Throughput & Latency Control in Ethernet Backplane Interconnects

Manoj Wadekar

Gary McAlpine

Intel

Date 3/16/04

Agenda

- Discuss Backplane challenges to Ethernet
- Simulation environment and definitions
- Preliminary Simulation results
- Illustrate current solutions/tradeoffs
- Summary

Ethernet : Backplane Concerns

- Frame Discard in Response to Congestion
  - TCP timeouts & retrans = big performance hits
    - Cluster traffic prone to frequent congestion events
  - Unacceptable in Cluster/Backplane Interconnect

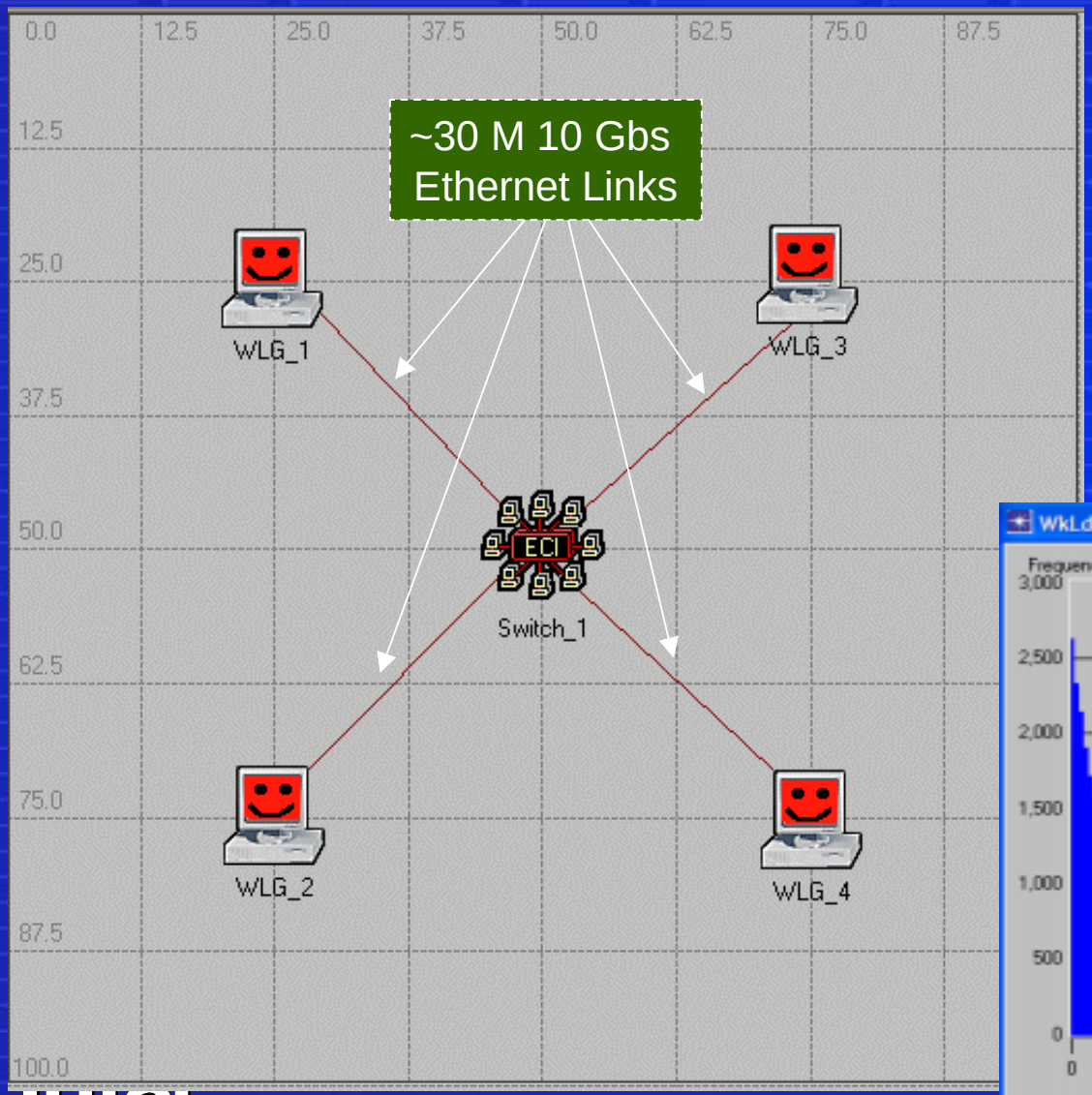
- LAN latencies (loaded) can be in milliseconds
  - Voice, video, real time apps sensitive to delay, delay var.
    - Additive over many hops
  - Servers targeting IPC with mean latency < 10 uS – loaded

- Large switch buffers
  - Smaller buffers enable tradeoff between smaller ports, cost
  - Backplane/Cluster networks – long links not required

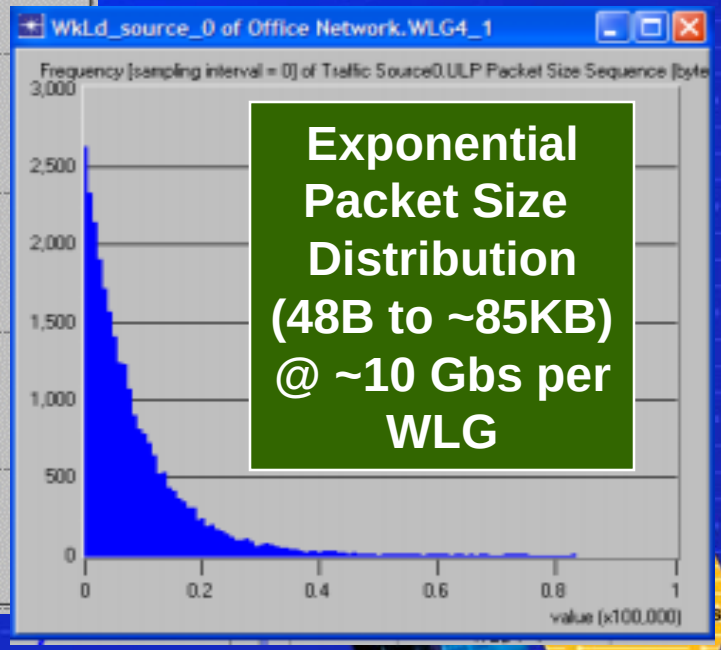
The Challenges
(The Evil Buts ...)

- 802.3x Flow Control may be available, but ...
  - Seldom used
  - Causes blocking, reducing throughput
- Frame discard avoids blocking, but ...
  - Causes timeouts & retransmissions
  - Requires large buffers to minimize discard frequency
- Large buffers support long links, high throughput, but ...
  - Allows large latencies & higher cost
  - Requires prioritization to enables low latency
- Prioritization can enable low latency on critical traffic, but
  - Not supported by most Apps

Simulation Environment



- 16 x 10Gbs Switch
  - 1.5 M Shared Buff
  - 10 Gbs 802.3 MACs
- 4 X 10Gbs Workload Gens
  - L2 WLG Source
  - NIC w/ 2KB Buff
  - MAC w/ 2KB Buff
  - WLG Sink
- Bursty Workload



Definitions

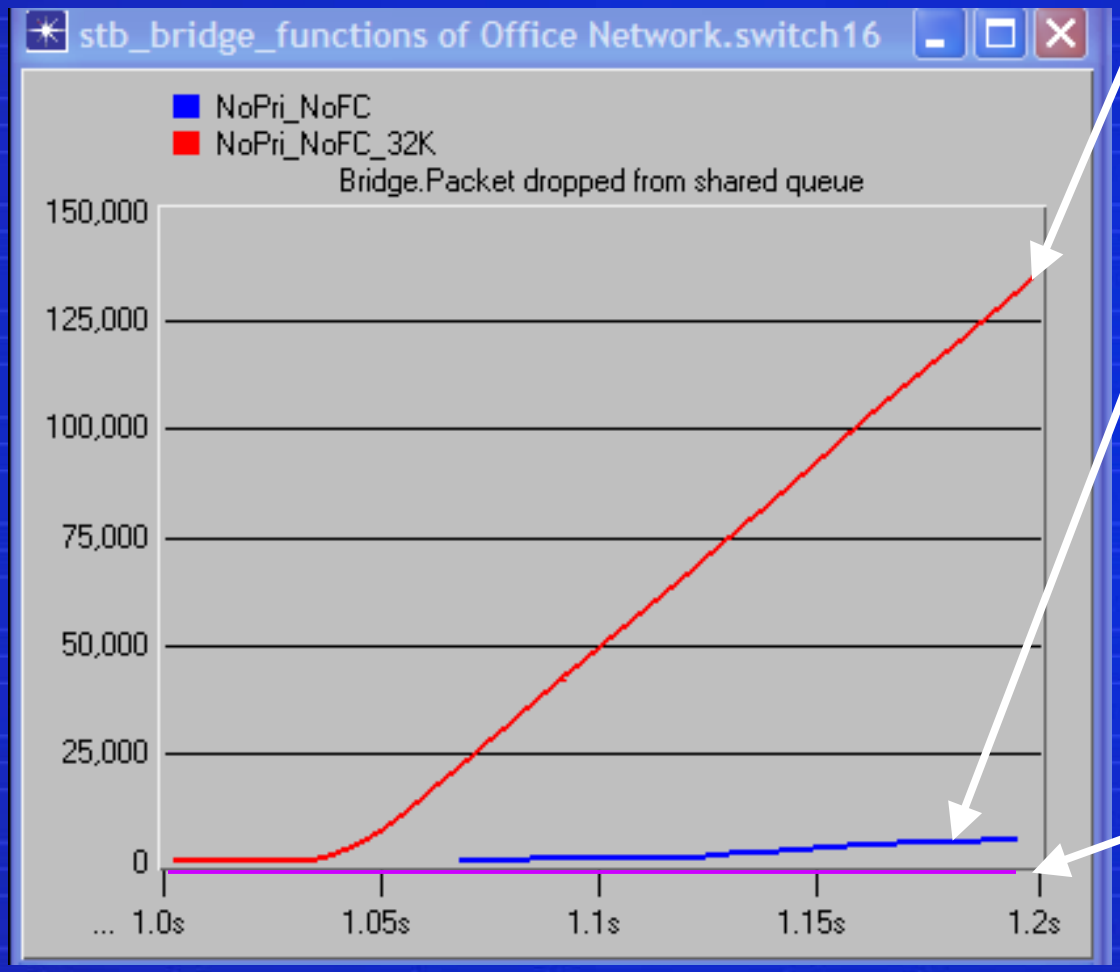
- Throughput
  - Aggregate Traffic (bps) received by all End Stations during test period
- Latency
  - End-to-End delay is measured per frame: First byte from Source App to last byte at Sink App
- Shared Memory Utilization
  - Maximum Memory utilization at switch during test
- Flow Control Thresholds
  - Memory High threshold at which switch starts sending XOFF
  - Memory Low Threshold at which switch sends XON
- Frames Discarded
  - Packets dropped at switch due to congestion

Modeling Scenarios

Throughput, latency and packet drop for:

- 1.5M Shared Memory w/o flow control
- 32K Shared Memory w/o flow control
- 32K Shared Memory with 802.3x Flow Control (Hi-Threshold = 16K)
- 1.5M Shared Memory with various Flow Control thresholds for 802.3x

Frame Discard



~135K Frames Discarded w/ 32 KB

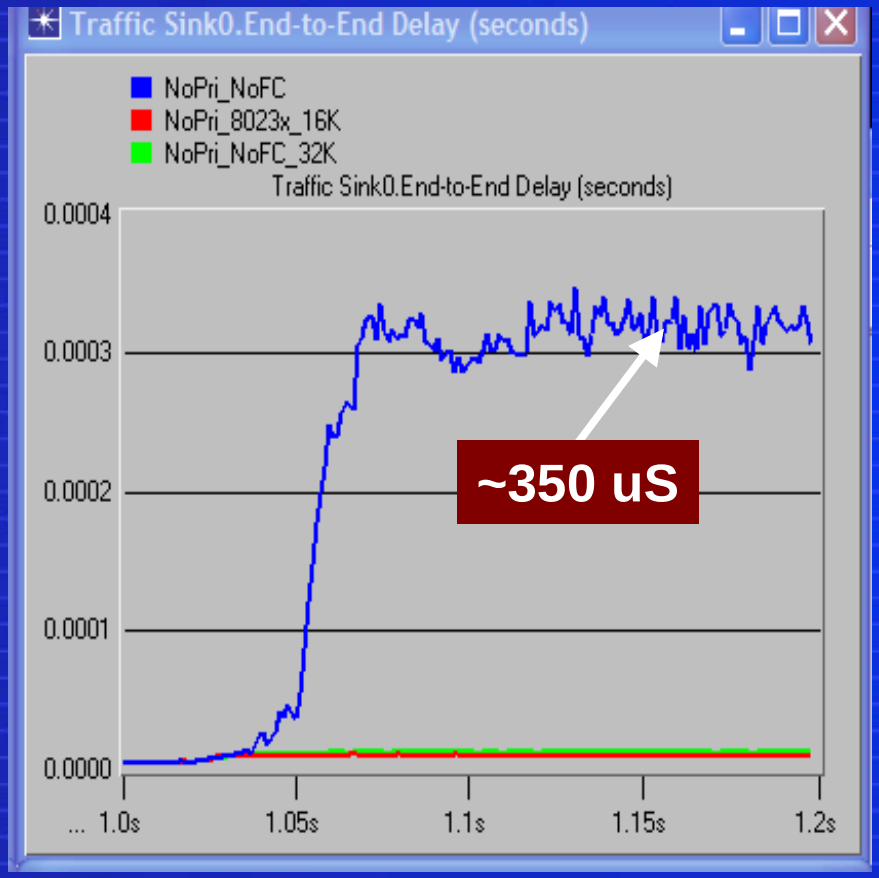
~5K Frames Discarded w/ 1.5 MB

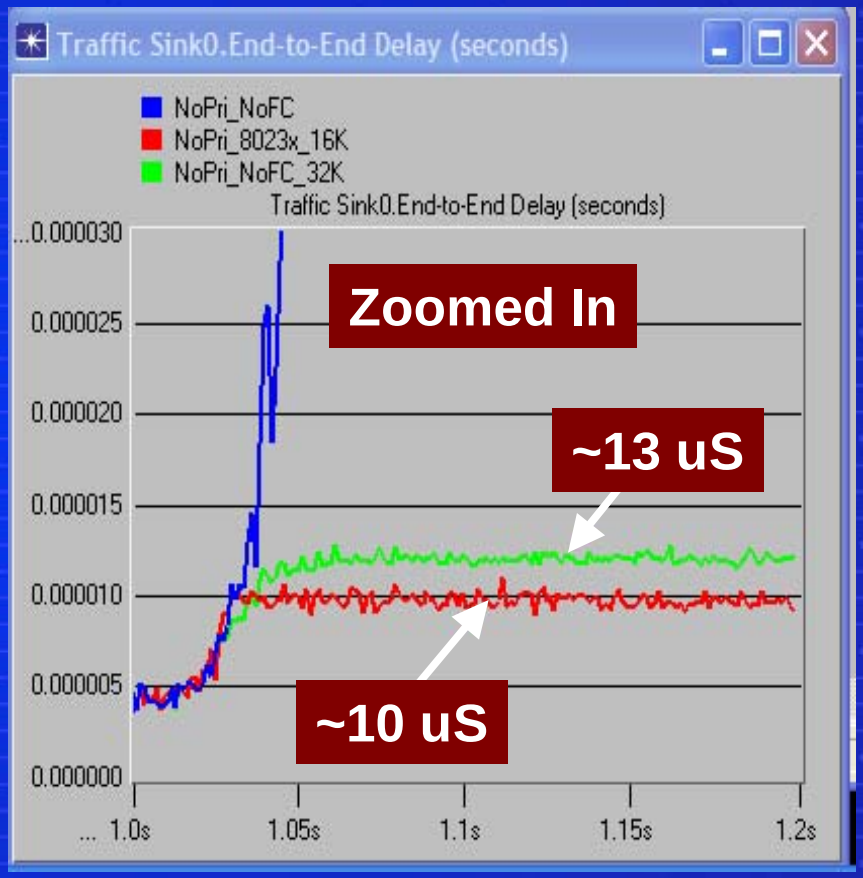
0 Frames Discarded w/ 802.3x

MAC Ctrl Frame Rate = ~165K F/S (One Link)

Frame discard can be completely eliminated with Flow Control even for smaller shared memory

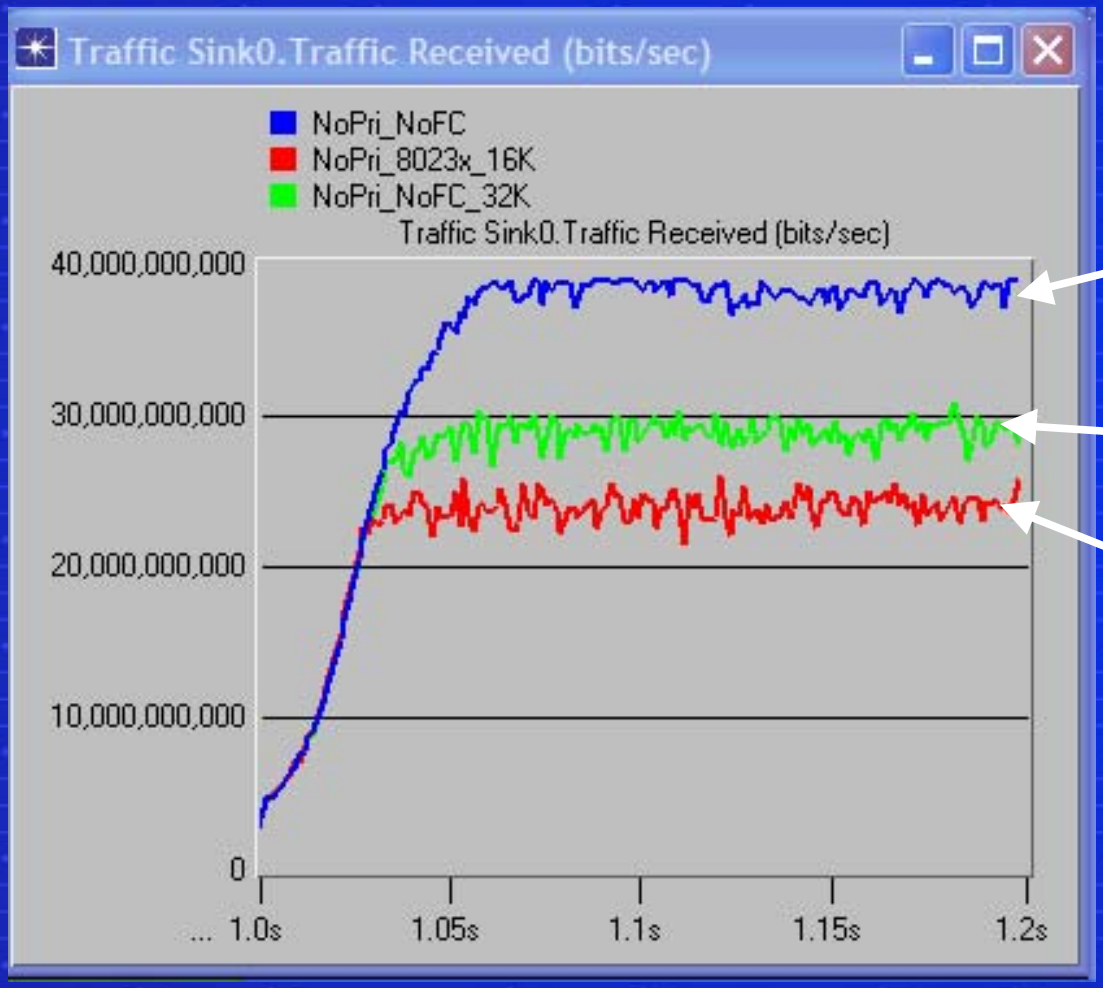
Latency





Latency can be constrained through use of Flow Control

Throughput



~38 Gbs (Max)

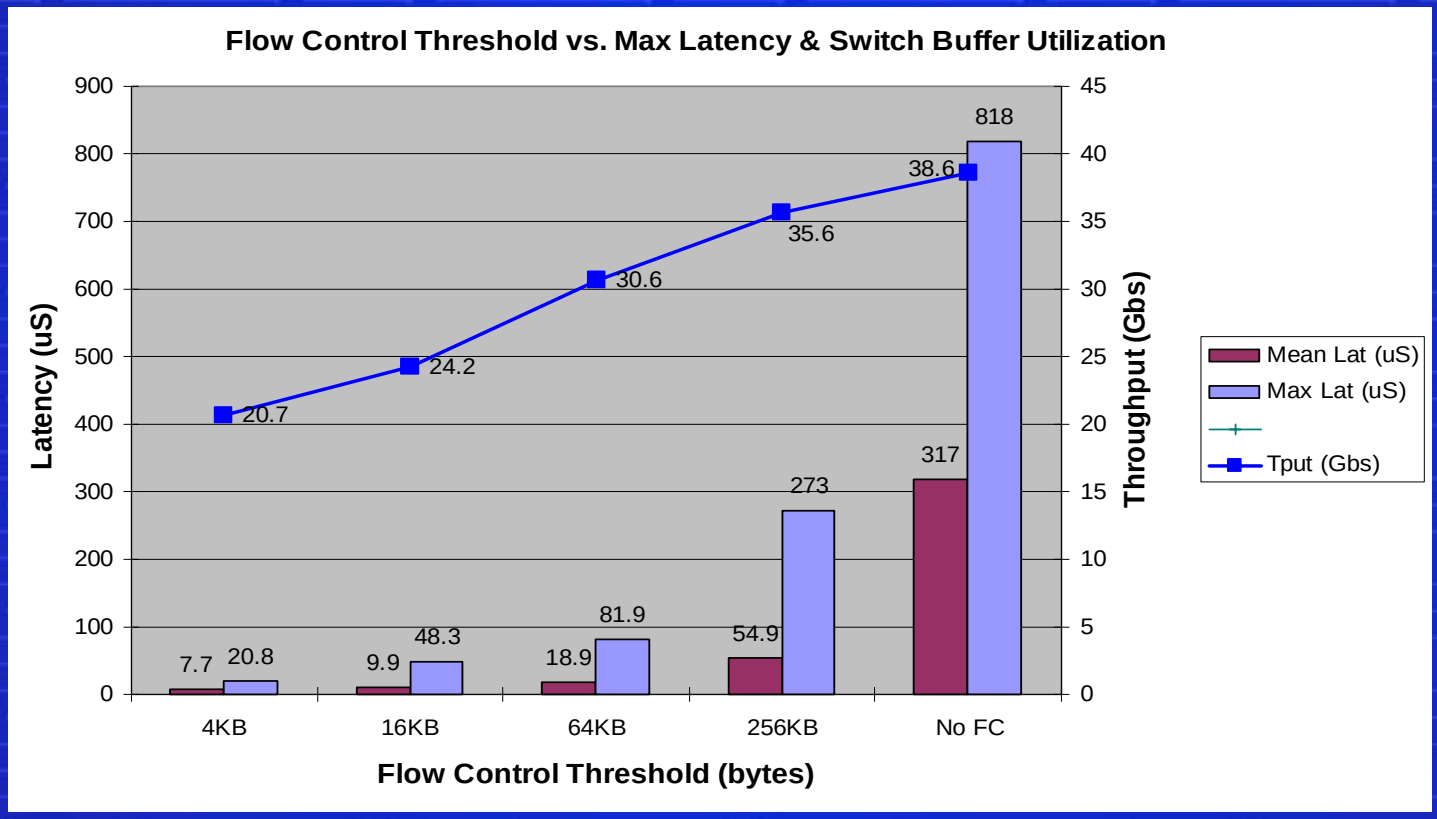
~29 Gbs (Due to Discard)

~24 Gbs (Due to Blocking)

Latency reduction with 802.3x(XON/XOFF) costs Throughput

Throughput & Latency Summary

Bursty Workload, 802.3x (XON/XOFF),
1.5MB Switch Mem, Hi-Thresh = 4KB, 16KB, 64KB, 256KB, None



802.3x currently supports throughput/latency trade-offs

Issues with 802.3x

- Latency can be controlled, but …
  - Throughput is lost in 2 ways:
    - Loss to XOFF blocking = 38 – 24 = ~ 14 Gbs or ~36%
    - Pause Frames @ 165K F/S = ~111 Mbs or ~1.11%
  - Initial testing of 802.1p traffic loads shows a Max latency issue on high priorities (still investigating)
- Latencies pushed back to end-points, but …
  - Enables higher layers to deal with it
  - Advances in ULP stacks will significantly diminish this issue
- How to handle QoS requirements?

Can throughput loss be avoided while keeping lower latency?

Summary & Next Steps

- 802.3x can constrain fabric latencies
- But … creates other issues
  - Throughput & Max latency issues remain
- Need to study simple enhancements to existing MAC Control Sub-layer
  - To reclaim throughput w/o sacrificing latency or packet delivery
  - To contain Max latency
- Next step to evaluate and simulate simple enhancements

Will present proposals and results in next plenary meeting