Proposal for 802.3 Enhancements for Congestion Management

Intel Corp.

Manoj Wadekar

Gary McAlpine

Tanmay Gupta

Agenda

- Nature of the problem
- Differentiated Service Support in 802.3 MAC
- Proposed Adaptive Rate Control Protocol
- Preliminary Simulation Results
- Summary

intel

Nature of the Problem

- In Switched Interconnects:
  - Even non-blocking switches experience congestion at TX ports
  - Typical reaction to congestion is frame discard, but ...
    - Unacceptable in some short range interconnects
  - 802.3x flow controls links to avoid overflow, but …
    - Increases BW loss and jitter
- The Basic Problems with 802.3x:
  - No priority awareness
  - All the priorities of traffic get equal punishment
    - Creates Challenges for Differential Service to various flows
  - Inserts dead time on the links
    - Costs BW
  - Punishment doled-out in big chunks (XOFF/XON)
    - Induces significant jitter

intel

Defining Congestion

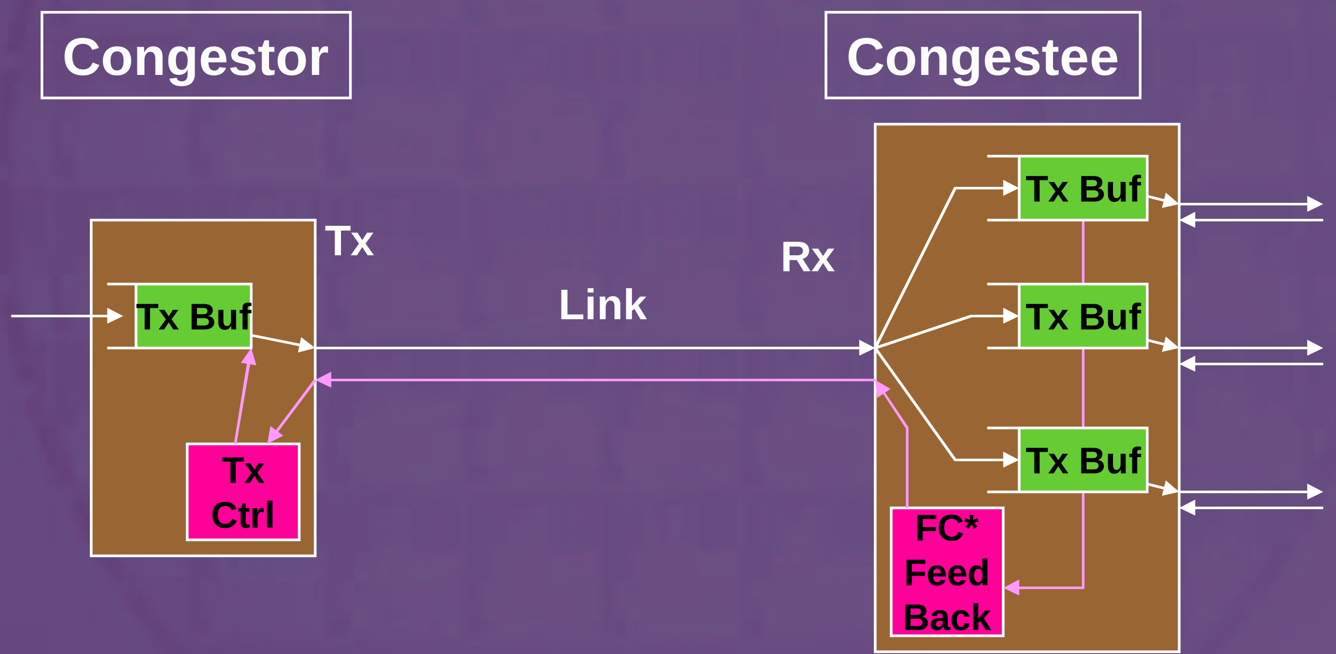
- Congestion is of two general types:
  - Transitory

    Traffic which can be smoothed over time, without frame drop because average bandwidth demand is less than capacity and peak demand that can be buffered
  - Oversubscription

    Traffic which cannot be smoothed over time and results in not being admitted to network (e.g., admission control), or either results in frame drop (e.g., buffer overflow, RED) or backs up into Source buffers

Current 802.3x Flow Control Model

- All priorities get queued in single Tx Buffer
- Congestee is assumed to be an output queued switch
- Flow Control feedback indicates a device is congested
- Tx Control - temporarily block all traffic flow in response



* FC = Flow Control

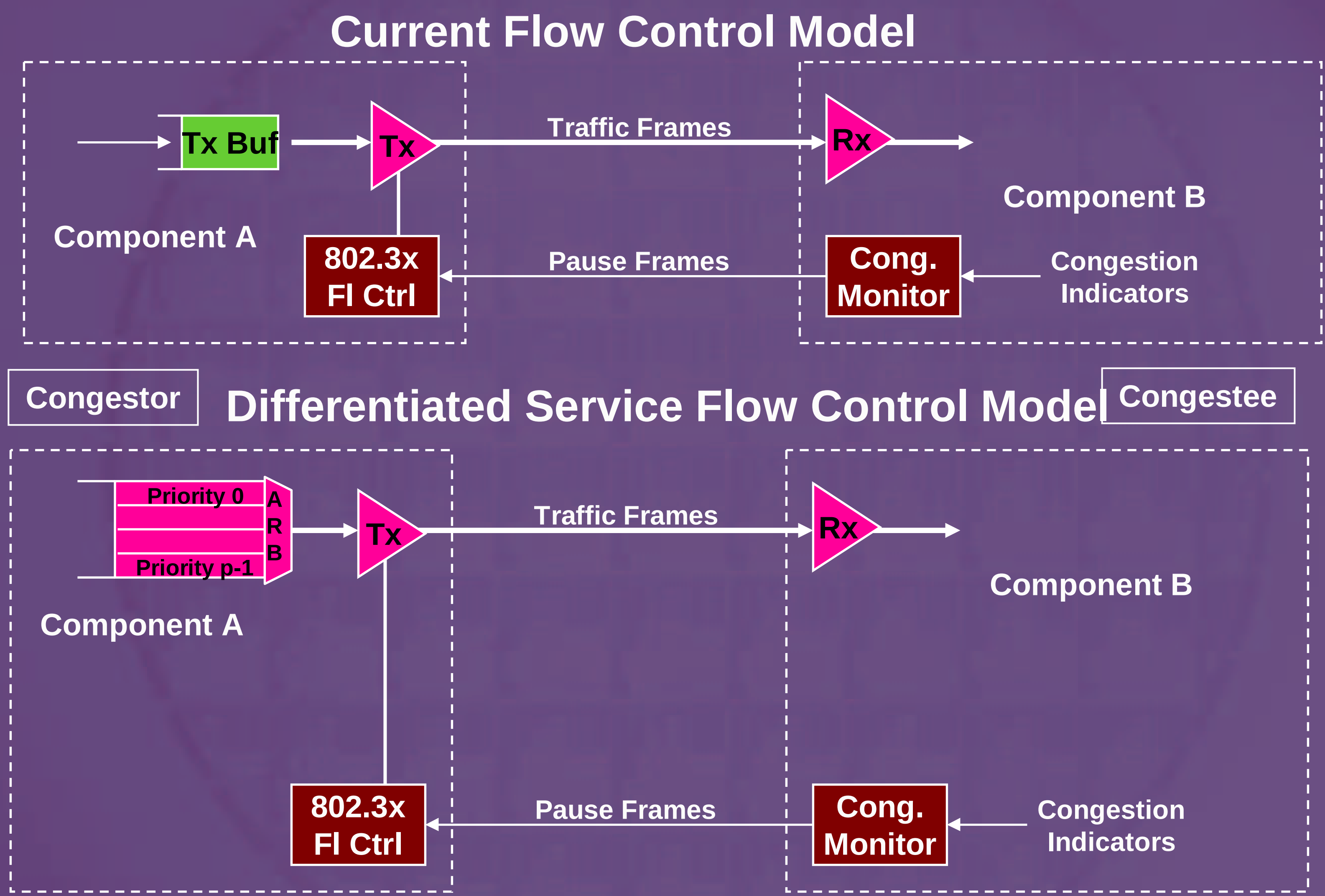
Possible Enhancements - Some early results

- Evolutionary changes in Ethernet that will:
  - Better support differentiated services
  - Reduce probability of Packet Drop at MAC Client
  - Improve throughput and latency characteristics
  - Reduce end-to-end latency in short range networks
- Look to differentiated service for high priority latency improvements
  - For Transitory Congestion
- Evaluate rate limiting protocols for total system performance improvement and for pushing congestions toward the source
  - For Oversubscription Congestion
- Following foils show preliminary simulation results

Differentiated Service

- How is this different than 802.1p?
  - 802.1p is not visible at 802.3 MAC Control Sub-layer
  - Single Transmit buffer scheduling
- Various classes of traffic from MAC Client need differentiated service
  - Enable differentiated rate control of the different priorities within the MAC Control Sub-layer
- Arbitration among different classes
  - High priority traffic gets priority in transmission

Flow Control Model Comparisons



Adaptive Rate Control (ARC)

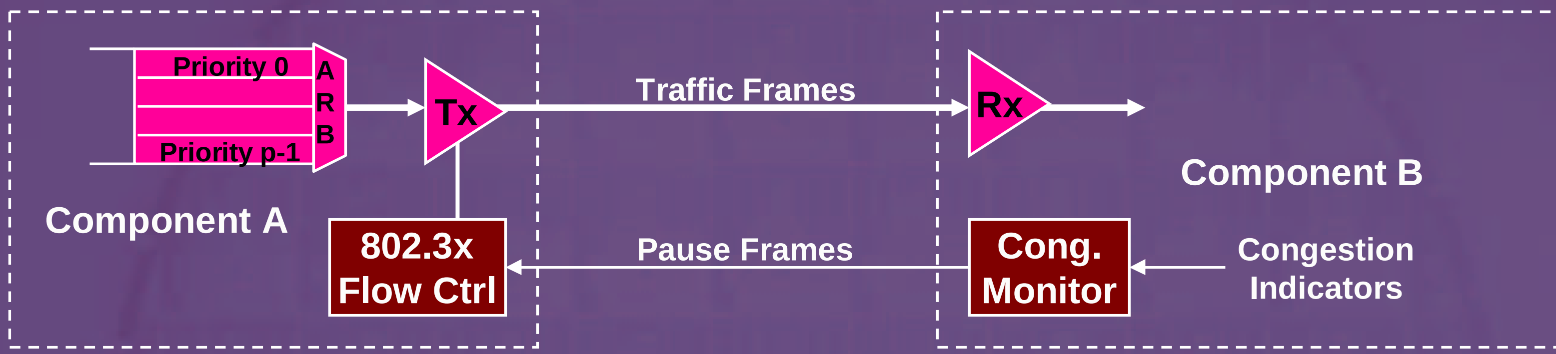
- Receiver (Congestee) provides Congestion feedback
  - Use XUP/XDOWN messages to control transmission rate
  - Granularity of feedback – per priority class
  - Multiple XUP/XDOWN may be generated for feedback

- Transmitter (Congestor) treats XUP/XDOWN messages as PUNISH/REWARD
  - Increases TX rate for given priority class for each XUP received
  - Decreases TX rate for given priority class for each XDOWN received

- Rate is controlled by inserting IPGs at individual queue outputs
  - IPG sizes determined by priority, punishment factor, & packet size
  - Punishment factor and affected class determined by Flow Control feedback

Flow Control Model Comparisons

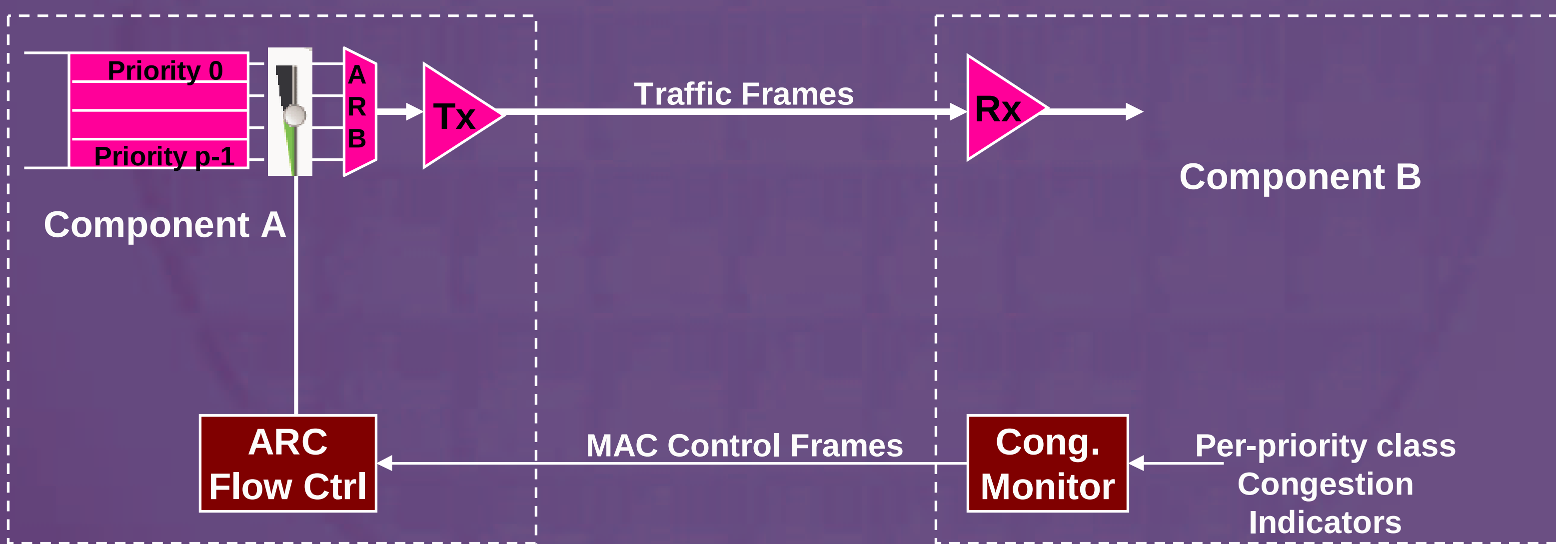
Differentiated Service Flow Control Model



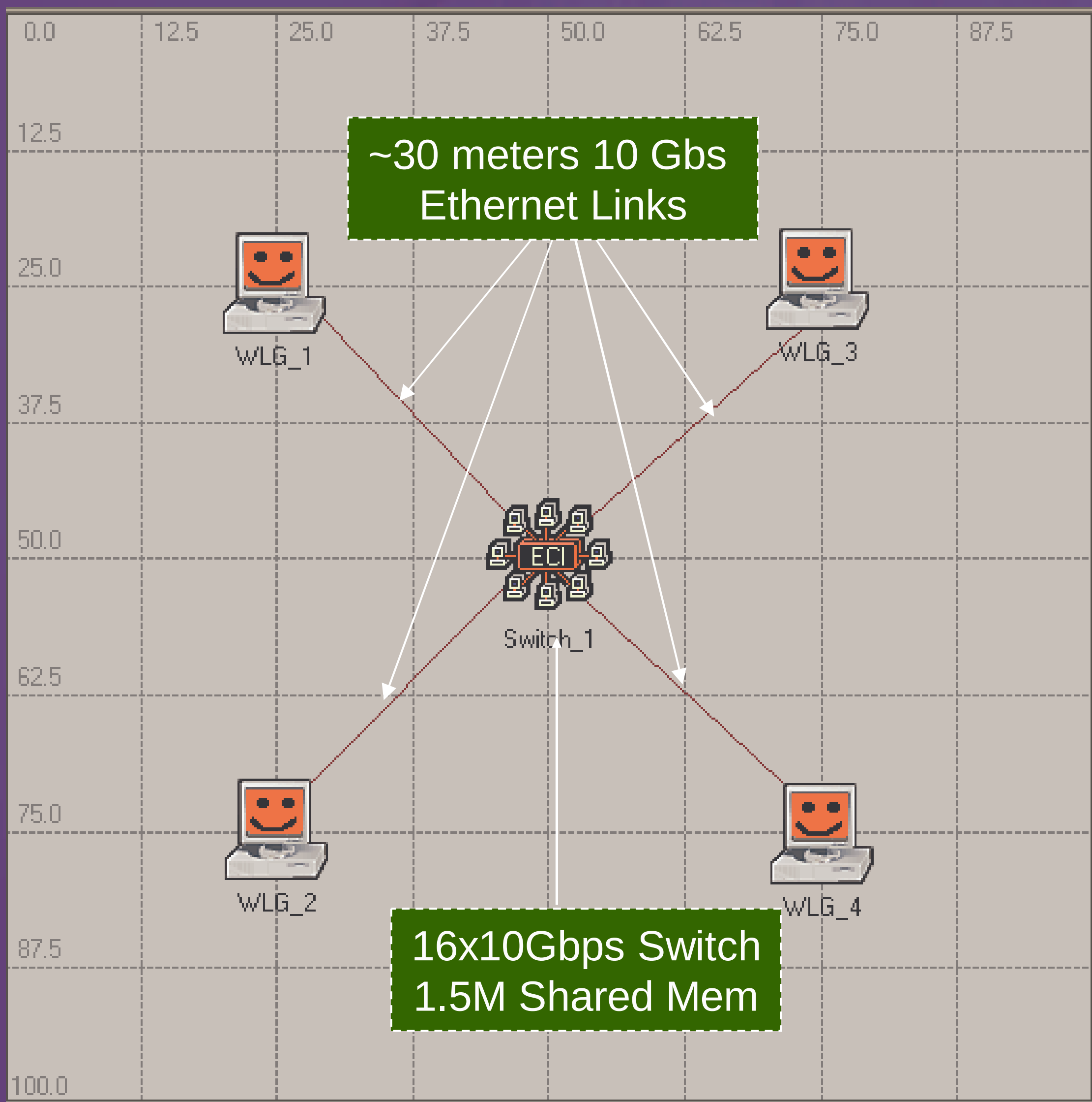
Congestor

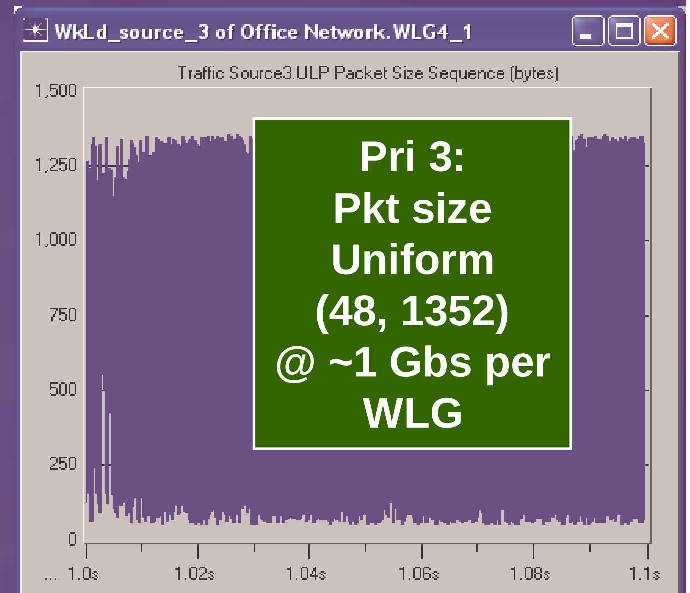
ARC Flow Control Model

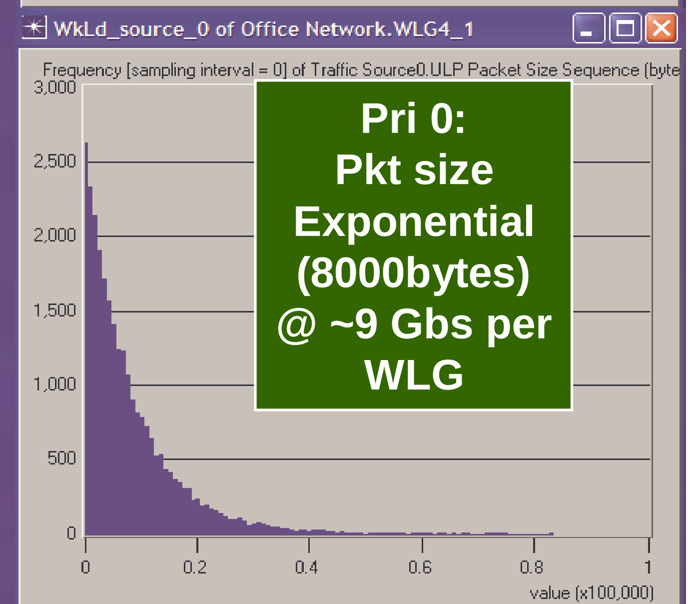
Congestee



Simulation Environment







Scenarios

- No Flow Control
- 802.3x Flow Control (Hi-Threshold = 16k)
- Adaptive Rate Control (Hi-Threshold = 16k)

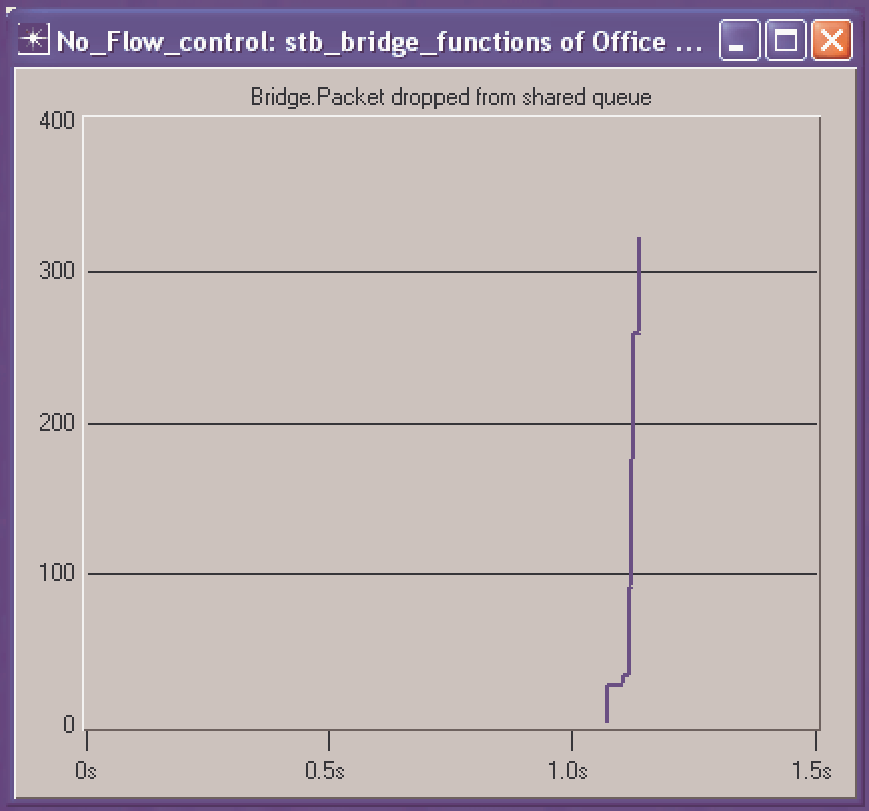
Note: ARC in the simulation does not have granular control over each priority

2 Priority Traffic Test

- 4 Workload Generators @ 10 Gbs each
  - Each generating 2 priorities of traffic
  - Priority 0 = Rand. ULP Pkt Sizes (48 to ~80000 Bytes)
    - Exponential distribution w/ mean of 8000 Bytes
    - 9 Gbs from each Workload Generator
    - Total 4 WLG = 36 Gbs Max
  - Pri 3 = Rand. ULP Pkt Sizes (48 to 1352 Bytes)
    - Uniform distribution w/ a mean of 700 Bytes
    - 1 Gbs from each Workload Generator
    - Total 4 WLG = 4 Gbs
- Latency measured per ULP segment (802.3 Frame)
  - 1st byte from source memory to last byte to sink memory
  - Includes source NIC read, 1st hop, Switch, 2nd hop, Dest NIC write

intel

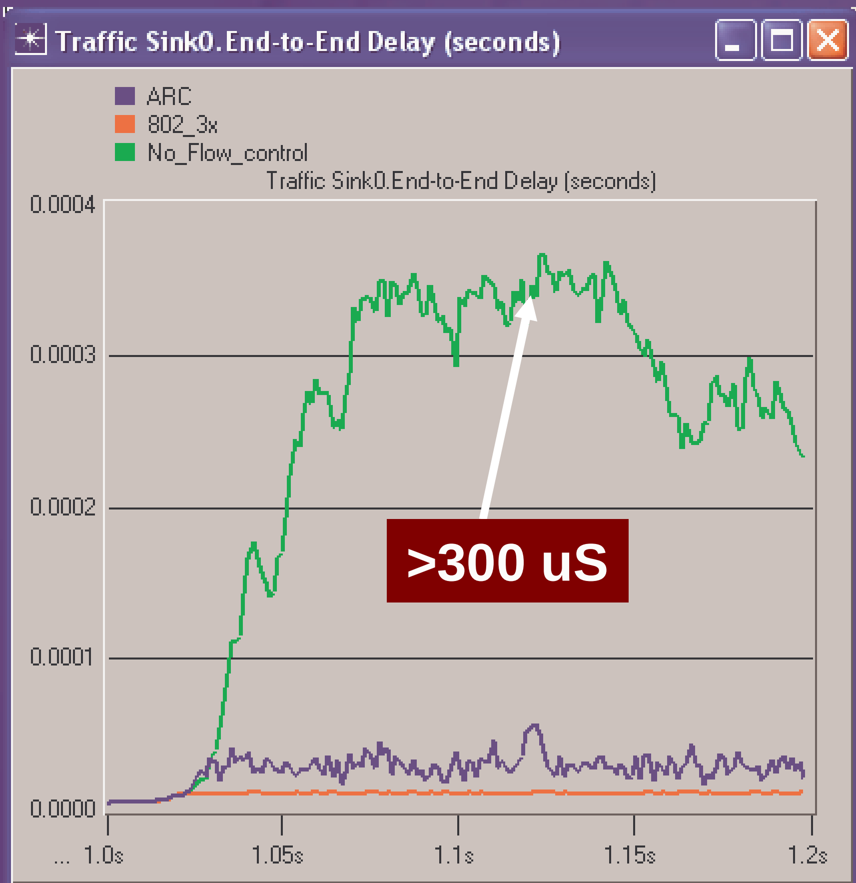
Packet Drop at the Bridge



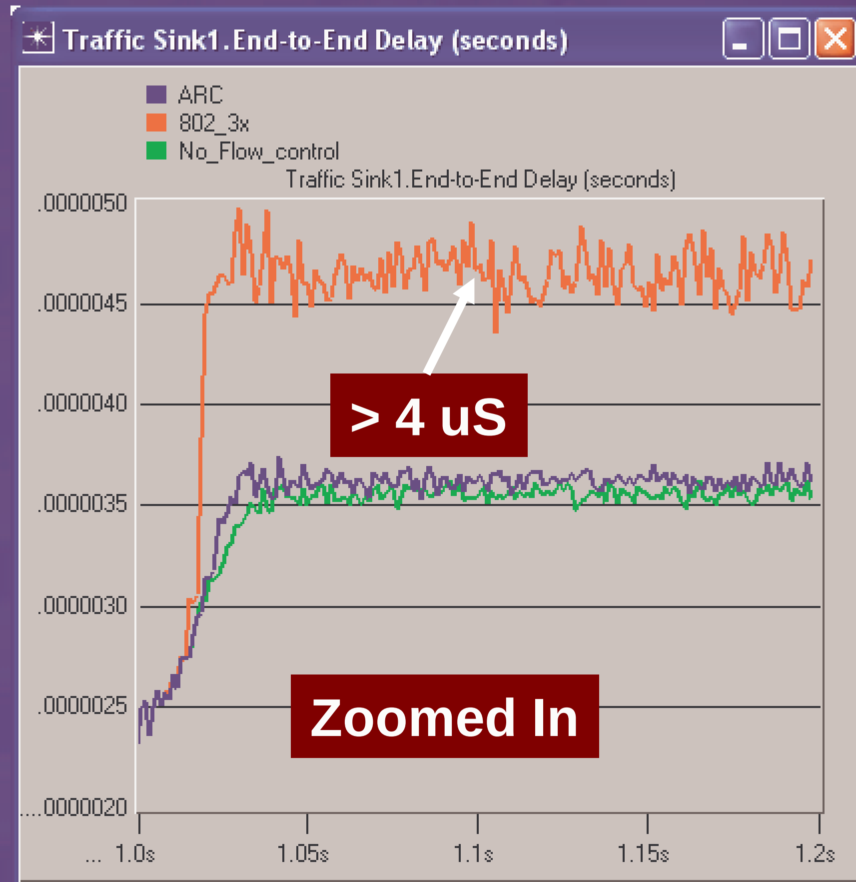
Rate and Flow Control Protocols avoid packet drop.
Packet drop increases end-to-end latency substantially

Latency Benefits

Low Priority Traffic



High Priority Traffic

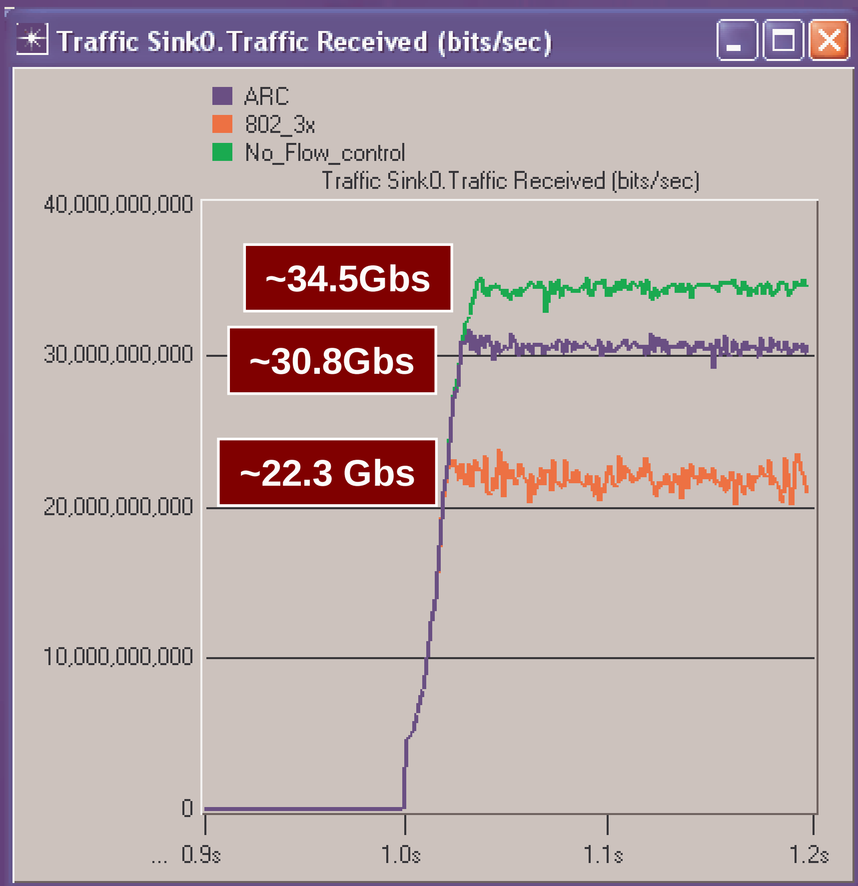


Better Congested Latency Characteristics than 802.3x or No FC

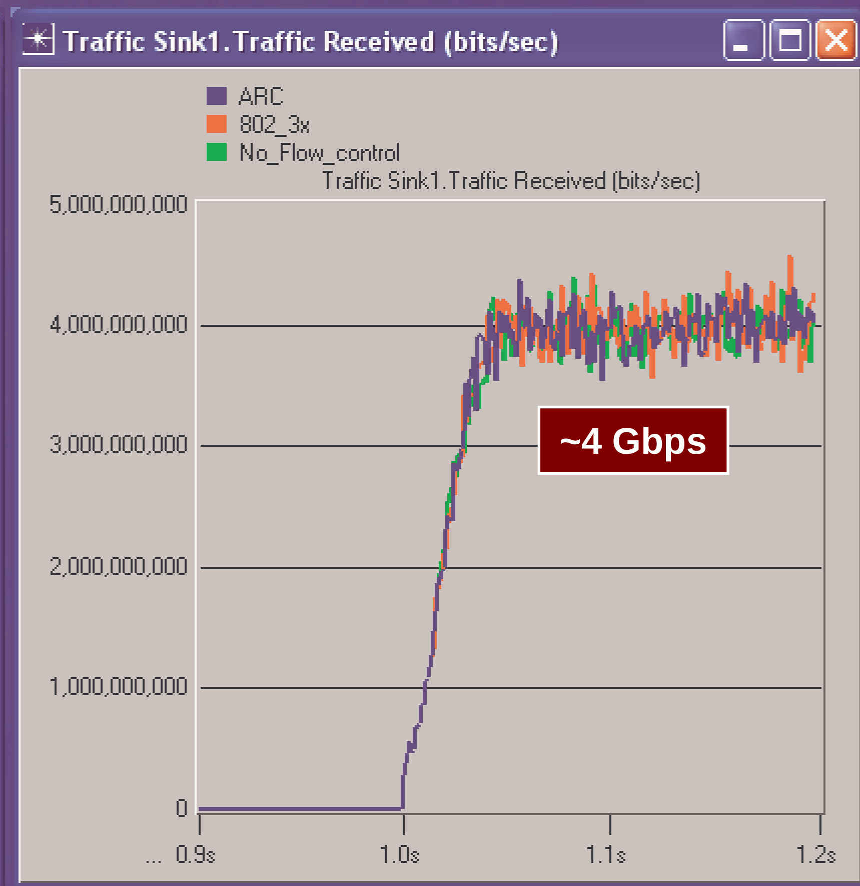
intel

Throughput Benefits

Low Priority Traffic



High Priority Traffic

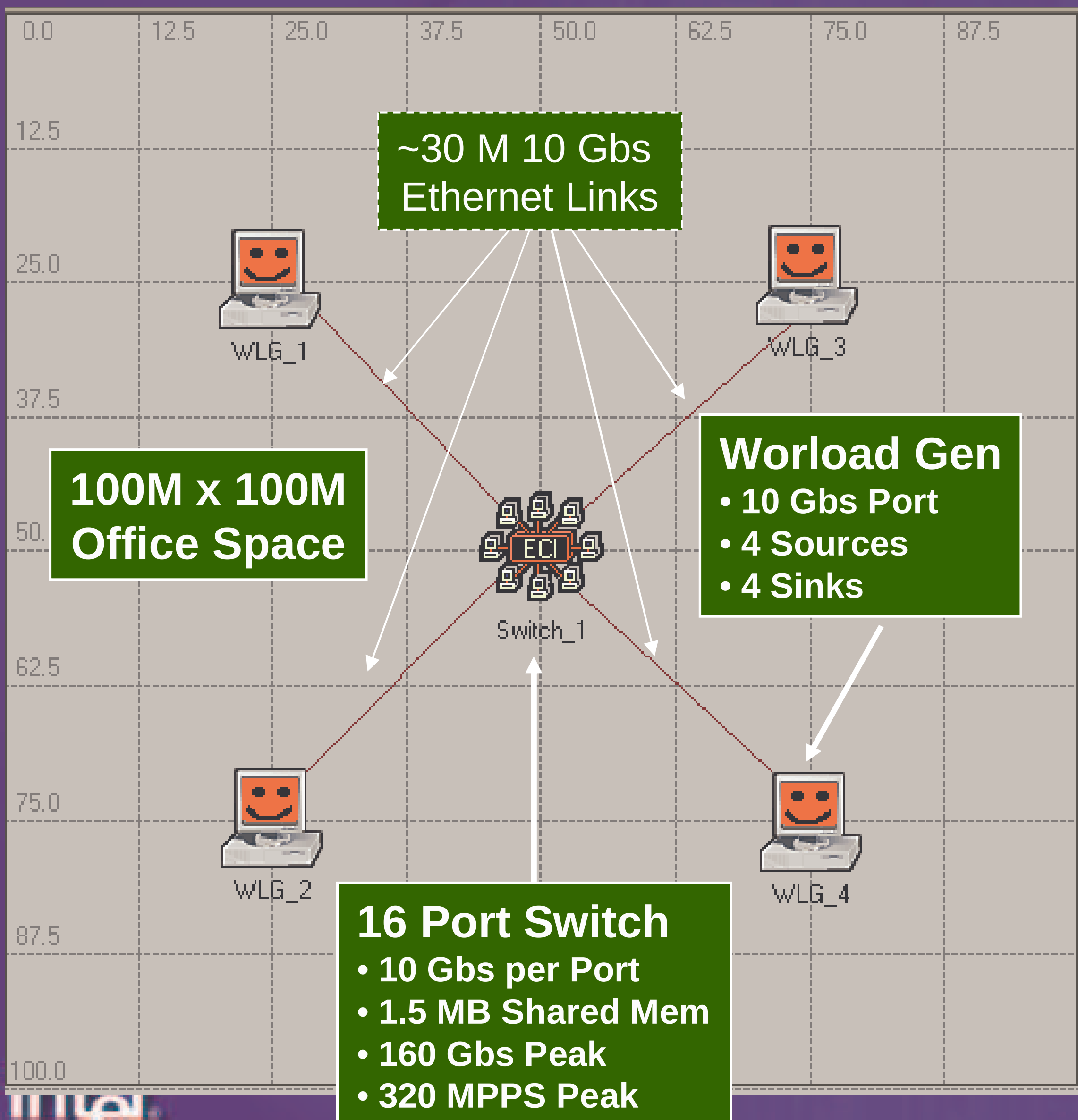


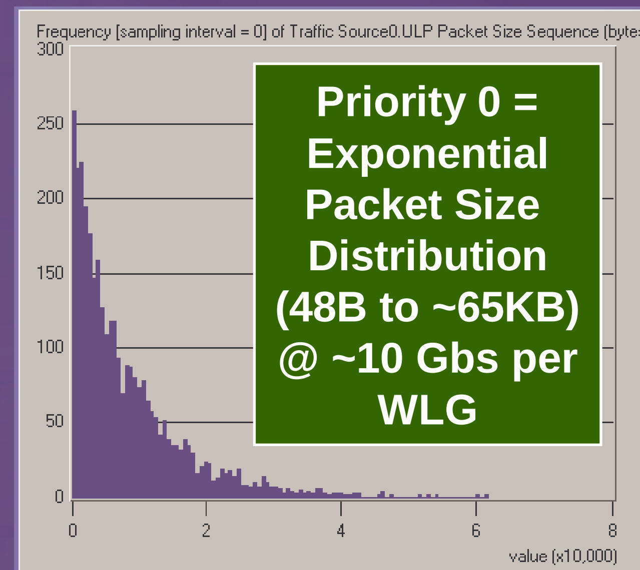
Adaptive Rate Control Provides better throughput than 802.3x.

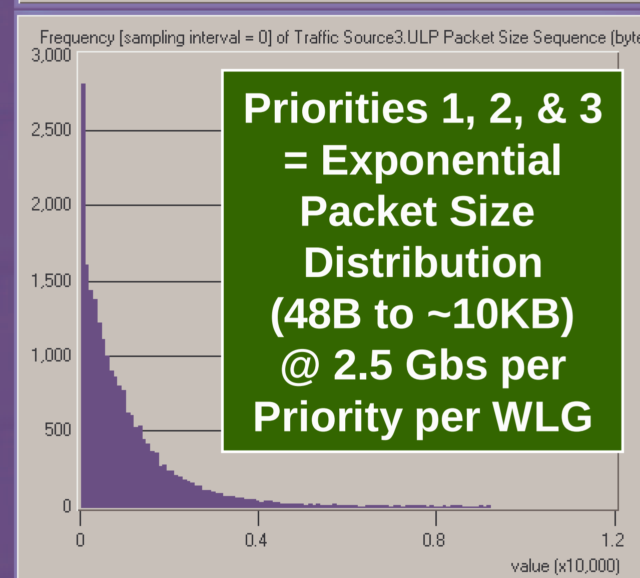
4 Priority Traffic Test

- 4 Workload Generators @ 10 Gbs each
  - Each generating 4 priorities of traffic
  - Priority 0 = Rand. ULP Pkt Sizes (48 to 65000 Bytes)
    - Exponential distribution w/ mean of 8000 Bytes
    - Provides background load, tries to hog all BW
  - Pri 1, 2, & 3 = Rand. ULP Pkt Sizes (48 to 10200 Bytes)
    - Exponential distribution w/ mean of 1000 Bytes
    - 2.5 Gbs of each priority from each Workload Generator
    - Total 4 WLG = 10 Gbs each pri X 3 priorities = 30 Gbs total
- Cut-through enabled

4 Priority Test Model & Workload

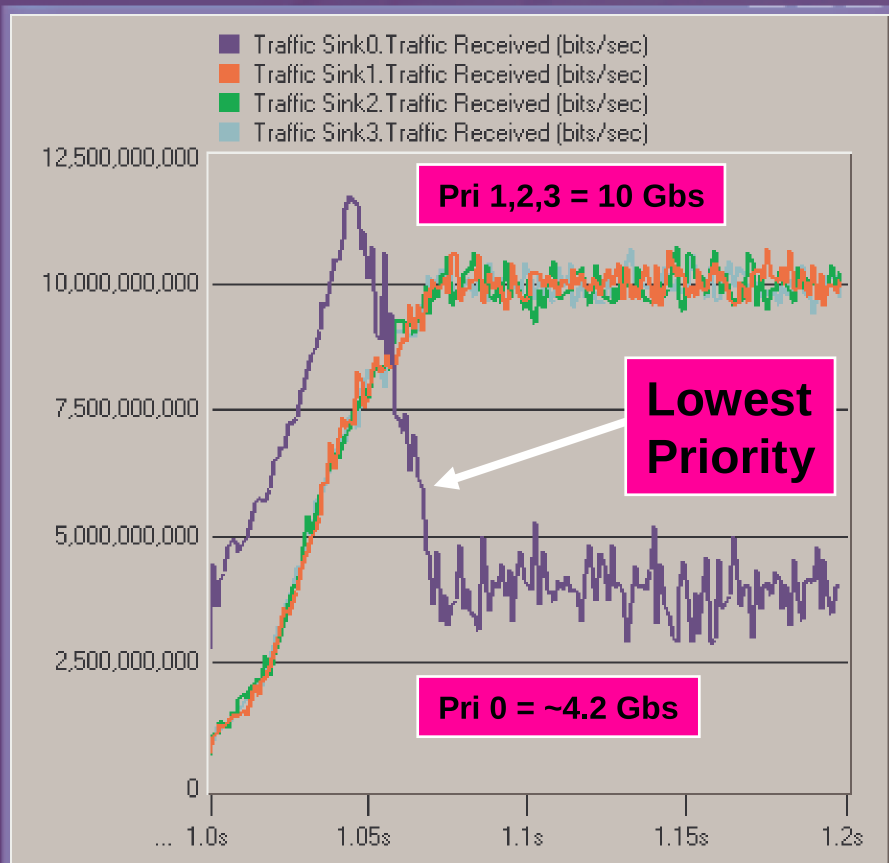




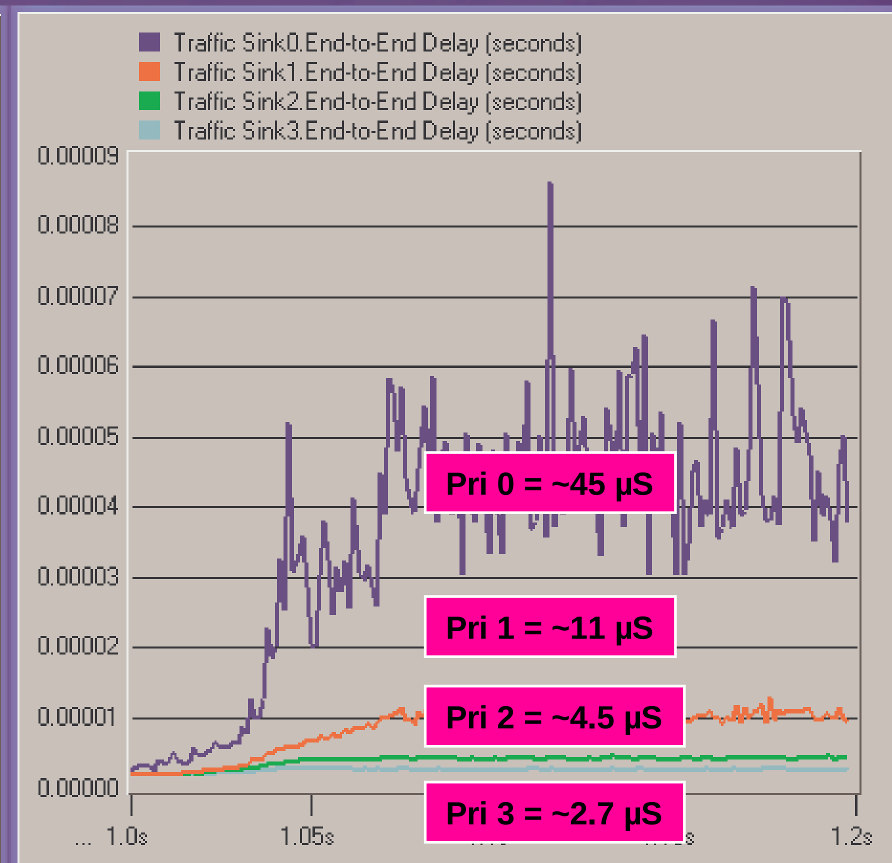


ARC – 4 Pri Throughput & Latency

Throughput

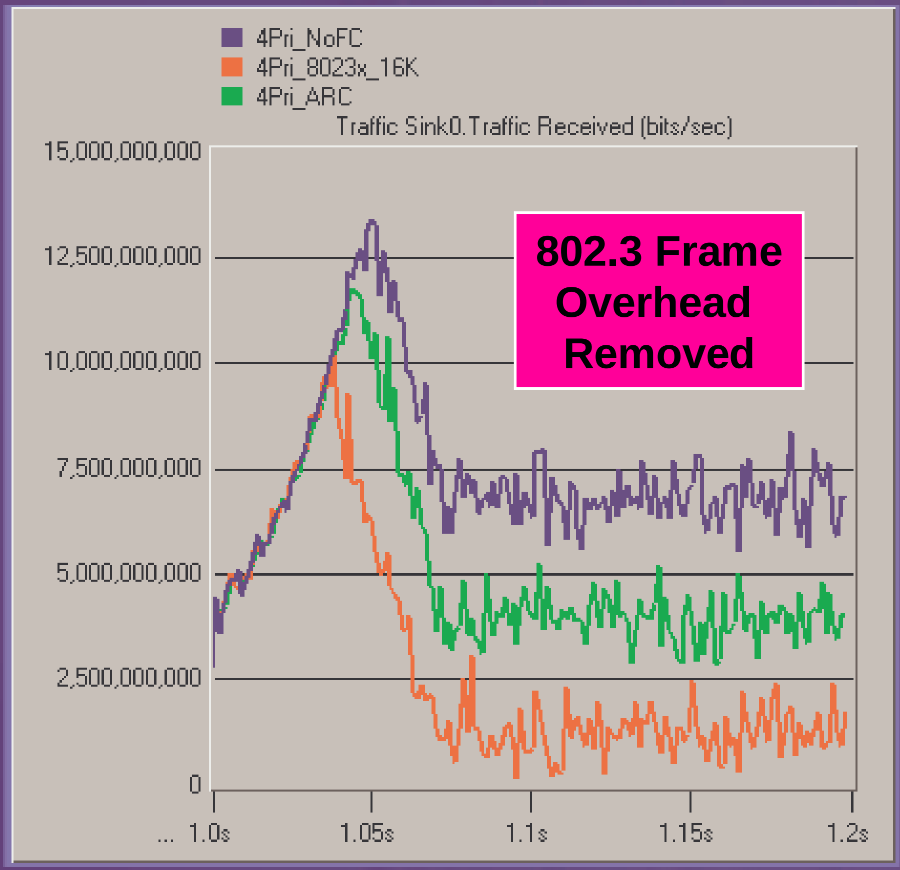


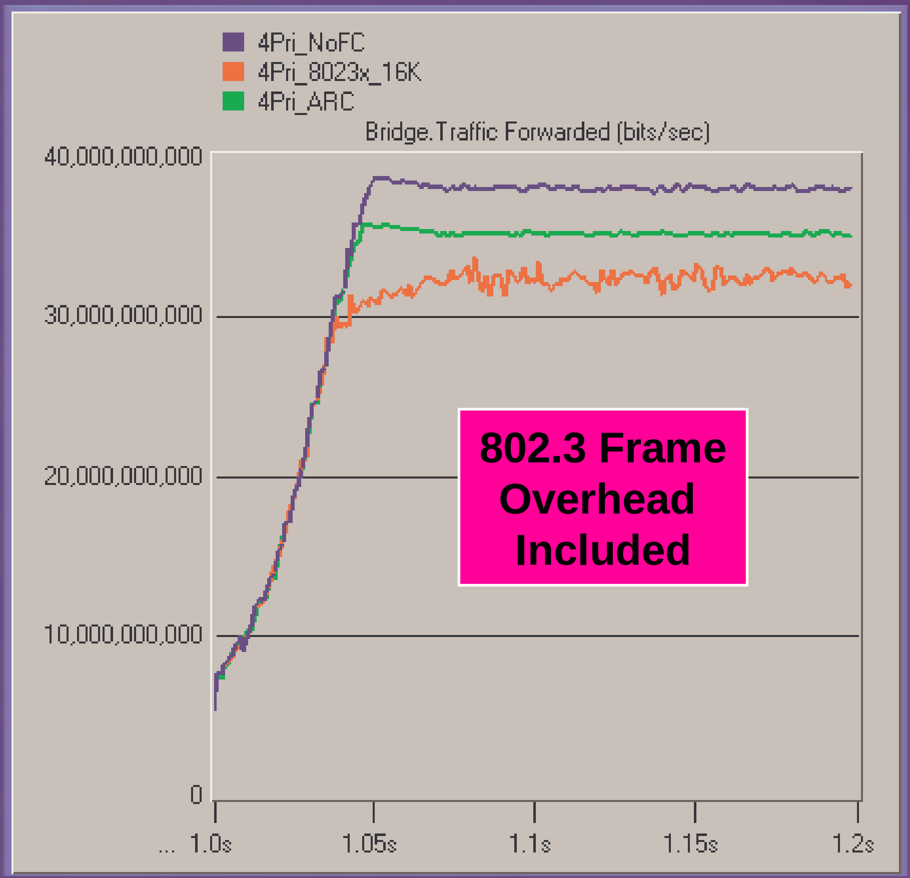
Mean Latency



Excellent differentiation characteristics during severe congestion

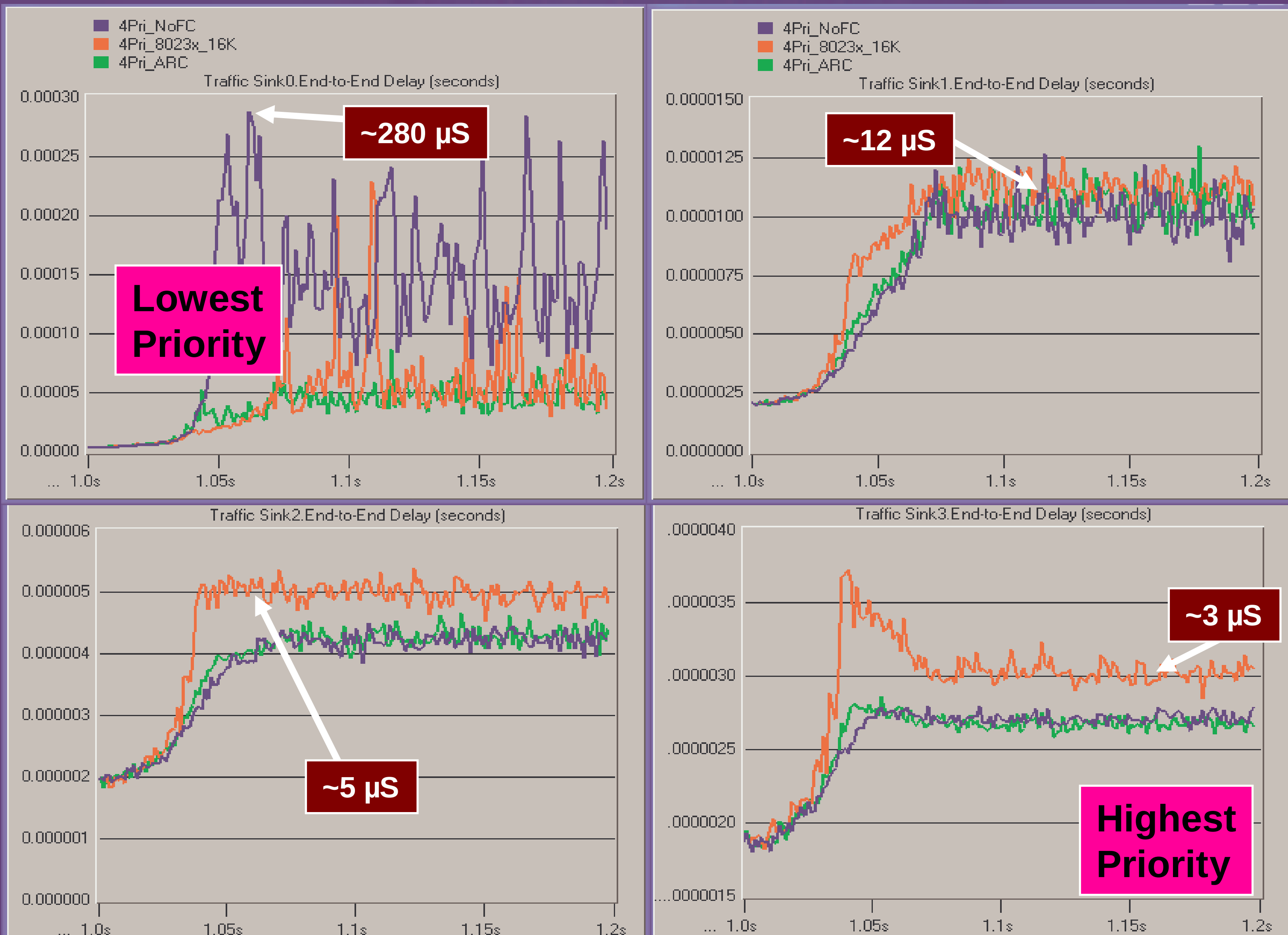
Pri 0 & Total Throughput Comparison





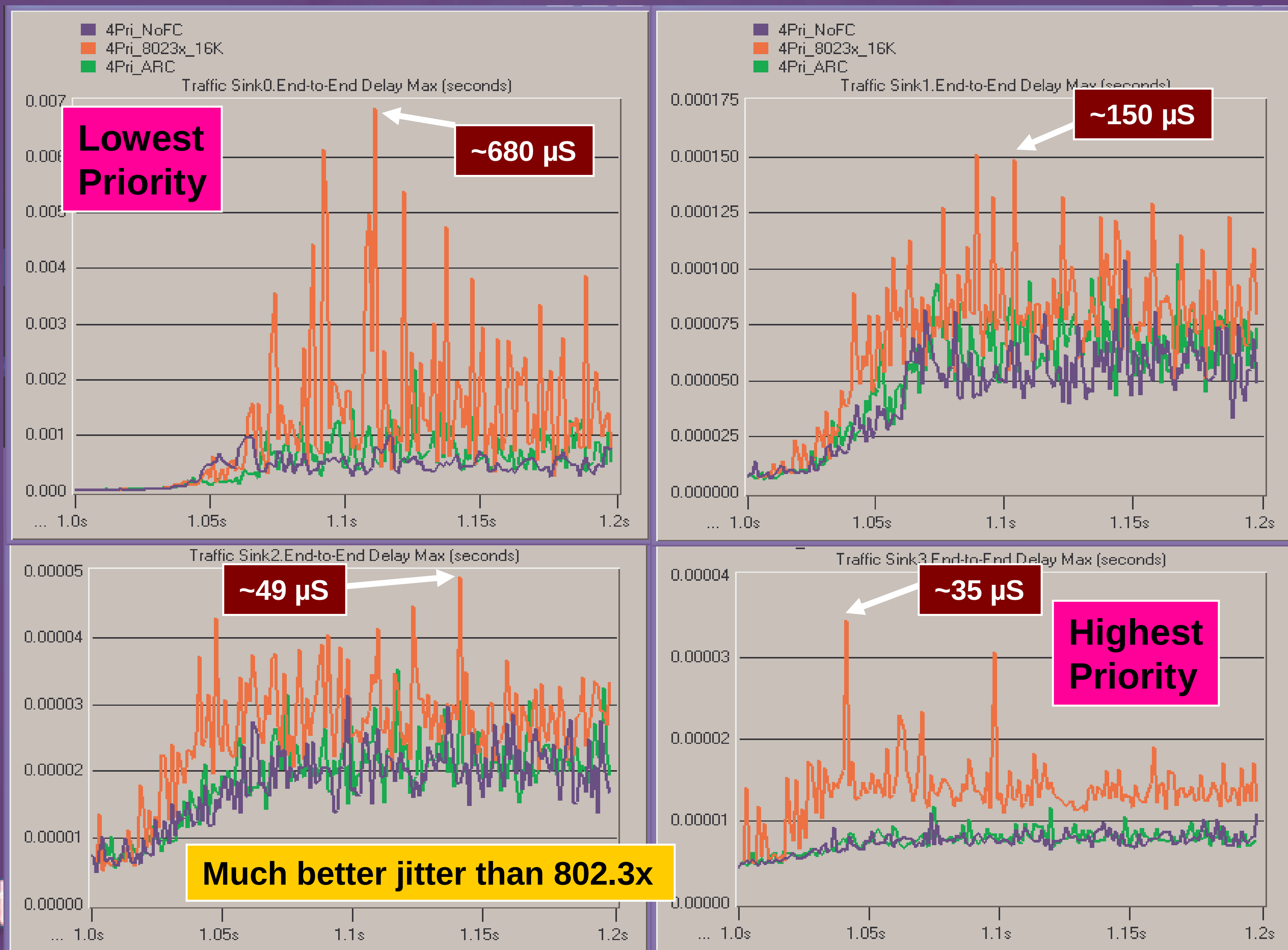
Better overall throughput characteristics than 802.3x during severe congestion

Mean Latency



Better mean latency than No FC or 802.3x

Max Latency - "J-J-J-i-t-t-t-t-e-r" Ind.



Summary & Next Steps

- 802.3x can constrain latencies
- But … creates other issues
  - Does not guarantee Differentiation in Transitory congestion
  - Throughput & Max latency issues remain
- Need to study simple enhancements to existing MAC Control Sub-layer
  - Provide for Differentiated Service within 802.3
  - Consider Rate Control protocols for Oversubscribed congestion
  - Preliminary simulation results show promise
  - Further simulation to study TCP/IP workloads