TCP and QCN in a Multihop Output Generated Hot Spot Scenario

Brad Matthews, Bruce Kwan & Ashvin Lakshmikantha

IEEE 802.1Qau Plenary Meeting (Orlando, FL) – March 2008

Goals

- Quantify performance of QCN + TCP interactions in terms of Innocent Flow and Hot Spot Flow throughput
  - 500ms congestion scenario
  - Periodic congestion scenario

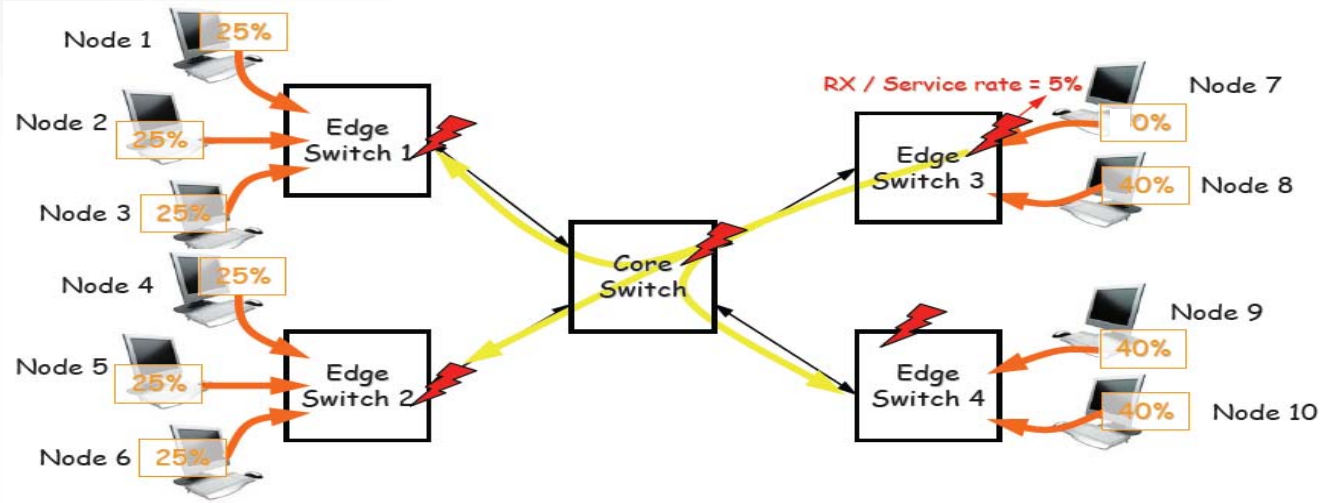
BROADCOM.
Connecting everything®

System Parameters

- Congestion Management Schemes
  - TCP Only
  - TCP + QCN
  - TCP + QCN + PAUSE
  - QCN + PAUSE
- Switch Parameters
  - Shared memory switch
    - Memory Size 2.4Mbytes
  - Partitioned memory per input, shared among all outputs
  - PAUSE Disabled
    - Output queue limit of 150kbytes
  - PAUSE Enabled
    - No output queue limit
    - Applied on a per input basis based on watermarks
    - Watermark_hi = 130kbytes
    - Watermark_lo = 110kbytes
- QCN Parameters
  - W = 2.0
  - Q_EQ = 26kbytes
  - Gd = 1/128 = 0.0078125
  - Base marking: once every 150kbytes
  - Jitter on marking: 30%
  - Runit = 1Mb/s
  - MIN_RATE = 10Mb/s
  - BC_LIMIT = 150kbytes
  - TIMER_PERIOD = 15ms
  - R_AI = 5Mbps
  - R_HAI = 50Mbps
  - FAST_RECOVERY_TH = 5
  - Quantized_Fb: 6 bits
  - Jitter at RP: 30% (byte counter and timer)
- TCP Parameters
  - New Reno [RFC3782]
  - No ECN/RED
  - Min RTO = 10ms
  - Delayed ACK Timeout = 1ms

BROADCOM.
Connecting everything®

Topology and Workload: 500ms Congestion Event

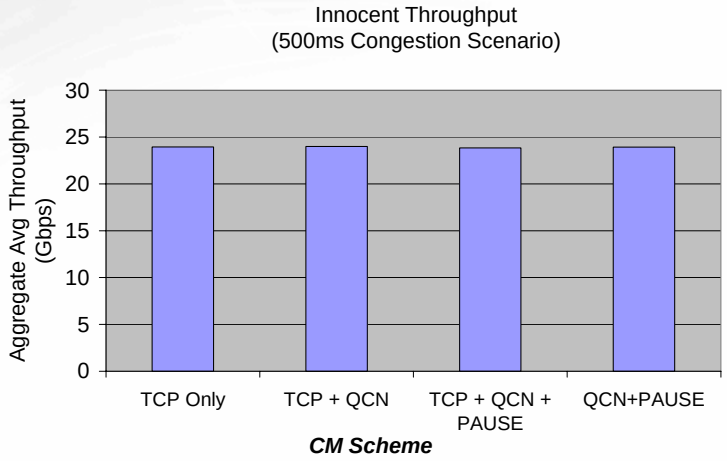


- Multi-stage Output-Generated Hotspot Scenario
  - Link Speed = 10Gbps for all links
  - Loop Latency = 16us
- Traffic Pattern
  - 9k byte transactions arriving with a Bernoulli distribution
  - Transport layer is either UDP or TCP
  - Destination Distribution: Uniform distribution to all nodes (except self)
  - Frame Size Distribution: Fixed length (1500bytes) frames
  - Offered Load
    - Nodes 1-6 = 25% (2.5Gbps)
    - Nodes 8-10 = 40% (4Gbps)
- Congestion Scenario
  - Node 7 temporarily reduces its service rate from 10Gbps to 500Mbps between [0-500ms]

** Topology and Workload based on Benchmark #2: OG HS Multi-Hop. Congestion*
Picture is from: http://www.ieee802.org/1/files/public/docs2006/au-sim-Zurich-Hotspot-Benchmark-OG-MS-r2.pdf

BROADCOM.
Connecting everything

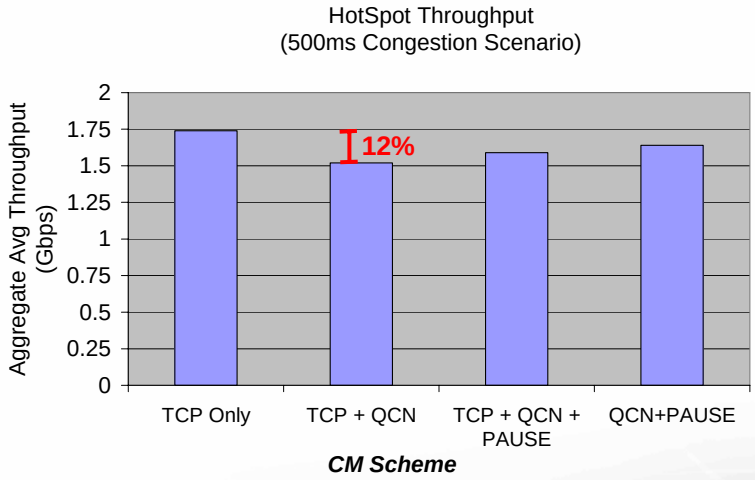
500ms Congestion Scenario
Innocent and Hot Spot Throughput



Ideal aggregate innocent flow throughput is 24Gbps.

For the 500ms Congestion Scenario, TCP performs well on its own.

Introducing QCN does not degrade innocent throughput for this scenario.



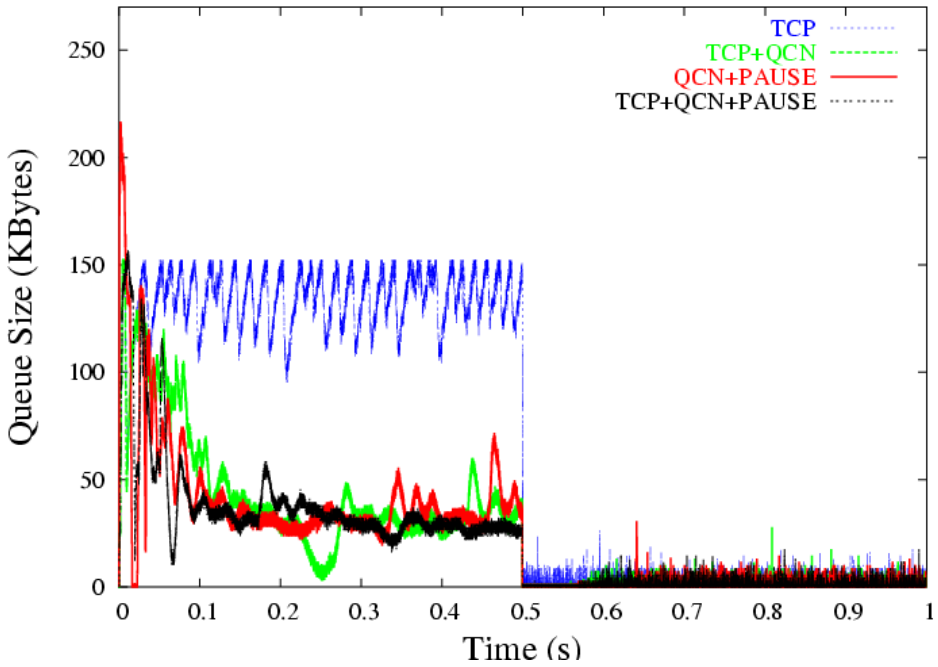
Ideal aggregate hot spot flow throughput is 1.75 Gbps.

For the 500ms Congestion Scenario, TCP performs well on its own.

Introducing QCN does appear to ***mildly degrade*** the hot spot throughput for this scenario.

BROADCOM.
Connecting everything®

500ms Congestion Scenario
Instantaneous Queue Size vs Time



TCP-only case relies on frame discards to manage congestion.

Use of standard AQM schemes need to be considered to provide a better comparison of queue behavior.

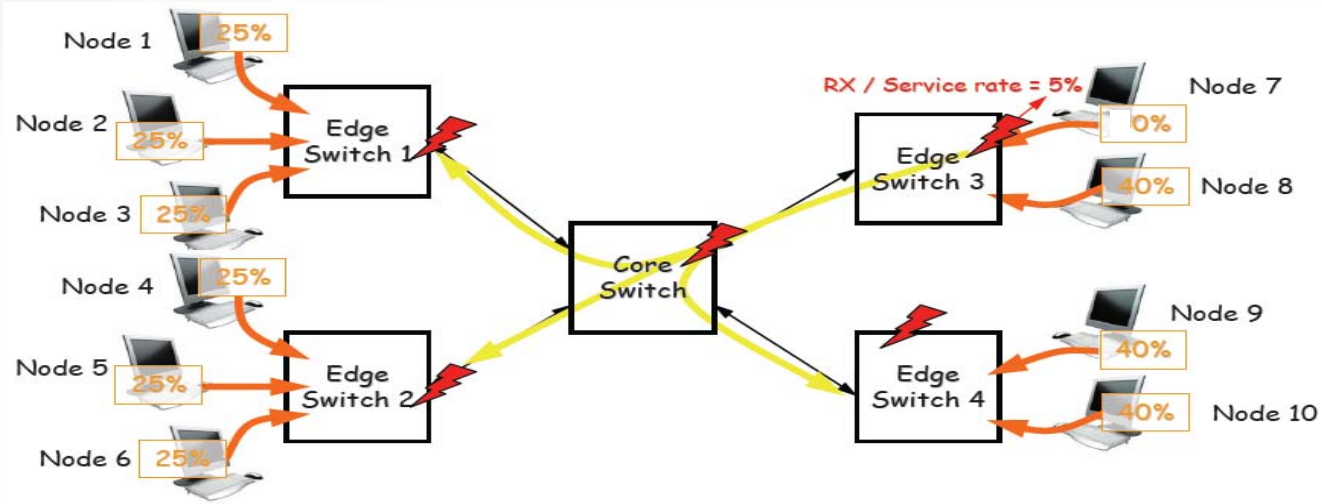
BROADCOM.
Connecting everything®

Comment on Differences from Previous TCP Simulation Results

- Past TCP Simulation Results [from Eric Geisler & Manoj Wadekar]
  - http://www.ieee802.org/1/files/public/docs2007/au-sim-geisler-cm-tcp-effects-1107-v1.pdf
  - TCP throughput performance appeared degraded due to retransmission timeouts that occurred due to frame loss
- Key Difference
  - TCP Stack: New Reno v.s. Reno
    - New Reno is a common TCP stack [RFC 3782] that provides an improved handling of recovering from multiple frame loss scenarios as compared with Reno
  - MinRTO
    - 100ms is used in the 11/2007 presentation noted above whereas 10ms is used in this presentation to capture the fact that RTT delays are smaller for Data Center networks than for Internet environments
- Impact
  - TCP's ability to recover well from multiple frame losses using New Reno is reflected in the improved Hot Spot throughput performance found in this presentation

BROADCOM.
Connecting everything®

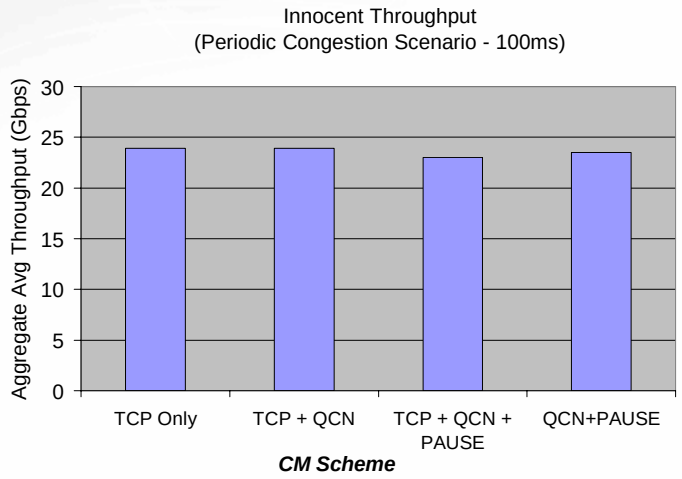
Topology and Workload: Periodic Congestion Events



- Traffic Pattern
  - Same as before
- Congestion Scenario
  - Node 7 periodically reduces its service rate from 10Gbps to 500Mbps
  - Congestion Duration: 100ms
    - Duty Cycle = 1/2
- Simulation Duration: 1 second
- Performance Metric: Aggregate Throughput
  - Ideal Aggregate Innocent Flow Throughput: 24Gbps
  - Ideal Aggregate Victim Flow Throughput: 500Mbps or 3Gbps (Avg = 1.75Gbps)

BROADCOM.
Connecting everything

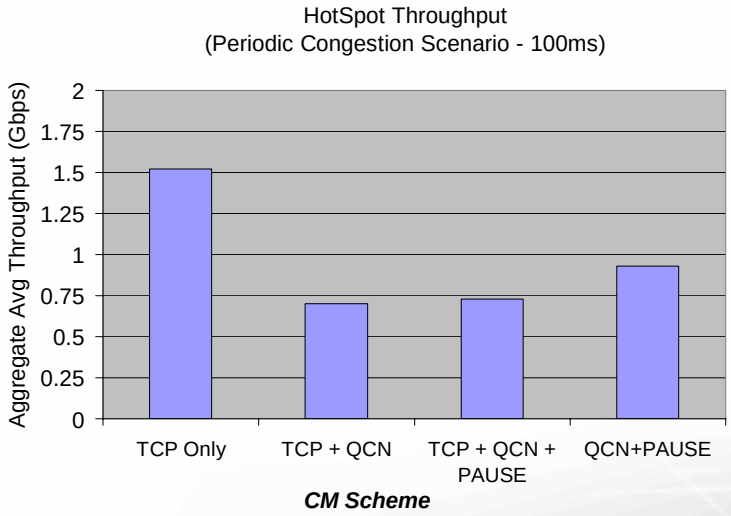
Periodic Congestion Scenario – 100ms Innocent and Hot Spot Throughput



Ideal aggregate innocent flow throughput is 24Gbps.

For the Periodic Congestion Scenario, TCP performs well on its own.

Introducing QCN does not degrade innocent throughput for this scenario.



Ideal aggregate hot spot flow throughput is 1.75 Gbps.

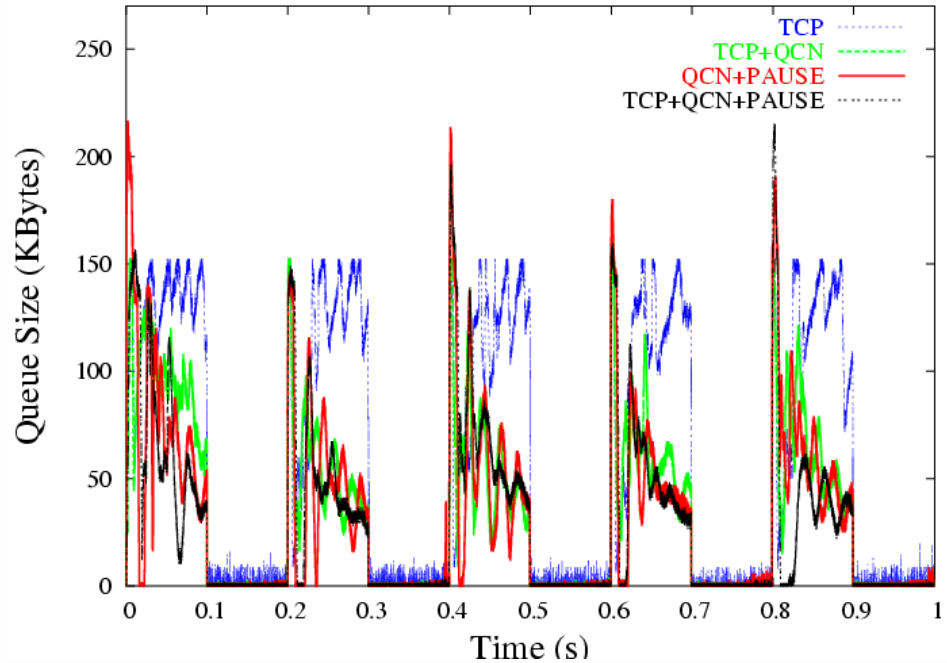
For the Periodic Congestion Scenario, TCP performs well on its own.

As previously reported*, QCN can lead to ***degradation*** in hot spot throughput when congestion periodically comes and goes at these timescales.

** By Mitch Gusat & Guenter Roeck*

BROADCOM.
Connecting everything®

Periodic Congestion Scenario – 100ms Instantaneous Queue Size vs Time



BROADCOM.
Connecting everything®

Conclusions

- TCP-Only
  - Performs well in terms of innocent/hot spot throughput for scenarios considered
  - Given scenario is TCP without AQM schemes (i.e. ECN or RED), queue resource use is higher than with QCN. ECN should mitigate this issue.

- TCP + QCN Interactions
  - Introducing QCN does not degrade innocent flow throughput when operating with TCP for the scenarios examined
  - Hot Spot throughput is degraded when QCN is enabled relative to TCP-Only case
    - This attribute is not specific to TCP interactions but has been noted in the past as a general issue

Next Steps

- Enable ECN/RED in TCP scenarios to observe interactions with QCN

BROADCOM.
Connecting everything®