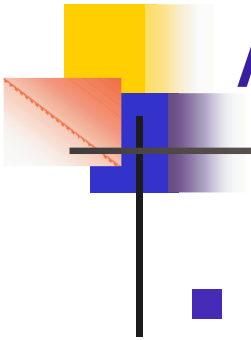
Rate Control in Short Range 802.3 Interconnects

Manoj Wadekar
Gary McAlpine
Tanmay Gupta

September 27, 2004

Intel



Agenda

- Goals and non-goals for presentation
- Known Problems
- Directional Rate Control
- L2-Congestion Indication
- Summary

Goals and non-goals

Goals

- Demonstrate "technical feasibility" of Congestion Management in Ethernet
- Demonstration of comparative advantages over existing options
- Trigger interest and discussion

Non-goals

- Drive specific solution proposals

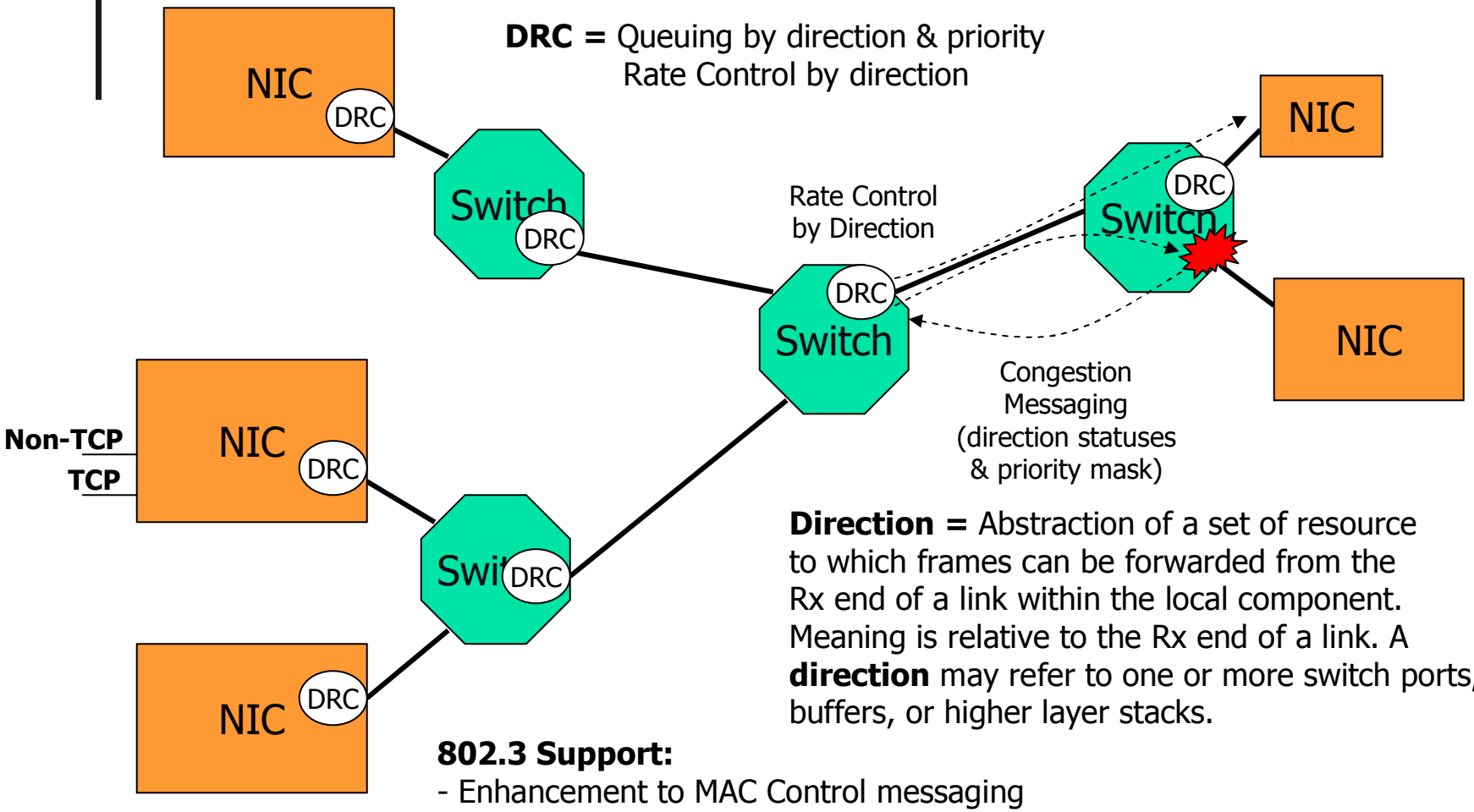
Known issues

- HOL Blocking at Tx ports
  - Biggest problem with 802.3x
  - RED reduces blocking with random drops
- TCP Rate Adaptation is "reactive" not "proactive"
  - Needs packet drop to "guess" congestion in L2 networks
- Frame drops spoil zero copy operation
  - Forces out-of-order processing & copies in ULP stacks
  - Affects iSCSI, RDMA, low latency sockets ... etc.
- Very large buffers not practical in backplane
- Coarse granularity of TCP Timers
- Mixed traffic types
  - Some sensitive to latency, latency variation, loss
  - Not all traffic is TCP

Improve Link Efficiency

- Issue: HOL blocking at Tx ports.
  - Transient congestion
- Method: Directional Queuing & Rate Control
  - Local to each link
  - Dynamically reacts to congestion at Tx ports
  - Optimizes traffic flow during congestion
  - Avoids local blocking due to congestion
  - Rate controls individual directions of flow

Direction Rate Control:
Reduce HOL Blocking at TX ports

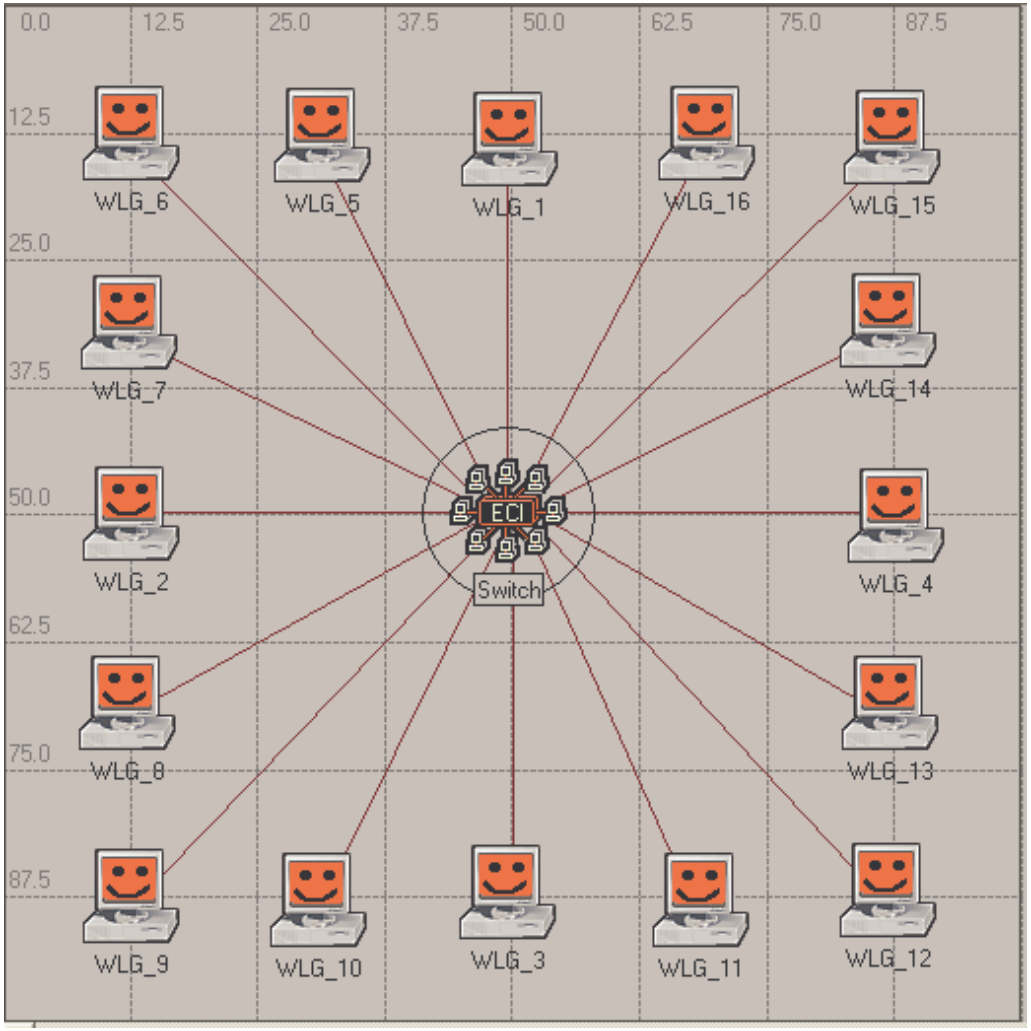


DRC = Queuing by direction & priority
Rate Control by direction

Direction = Abstraction of a set of resource to which frames can be forwarded from the Rx end of a link within the local component. Meaning is relative to the Rx end of a link. A **direction** may refer to one or more switch ports, buffers, or higher layer stacks.

802.3 Support:
- Enhancement to MAC Control messaging
- Enhancement to MAC Client Service Interface

Backplane Switch Test Model



All Links are 10 Gbs

L2-only workloads

Peak Throughput = 155.64 Gbs above L2

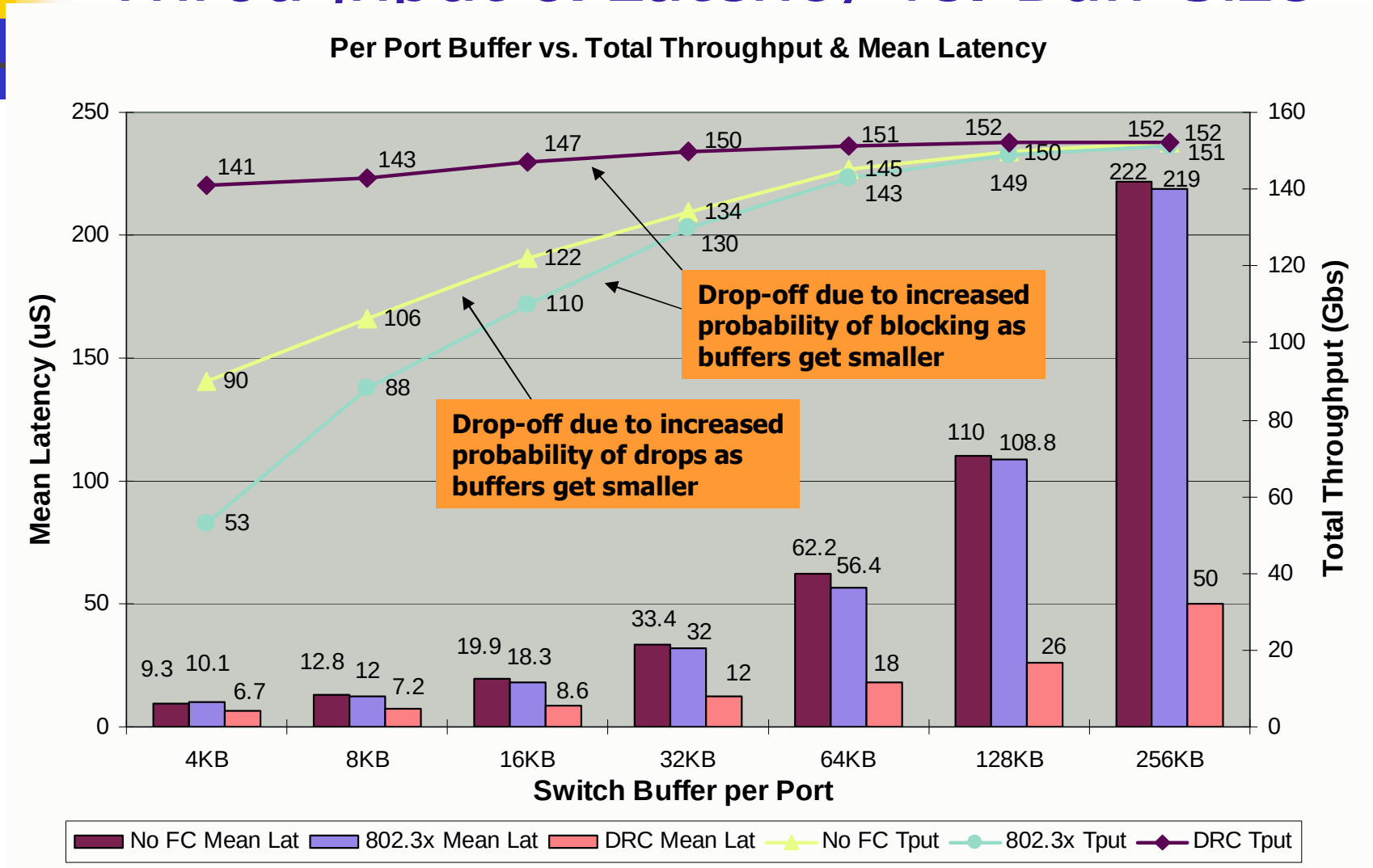
Workload distribution = Exponential (8000)

ULP Packet Sizes = 46 Bytes to ~64KB

All WLGs sending to all others randomly as fast as possible

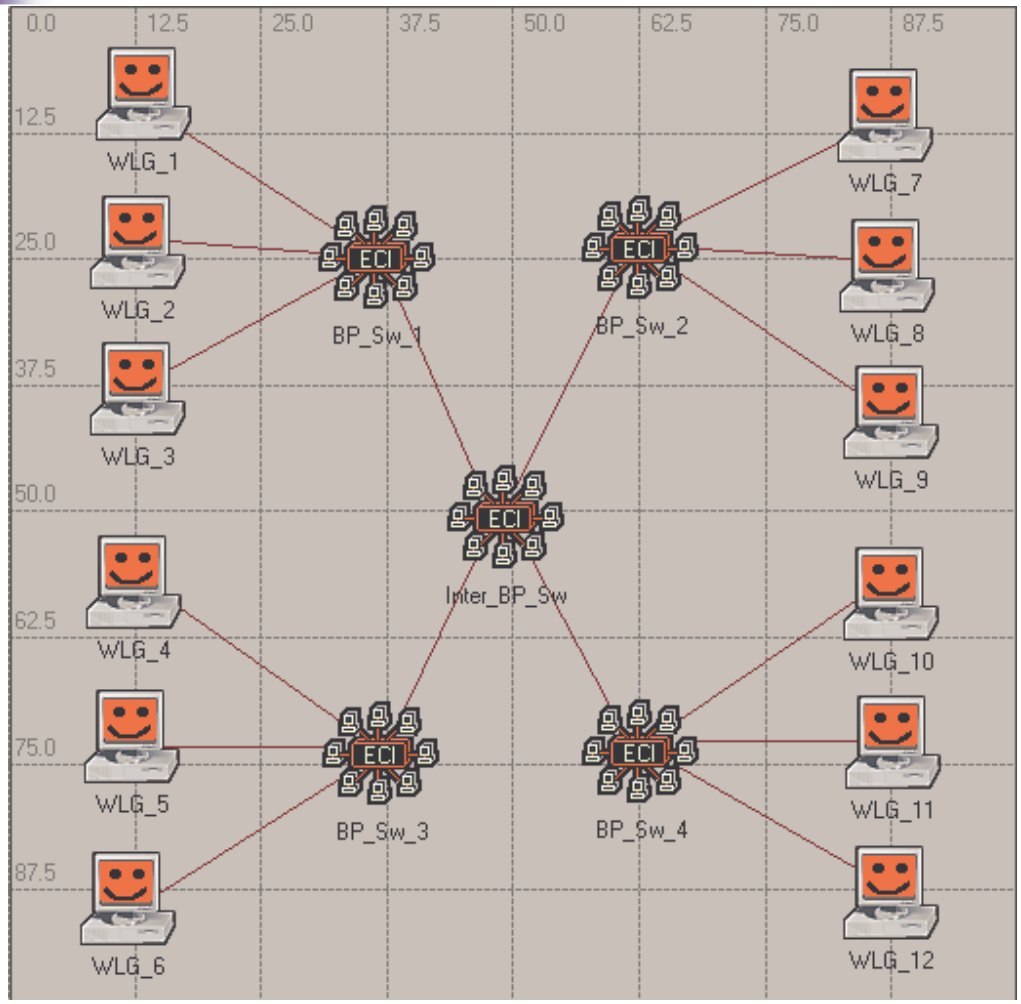
Latency is per frame = 1st byte of frame from source memory to last byte in dest memory.

Throughput & Latency vs. Buff Size



Directional Mechanisms reduce HOL Blocking at Tx Ports

Multi-stage L2 Test Model



All Links are 10 Gbs

L2-only workloads

Peak Throughput = 116.73 Gbs above L2

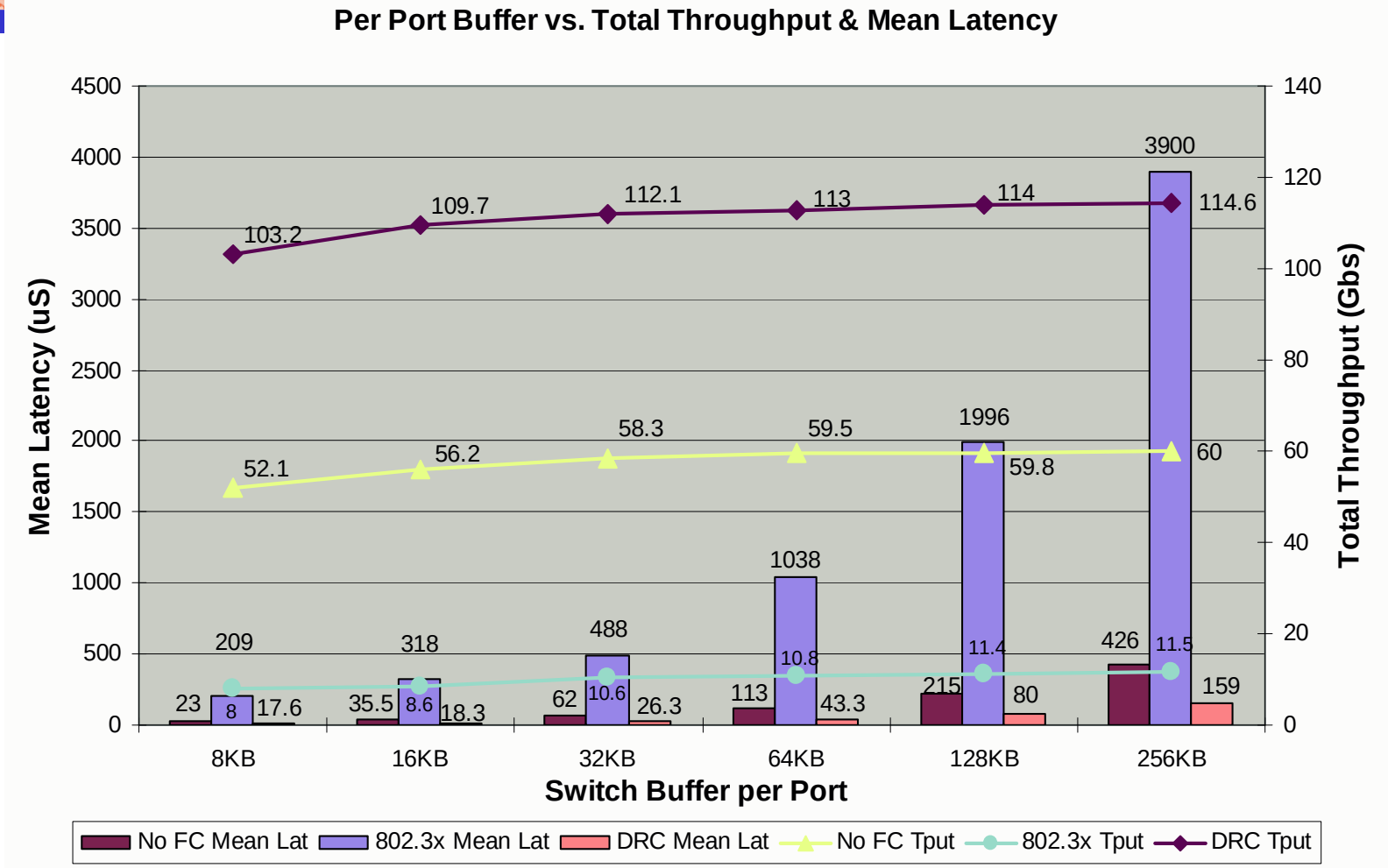
Workload distribution = Exponential (8000)

ULP Packet Sizes = 46 Bytes to ~64KB

All WLGs sending to all others randomly as fast as possible

Latency is per frame = 1st byte of frame from source memory to last byte in dest memory.

Throughput & Latency vs. Buffer Size



Directional Mechanisms also show benefit in Multi-stage Topologies

L2– Congestion Indication

Issue:

- Congestion due to oversubscription
- "Reactive" rate control in TCP

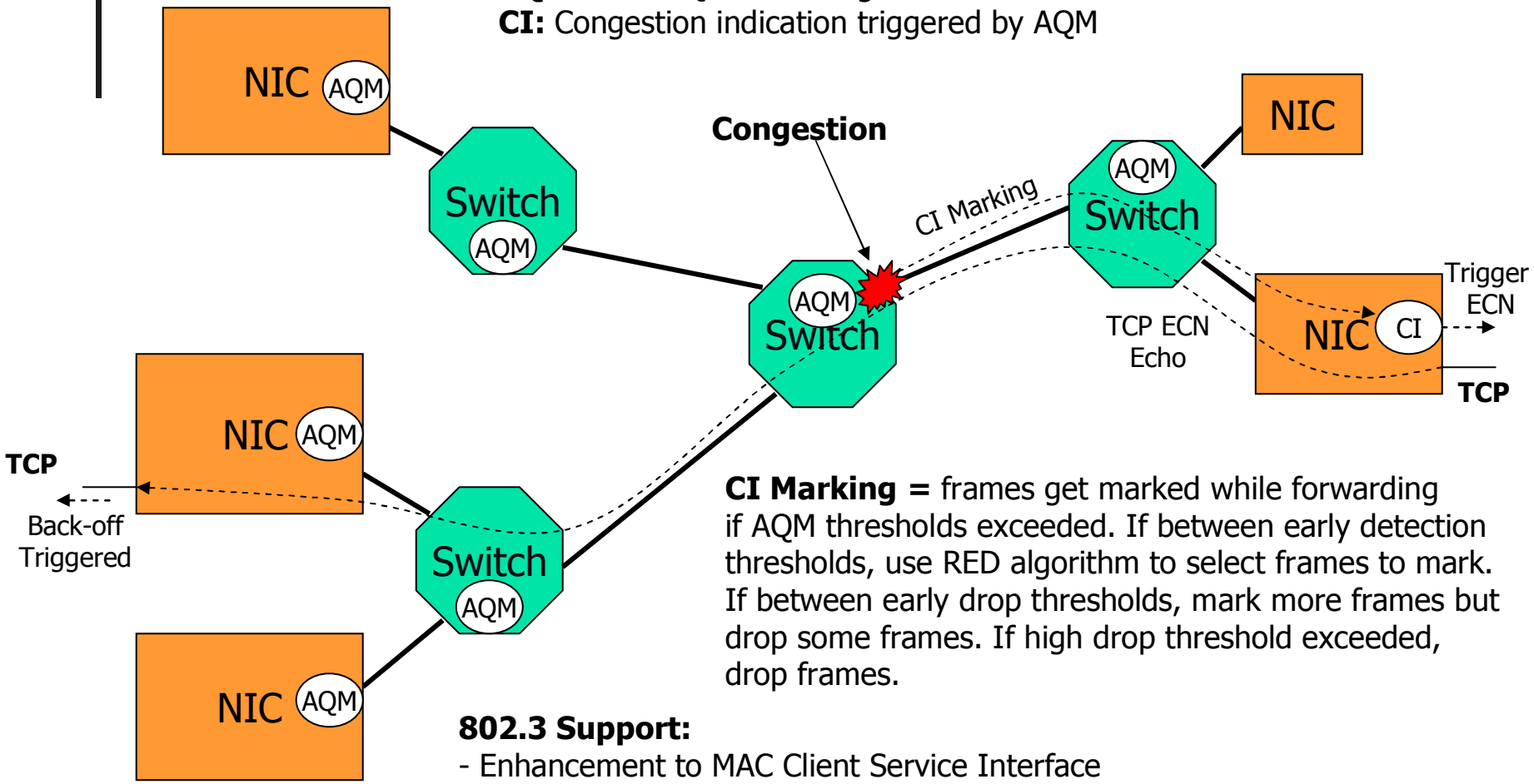
Method:

- "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
- 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

Intel Corp.

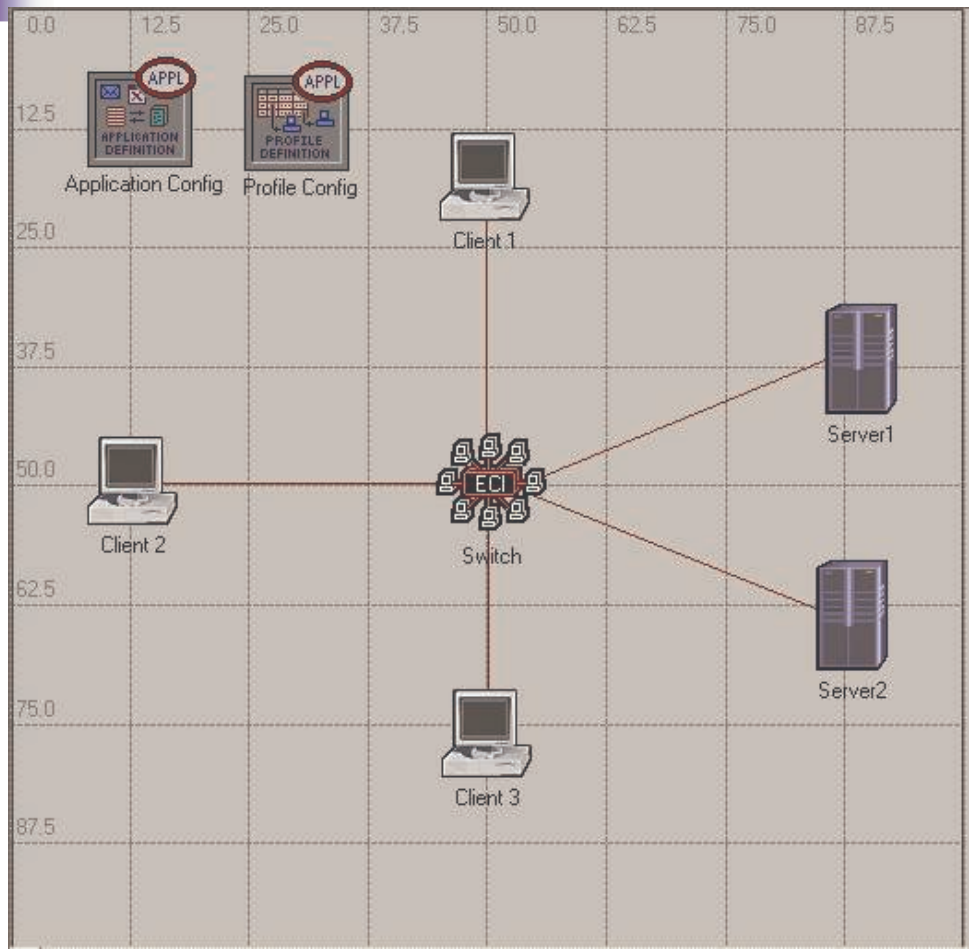
AQM: Active Queue Management
CI: Congestion indication triggered by AQM



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.

802.3 Support:
- Enhancement to MAC Client Service Interface
- If messaging used, Enhancement to MAC Control messaging

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

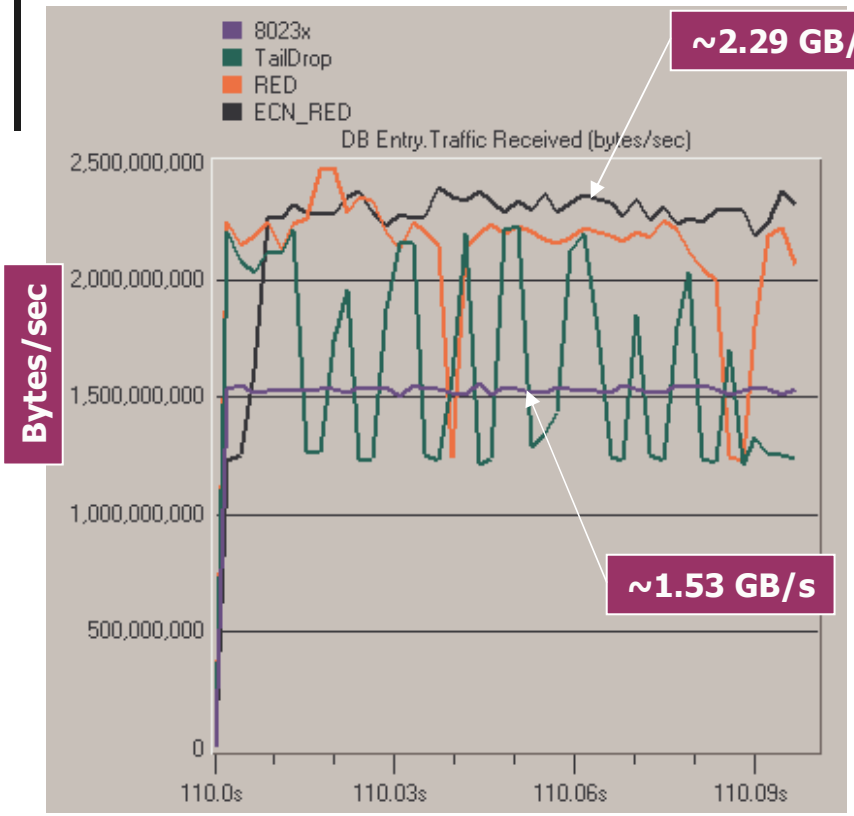
Client 1 sending to both servers

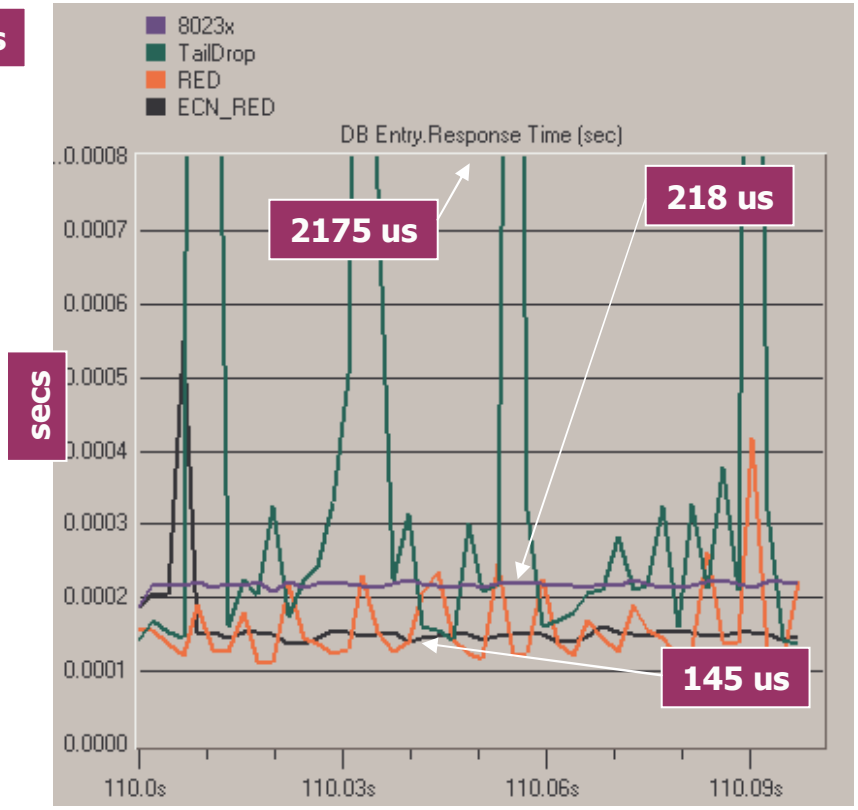
Clients 2 & 3 sending to Server 1

TCP Delay = DB Entry request to completion

HOL Blocking at Client1 for Client1-Server2 traffic

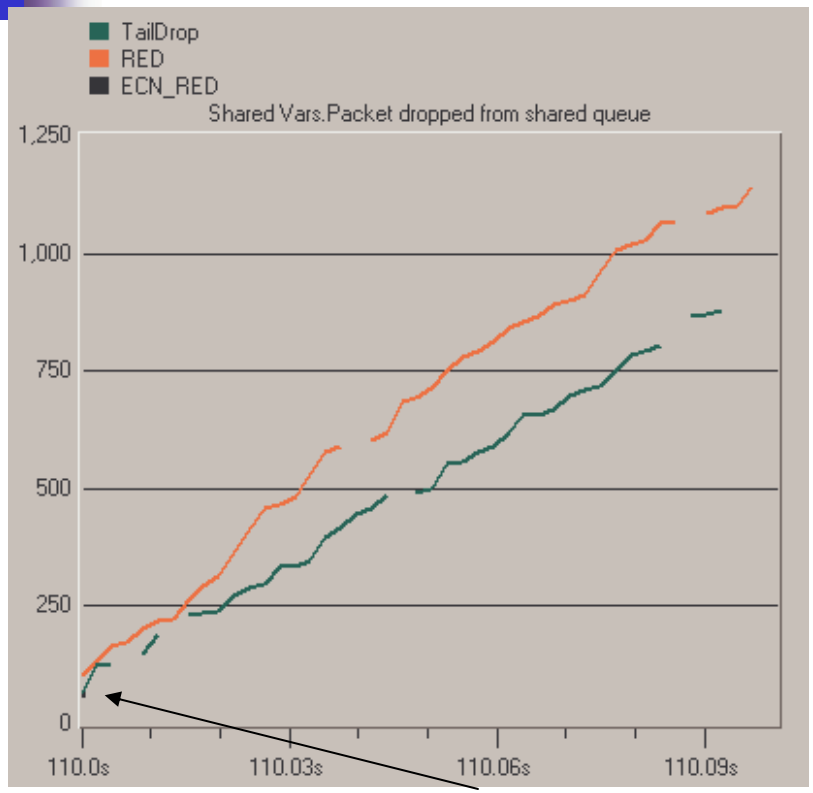
Application Throughput & Response Time



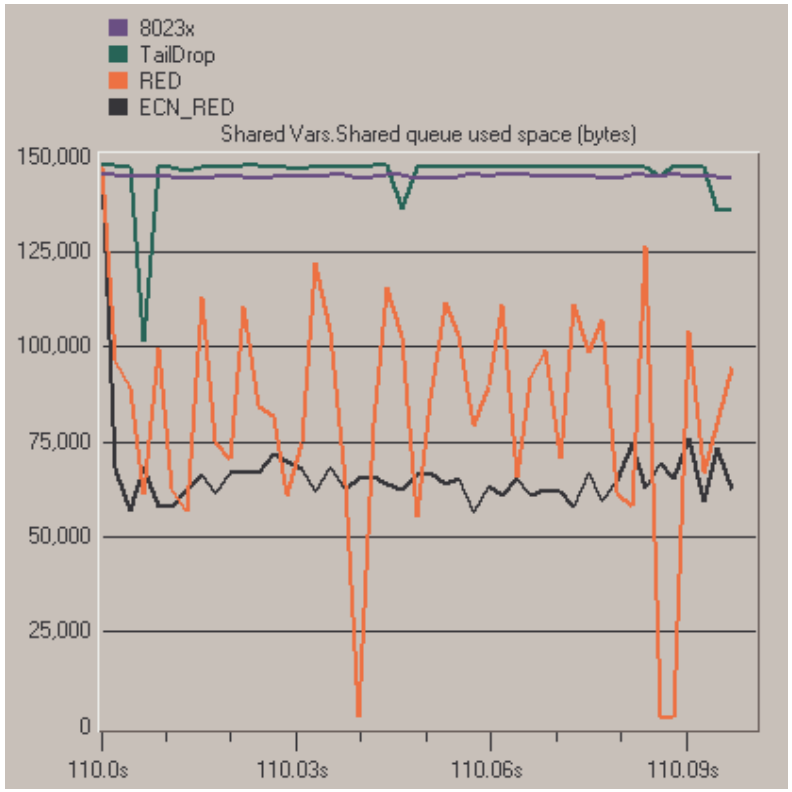


L2-CI with ECN improves TCP Performance

Shared Memory Utilization and Packet Drop at the Switch



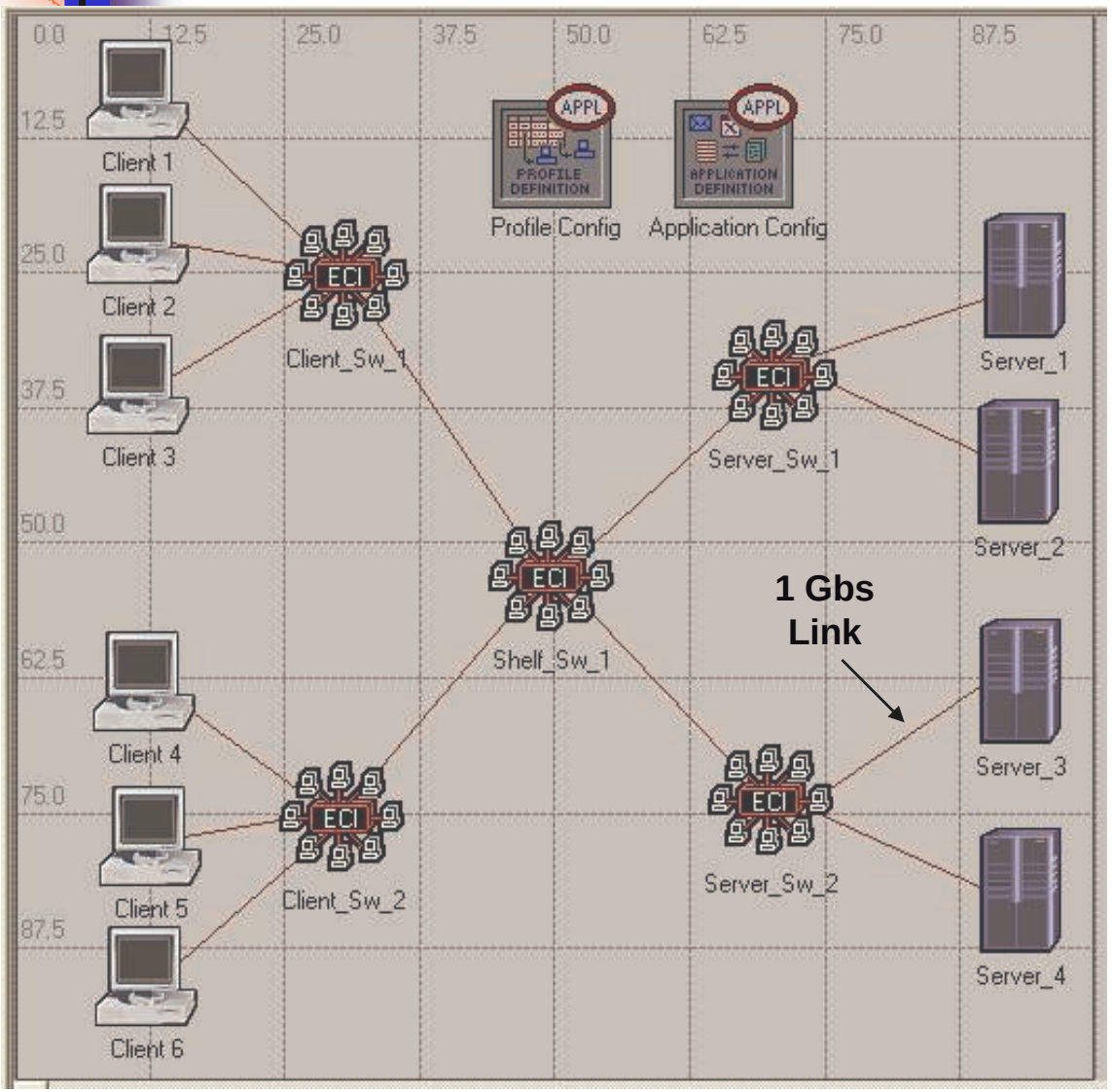
Number of drops



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

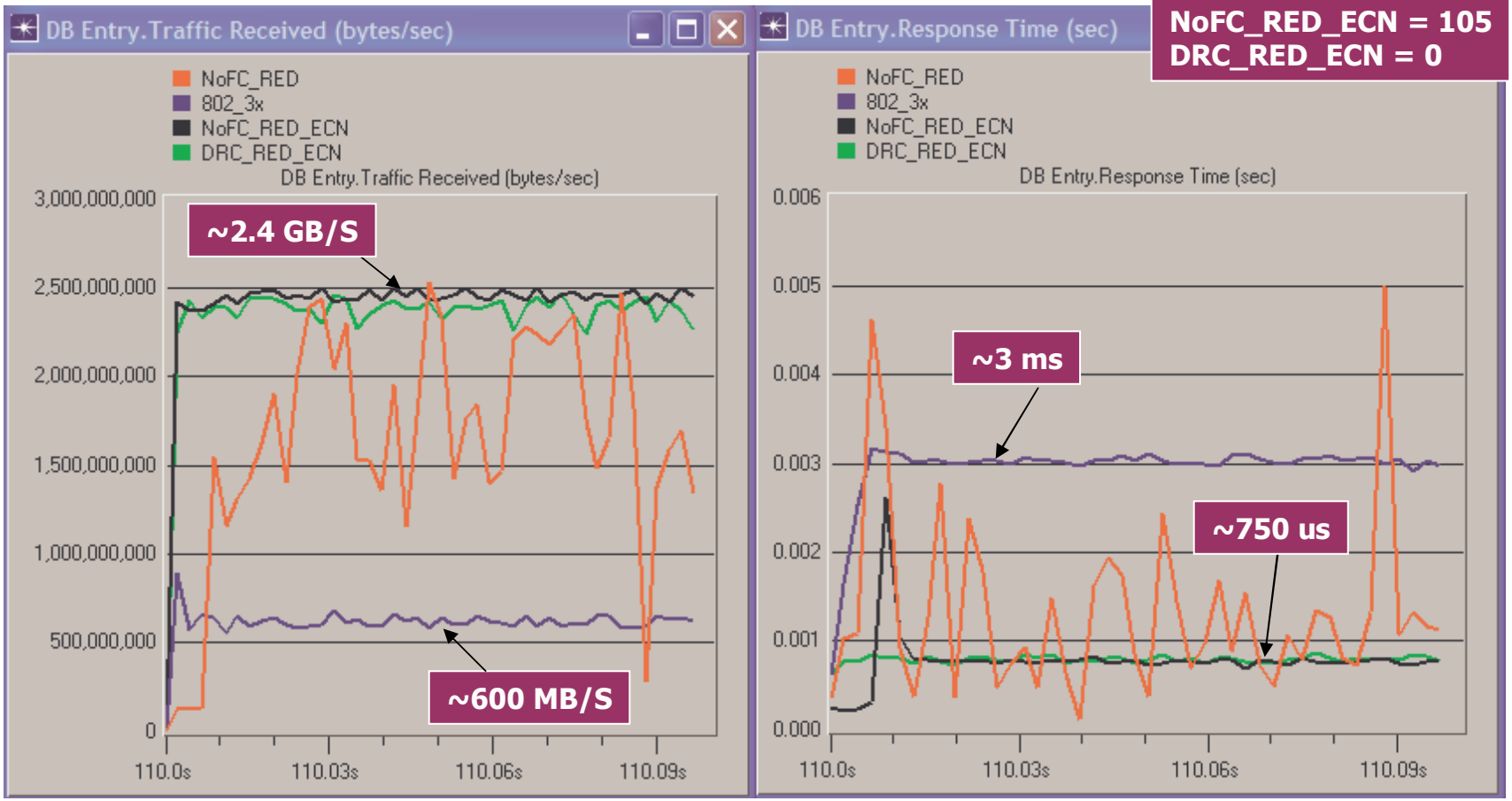
Application Throughput & Response Time (Buffer = 64 KB per Switch Port)

Drops:
NoFC_RED = 2554
802.3x = 0
NoFC_RED_ECN = 72
DRC_RED_ECN = 0

L2-CI/ECN shows excellent characteristic for short range TCP.
Addition of DRC eliminates drops and response spikes.

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

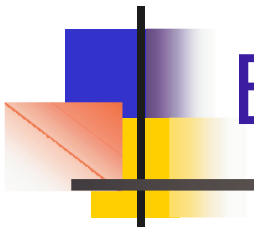
Drops:
NoFC_RED = 2373
802.3x = 0
NoFC_RED_ECN = 105
DRC_RED_ECN = 0



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

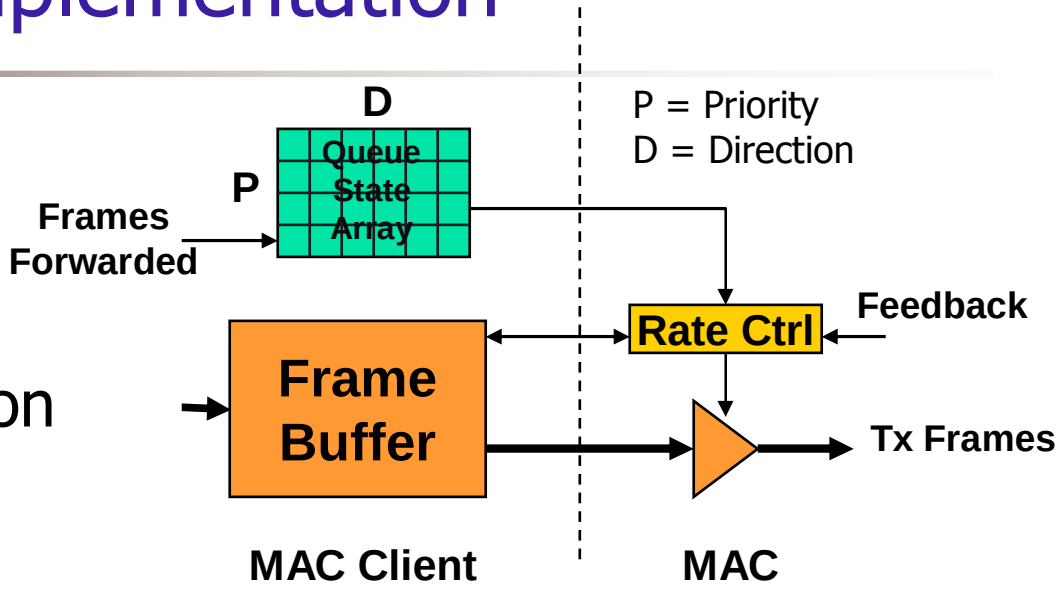
Summary

- Examples presented show "technical feasibility" of Congestion Management in Ethernet
- Can allow MAC Clients to take proactive actions based on congestion information via 802.3
- Facilitate & take advantage of higher layer CM mechanisms
- Simulations show significant comparative improvements



Backup

DRC Model Implementation



- Link Peer supplies:
  - Directional congestion feedback
    - Severity indicators
    - 2 bits per direction
  - Priority Mask
    - Rate control enables
    - 1 bit per priority
- Rate Control:
  - Controls rate of Tx from individual Queues
  - Determines order of Tx from Frame Buffer

Feedback Msg

MAC Ctrl Addr
Source Addr
DRC Type
Direction Statuses
Priority Mask
Padding
FCS