
- 802.3x PAUSE protocol provides simplicity but
  - Increases latency and Jitter
  - Decreases throughput
- 'Intelligent' rate limiting may be required
  - However system complexity and cost needs to be understood
- Must respect 802.1p Class of Service
- High availability requirements like fast switch-over, and minimum packet loss must not be compromised due to any congestion management solution
- Use of Ethernet as a backplane technology requires understanding and solving these concerns

Distinct Identity & Joint work between 802.3 and 802.1

Hugh Barrass

Cisco Systems

Distinct identity

CMSG has focused primarily on solutions to improve performance of short range networks in the presence of congestion

Data center networks demonstrate the distinctive nature of short range networks

<u>Typical (and arbitrary) characteristics</u>

	Data Center	Enterprise LAN	WAN
End to end latency	low	medium	high
Session duration	medium	short	long
# of sessions / node	low	medium	high
Sustained data rate	high	medium	low

Ethernet networks

To improve congestion performance in Ethernet networks, we need to define what we mean by "Ethernet Networks."

IEEE 802.3 defines the Ethernet MAC, Ethernet PHYs and some other related stuff – this is the traditional definition of "Ethernet."

Almost all instances of Ethernet today include more than 802.3:

- **IEEE 802.1 defines bridging, including priority, VLANs, spanning tree etc.**
- **Most Ethernet networks use Internet Protocol (as defined by IETF)**
- **Although TCP is common, many transport protocols are supported**

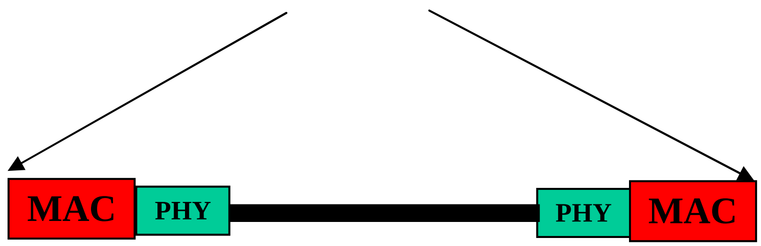
"Ethernet Networks" could be used to describe networks using 802.3 links, connected together by 802.1 bridges.

The Ethernet guarantee

802.3 can offer a guarantee for QOS
(For point-to-point Ethernet links)

Ethernet never drops a packet

MAC service interfaces

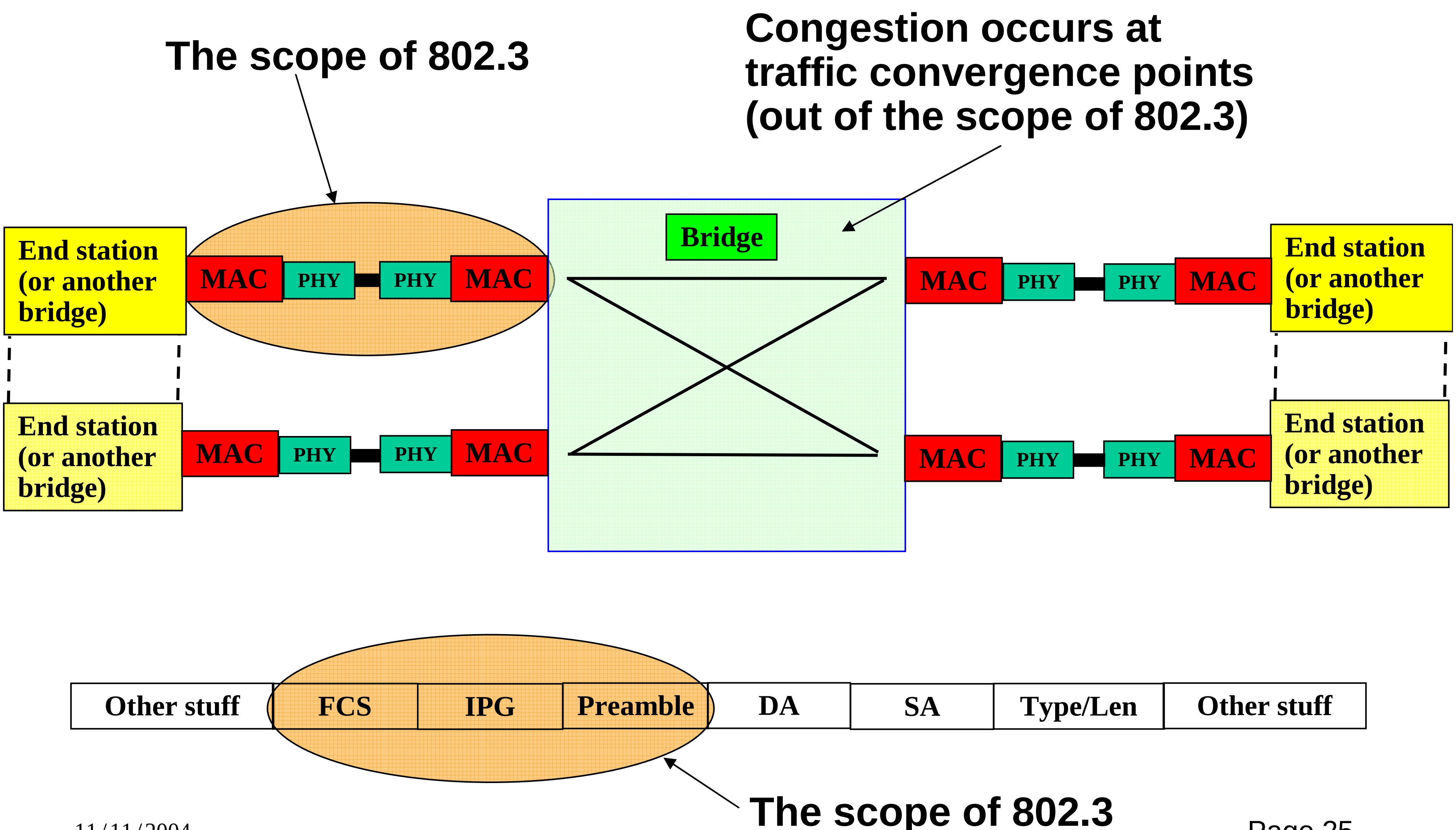


Every single packet put in at one end

Comes out at the other end*

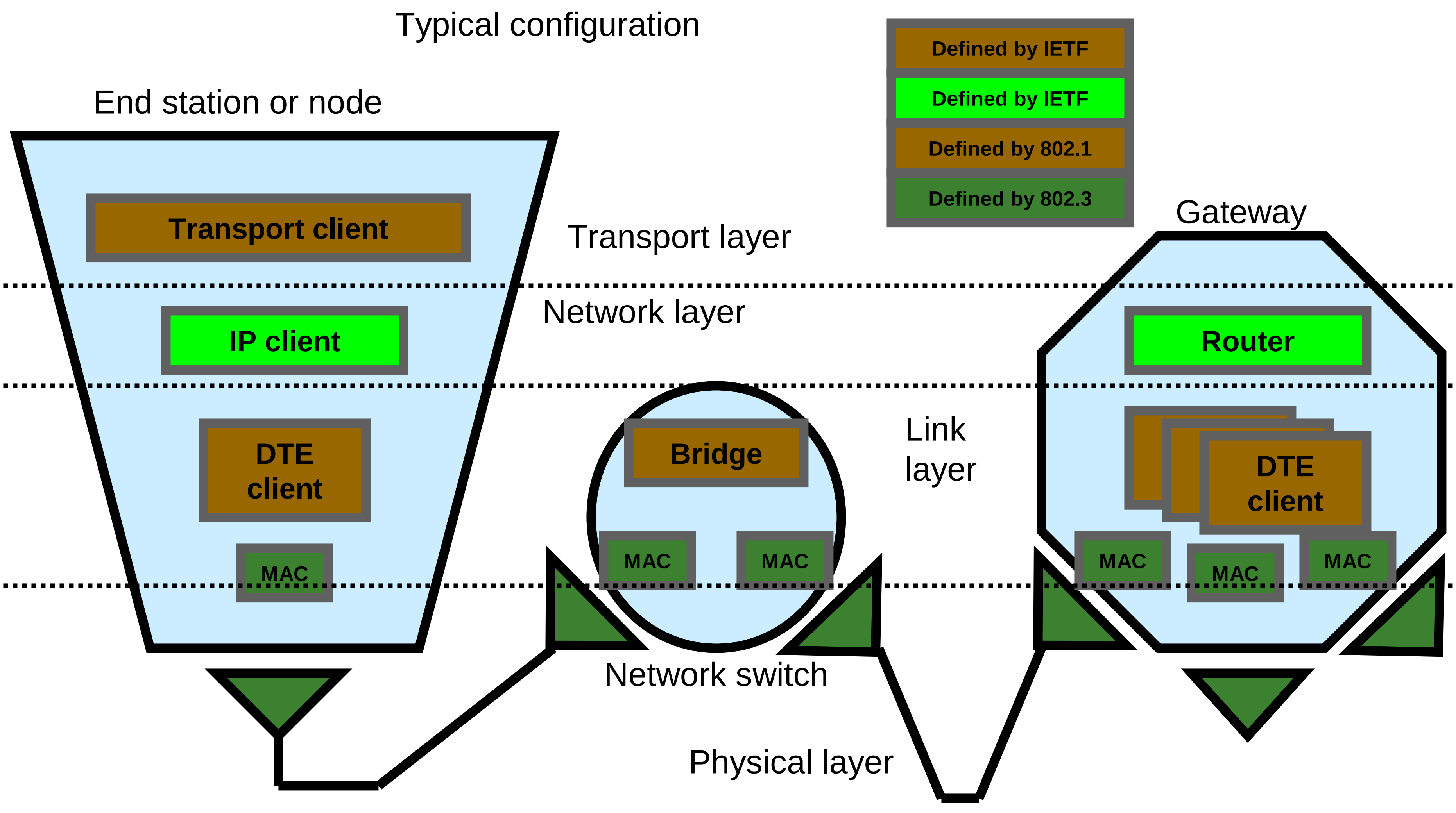
***Subject to restrictions imposed by the laws of physics and the Bit Error Ratio**
All other offers notwithstanding
Your mileage may vary

But…

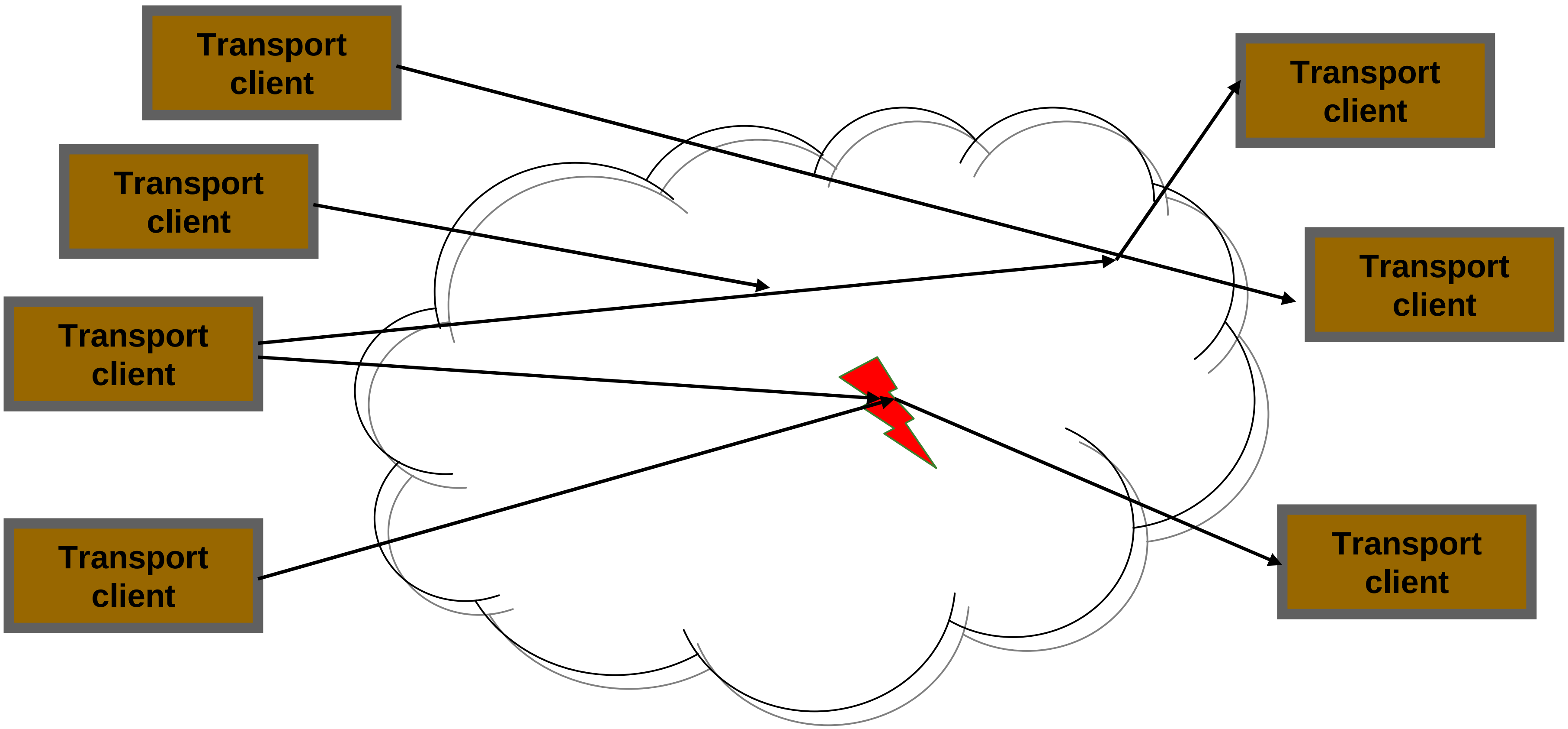


The scope of 802.3
Congestion occurs at traffic convergence points (out of the scope of 802.3)
Bridge
End station (or another bridge)
MAC
PHY
PHY
MAC
End station (or another bridge)
MAC
PHY
PHY
MAC
MAC
PHY
PHY
MAC
End station (or another bridge)
MAC
PHY
PHY
MAC
End station (or another bridge)
Other stuff
FCS
IPG
Preamble
DA
SA
Type/Len
Other stuff
The scope of 802.3

So, who is responsible?



Congestion in the network cloud

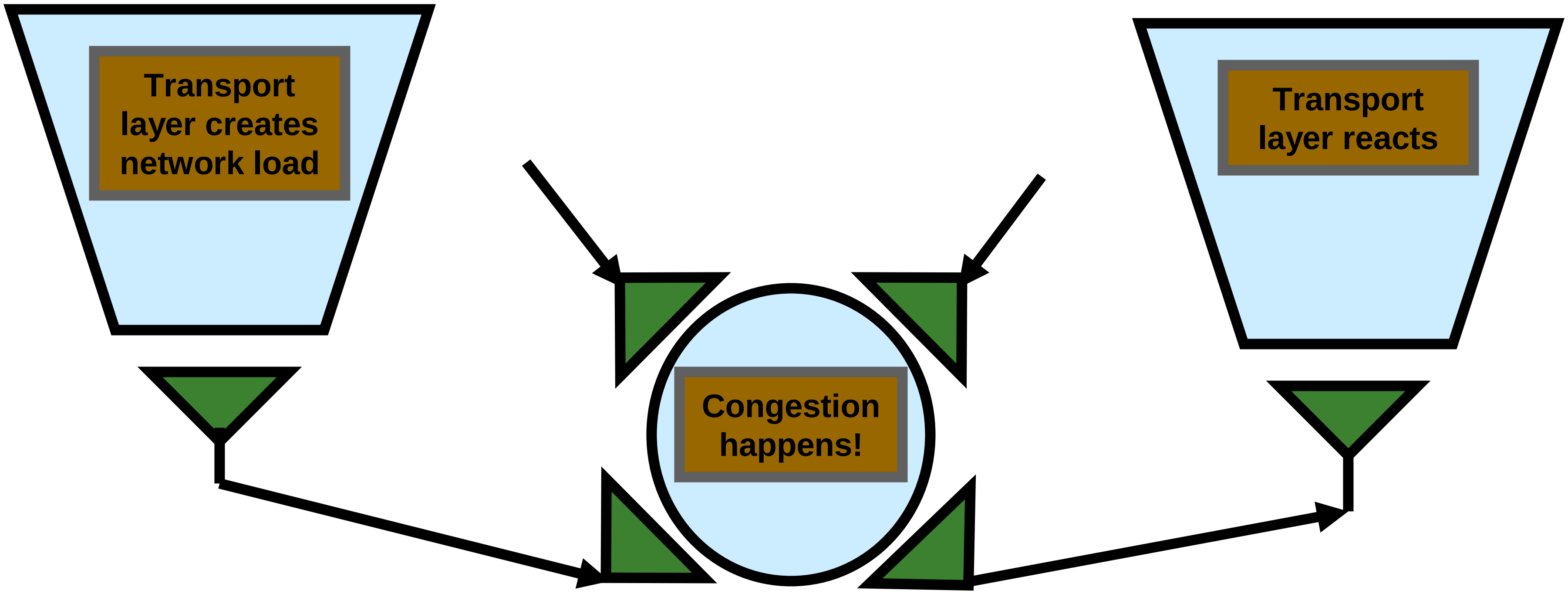


In arbitrary network topology connectivity cannot be assumed

Only by adjusting effected transport can congestion be remedied…

… without perturbing innocent conversations

When congestion happens!



Transport layer sends data into the network,
Congestion happens in the bridge,
Causing a reaction in the transport layer

Problems with transport adjustment mechanisms

- Transport adjustment often relies on packet loss
  - Retries are expensive – timeouts are disastrous for data center traffic!
  - Not only a problem with TCP
- Transport adjustment mechanisms are generally optimized for internet-like topologies
  - Transport windows are very large, requiring large network buffers
  - Reaction times are slow
- Data center traffic is bursty in time & space
  - Typically clients send bursts to various destinations
  - Causes congestion points to move
  - Needs fast reaction times in transport to avoid "misadjustment"

So where do we fix the problem?

- Congestion happens at convergence points
  - 802.1 defines the bridges that include the congestion (for L2 networks)
  - Notification should be defined in 802.1
- Reactions required in end stations
  - Need for definition of end station behavior
  - Where should that reside & what needs to be defined?
- What can be done in 802.3?
  - … anything that effects a single link
  - e.g. controlling the rate of a link
- n.b.
  - 802.3 is also the home of "willing" volunteers for simulation etc.

Technical Feasibility / Modeling data

Manoj Wadekar

Intel Corp.

An Example Approach:
L2– Congestion Indication

Issue:

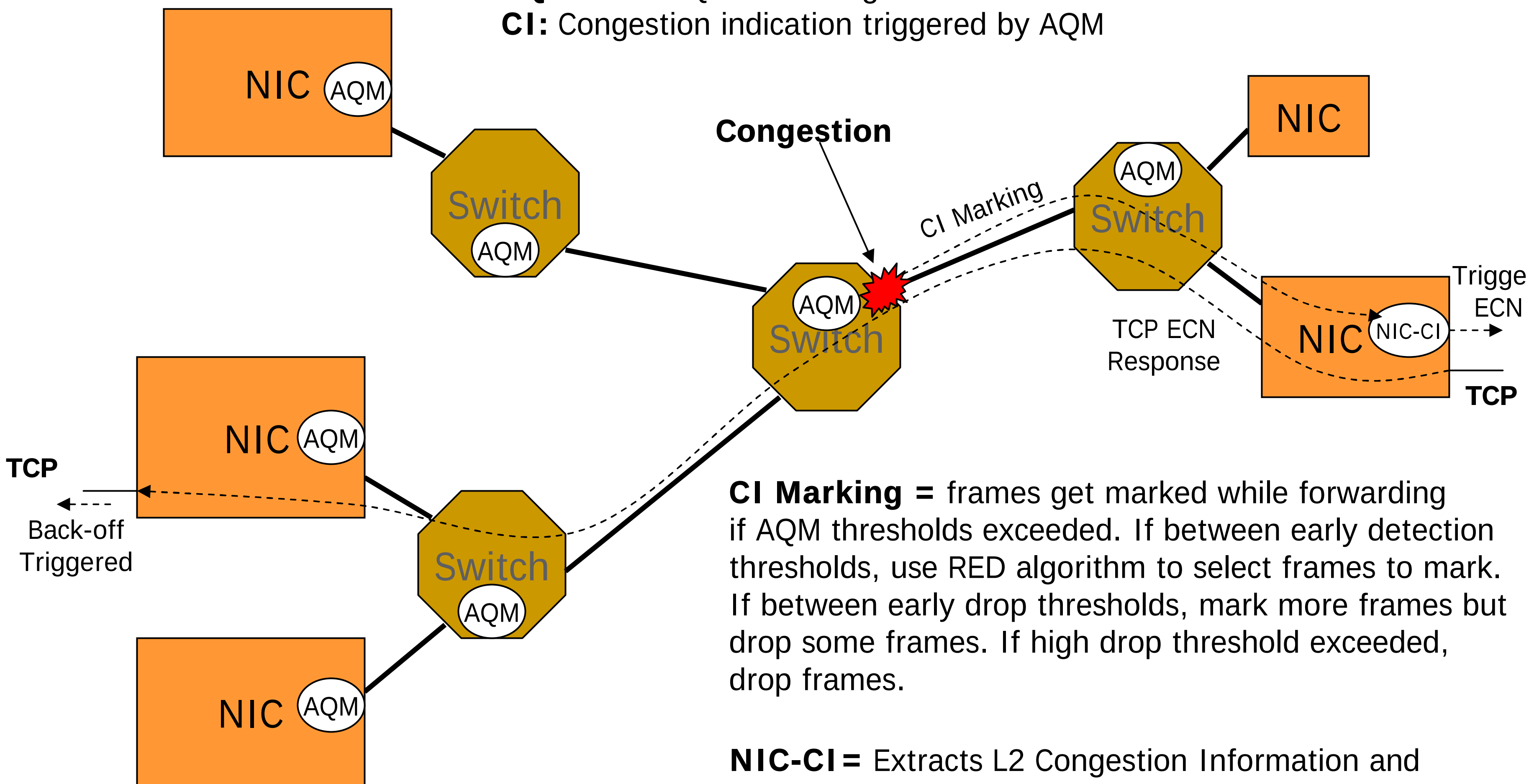
- Congestion due to oversubscription
- “Reactive” rate control in TCP

Method:

- “Rate Control” is done at end-points based on congestion information provided by L2 network
  - Provide Congestion Information from the network devices to the edges
  - Modification to NIC Driver to pass congestion information to protocols
- Various mechanisms possible for Congestion Indication
  - Marking, control packet, forward/backward/both
- TCP applications can benefit
  - ECN can be triggered even by L2 congestion
  - “Proactive” action by TCP, avoids packet drop
- Non-TCP applications can leverage
  - New mechanism to respond to congestion

Model Implementation:
L2 Congestion Indication

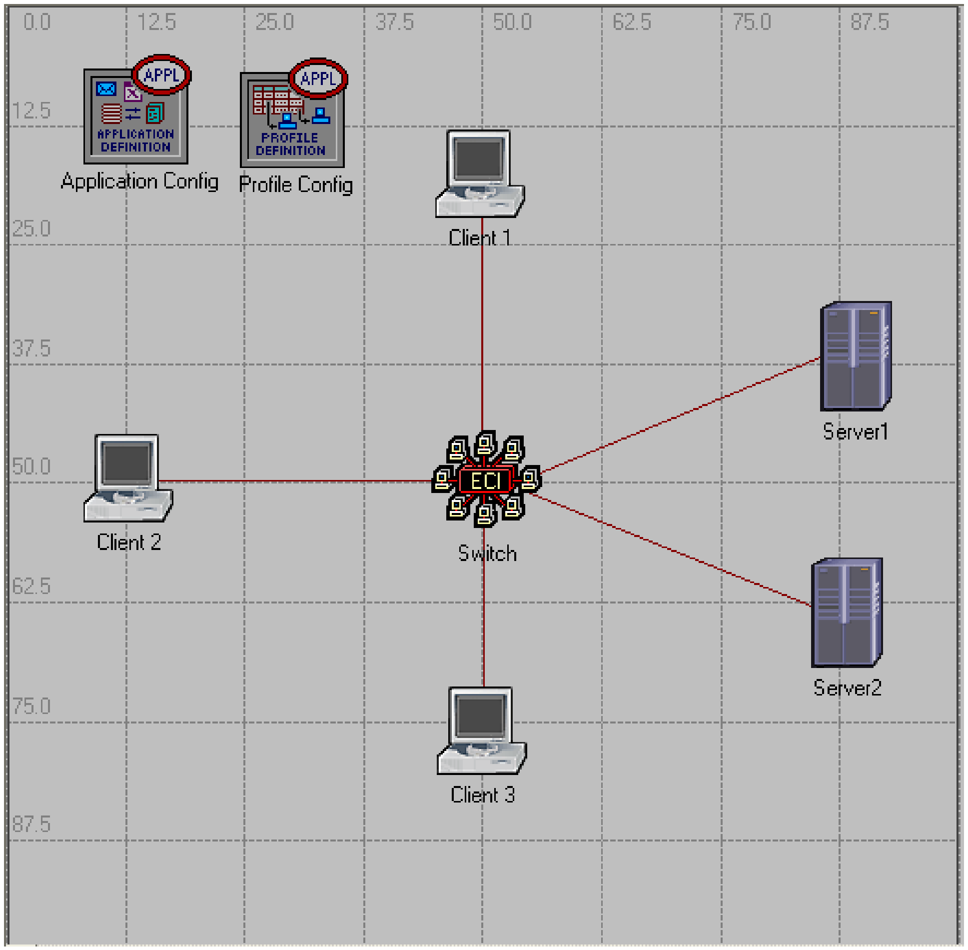
AQM: Active Queue Management
CI: Congestion indication triggered by AQM



CI Marking = frames get marked while forwarding if AQM thresholds exceeded. If between early detection thresholds, use RED algorithm to select frames to mark. If between early drop thresholds, mark more frames but drop some frames. If high drop threshold exceeded, drop frames.

NIC-CI = Extracts L2 Congestion Information and passes on to upper protocols

Simple Topology



All Links are 10 Gbs

Shared Memory 150KB

App = Database Entry over full TCP/IP stack

Workload distribution = Exponential (8000)

ULP Packet Sizes = 1 Bytes to ~85KB

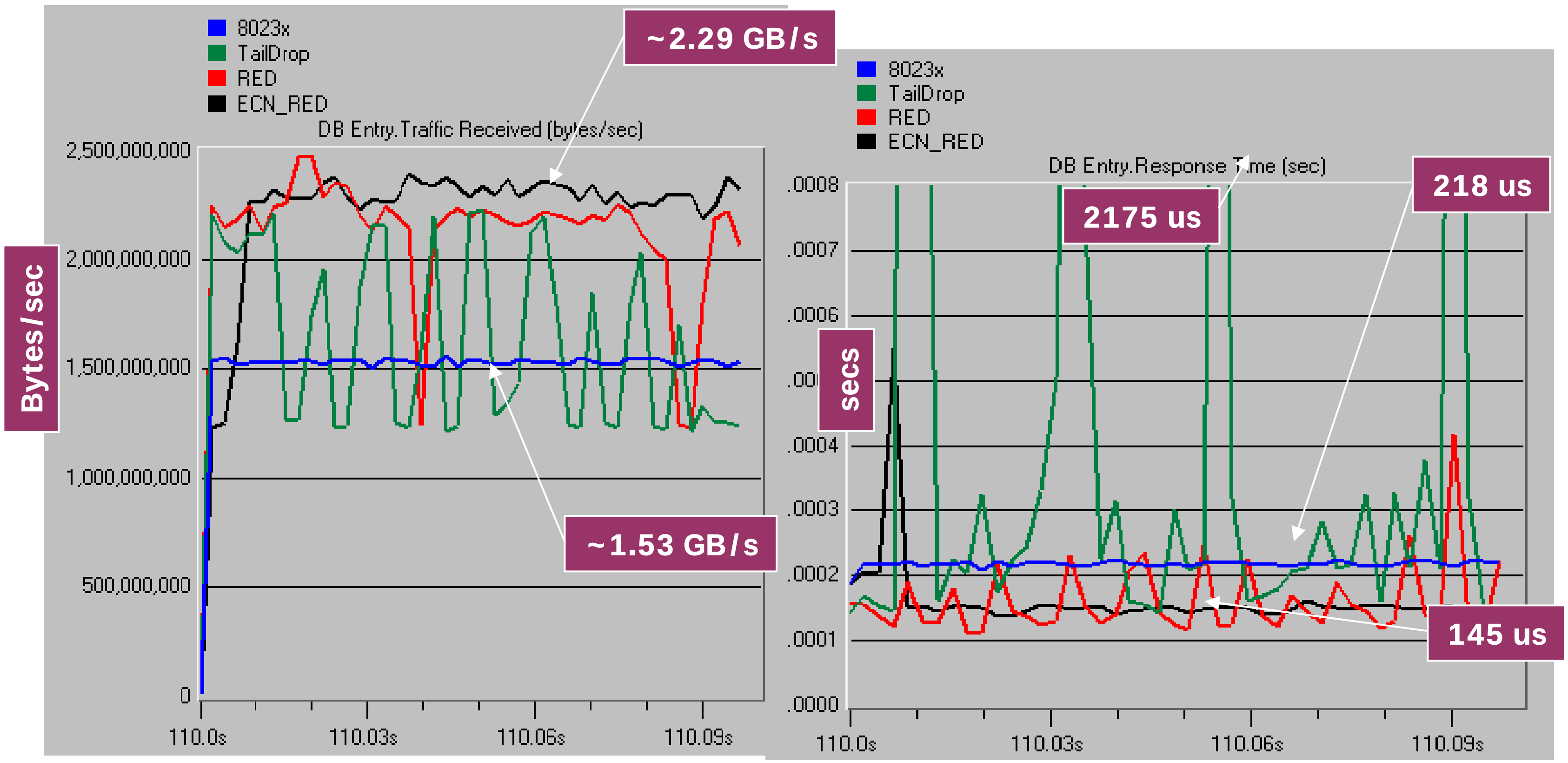
Client 1 sending to both servers

Clients 2 & 3 sending to Server 1

TCP Delay = DB Entry request to completion

HOL Blocking at Client1 for Client1-Server2 traffic

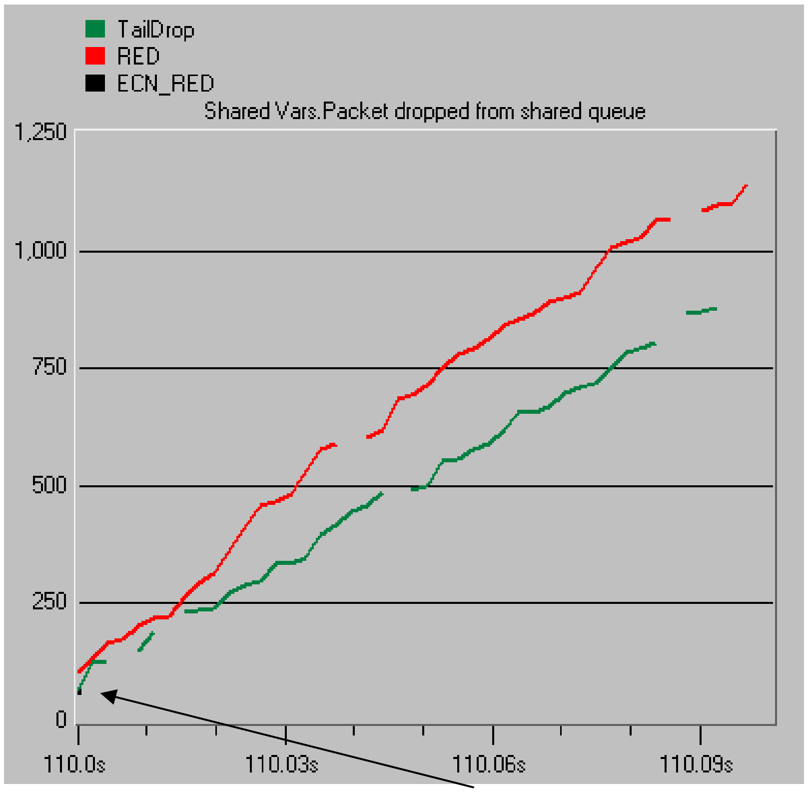
Application Throughput & Response Time

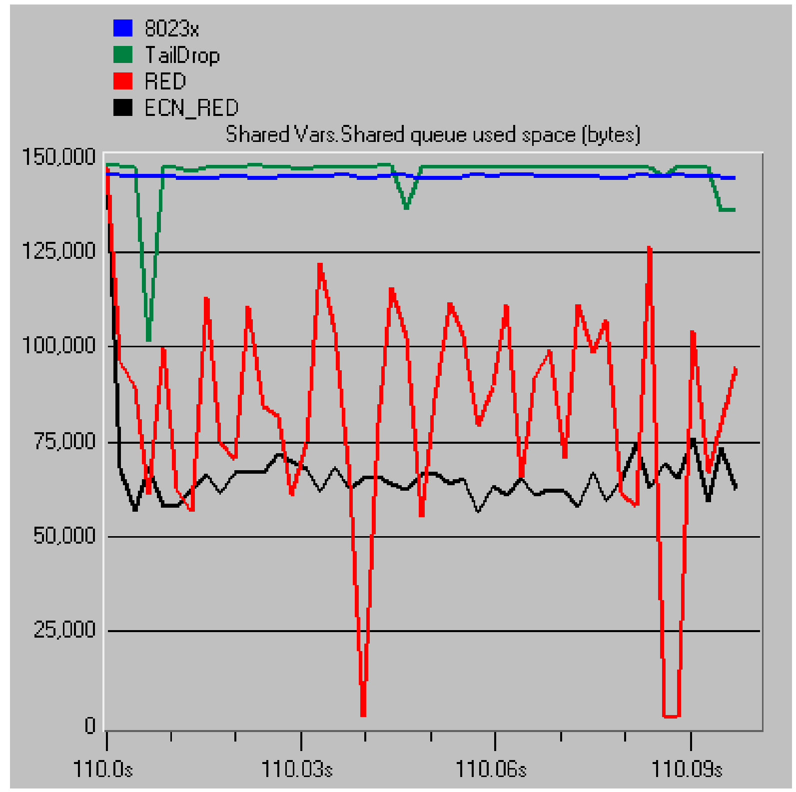


L2-CI with ECN improves TCP Performance

Shared Memory Utilization and Packet Drop at the Switch

Number of drops

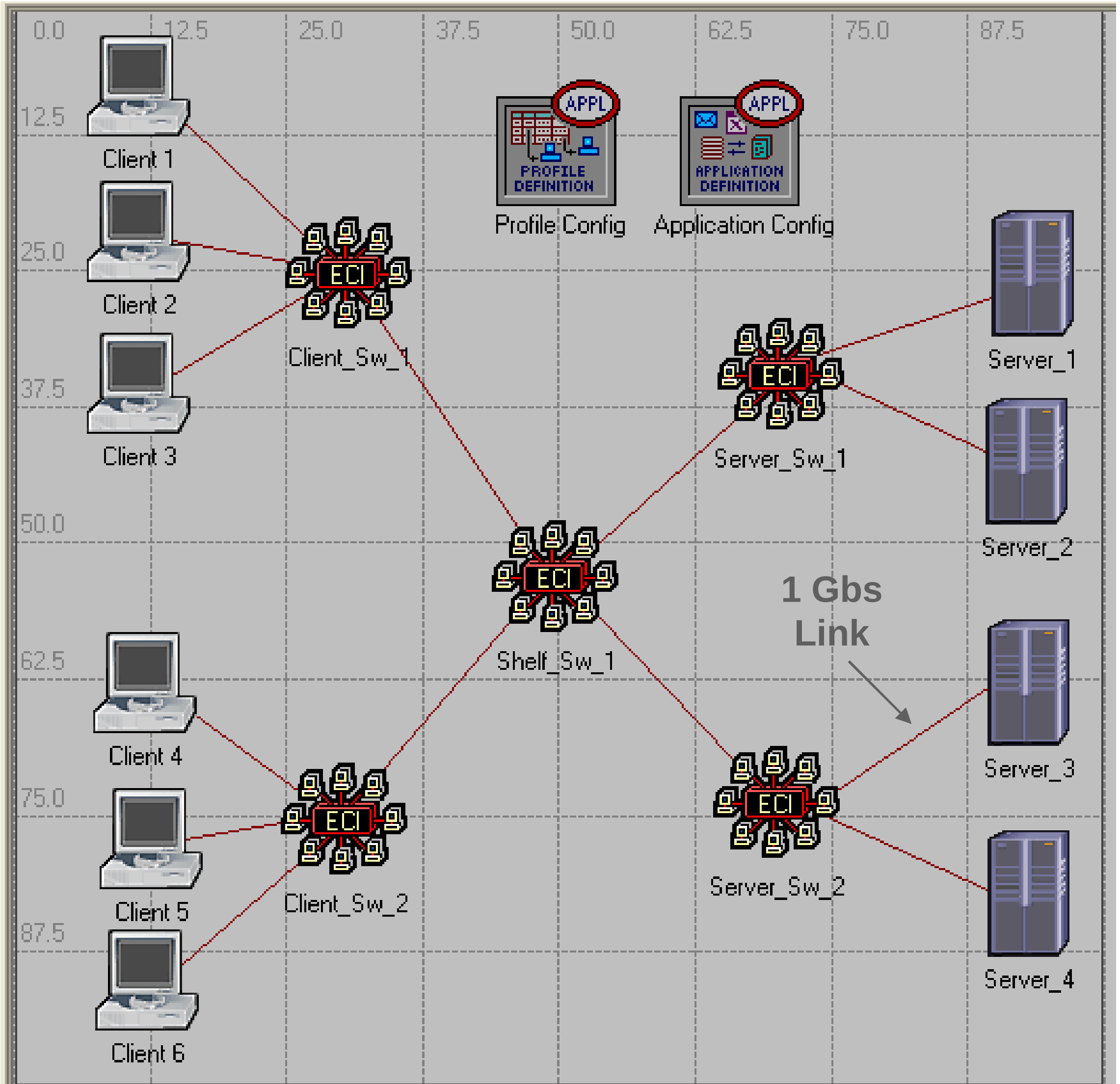




Some initial drops with ECN when it is stabilizing its average Q size

L2-CI can significantly reduce packet drops & reduce buffer requirements

Multi-stage system w/ mixed link speeds



All Links except one are 10 Gbs

Peak Throughput = 2.434 Gigabytes / Sec

App = Database Entry over the full TCP/IP stack

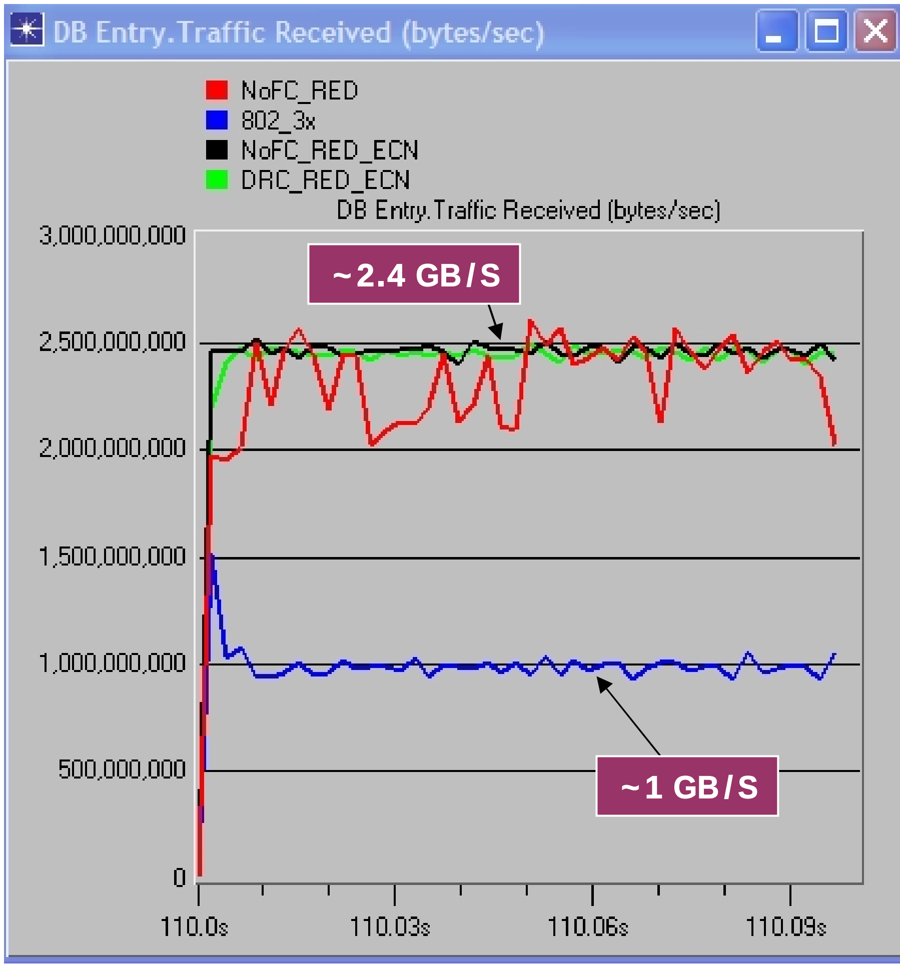
Workload distribution = Exponential (8000)

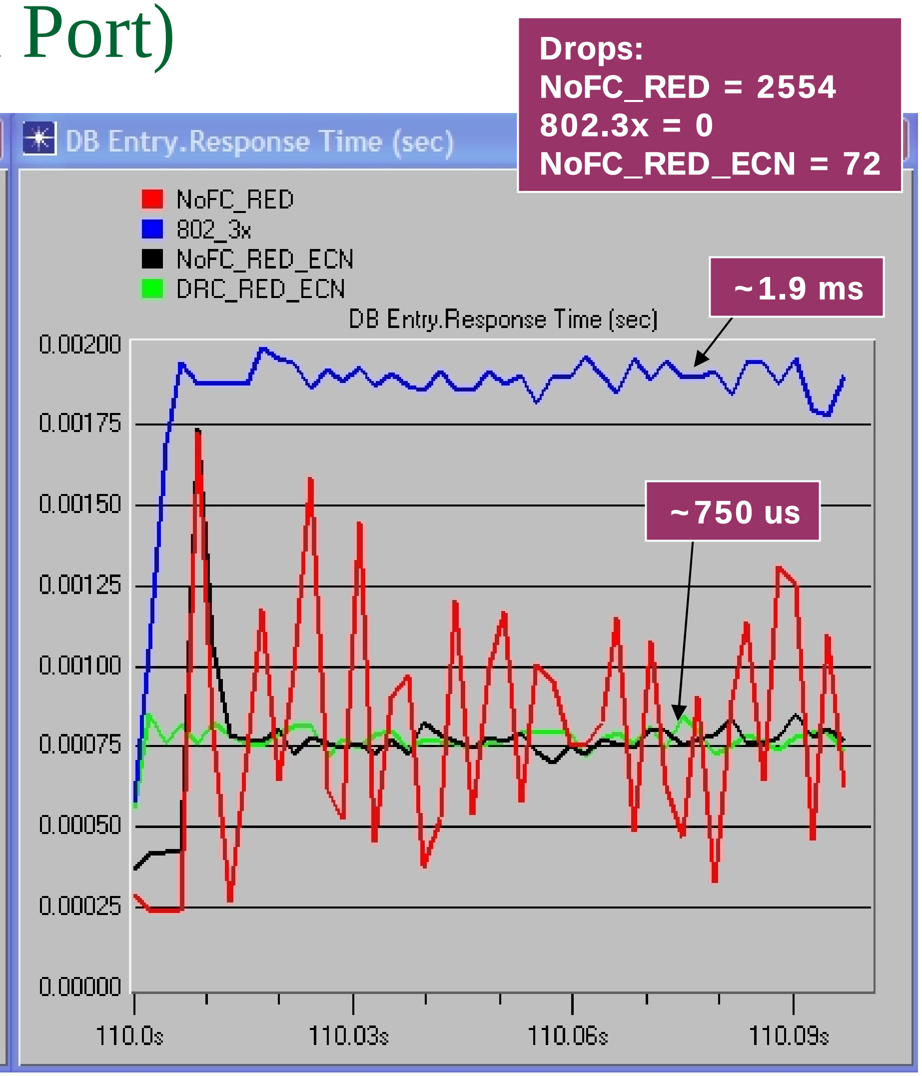
ULP Packet Sizes = 1 Byte to ~85KB

TCP Window size = 64KB

All clients sending database entries to all servers

Application Throughput & Response Time (Buffer = 64 KB per Switch Port)

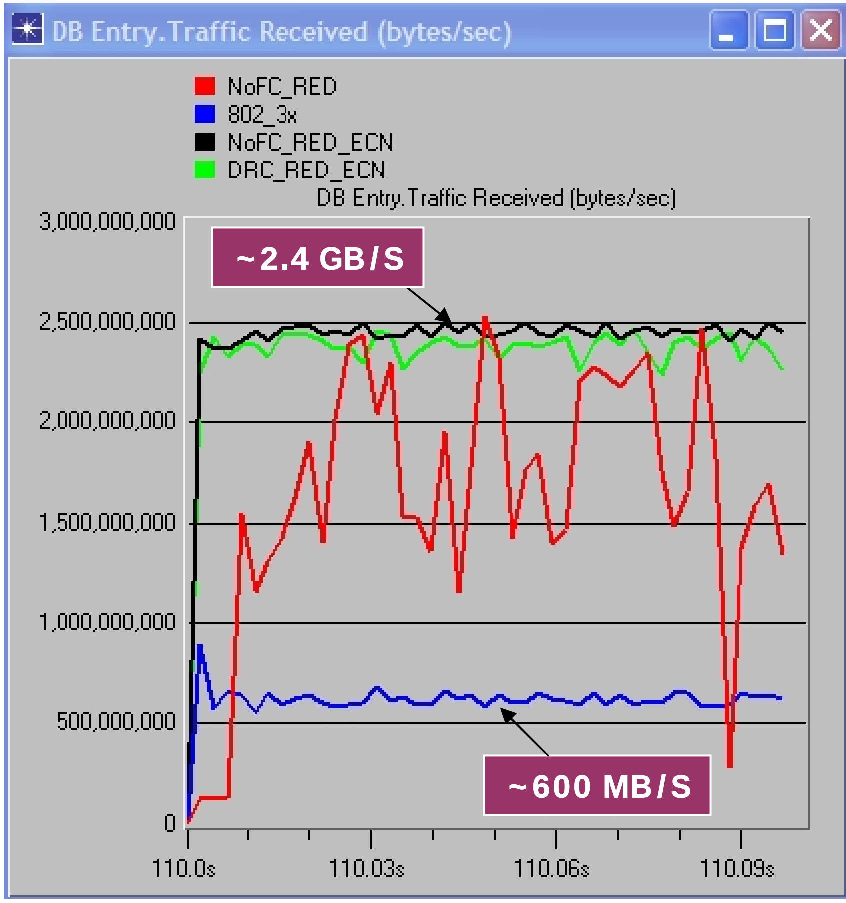


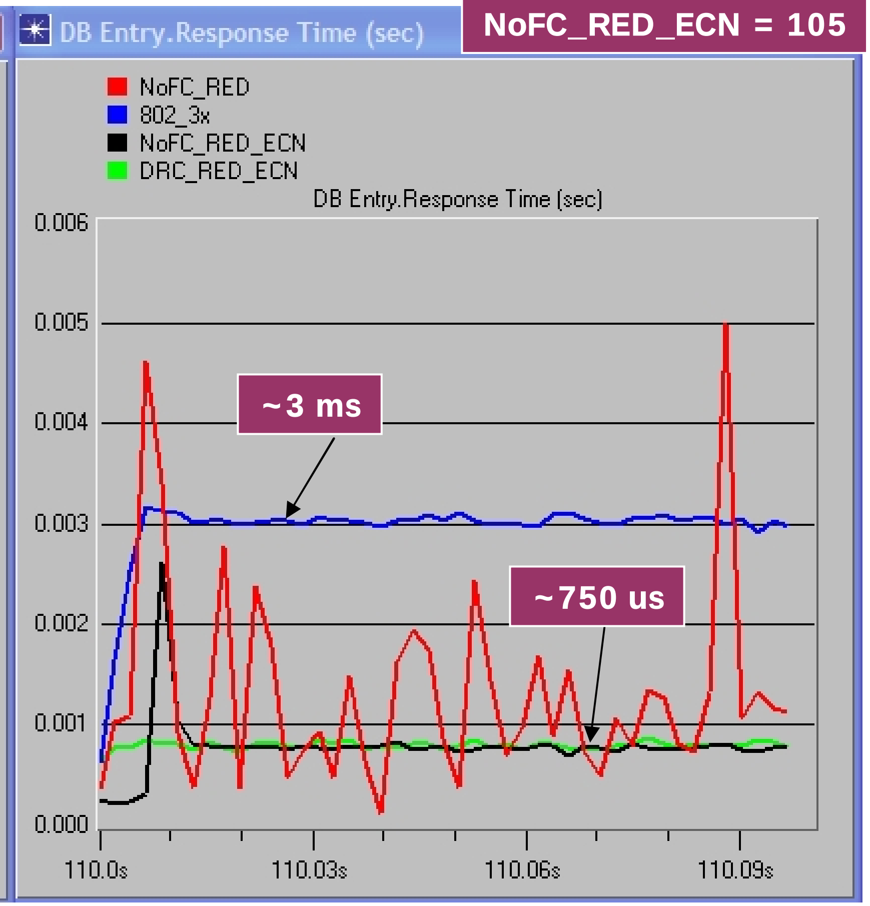


L2-CI/ECN shows excellent characteristic for short range TCP.

Application Throughput & Response Time (Buffer = 32 KB per Switch Port)

Drops:
NoFC_RED = 2373
802.3x = 0
NoFC_RED_ECN = 105





L2-CI/ECN maintains performance even with small switch buffers

Simulation Summary

- Example presented show “Technical Feasibility” of Congestion Management in Ethernet
- Can allow MAC Clients to take proactive actions based on Congestion Information
- Facilitate & take advantage of higher layer CM mechanisms
- Simulations show significant comparative improvements

Wrap-Up and Q&A