Evaluating Transient Duration in a Multi-hop Output Generated Hotspot Scenario

IEEE Interim (Monterey, CA)
Bruce Kwan, Jin Ding, & Brad Matthews
January 22-25, 2007

Overview

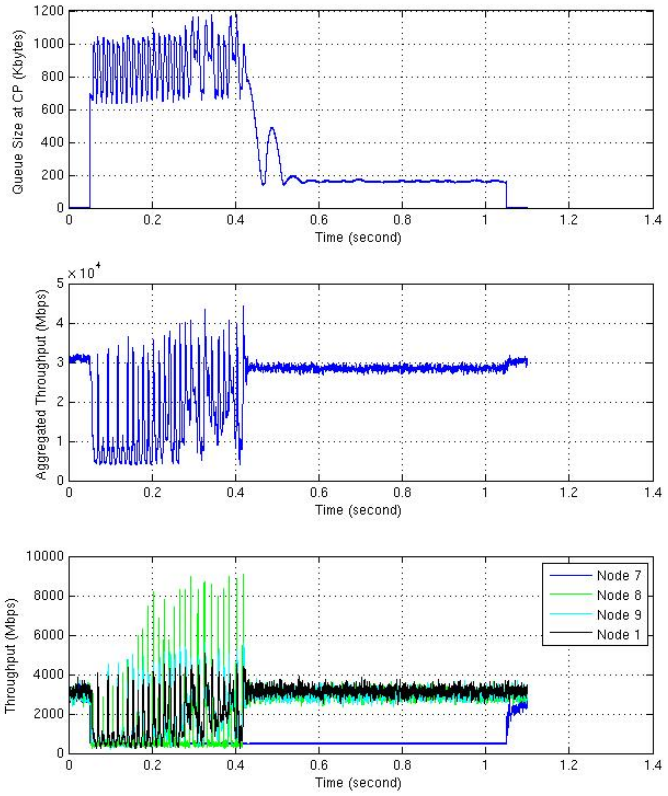
- Goals
- Observations
- Simulation Setup
- Results
- Summary & Next Steps

BROADCOM.
Connecting everything®

Experiment Goals

- Quantify Transient Duration
  - Quantify how long it takes for BCN to resolve a congestion event so that throughput in the system achieves steady state.
  - Similar approach taken in Cisco presentation on Baseline scenario
    - http://www.ieee802.org/1/files/public/docs2006/au-sim-bergamasco-bcnmax-comparison-110906v2.pdf
- Trigger Discussions
  - What should the target be?
  - What is required for our target data center applications?
  - What severe-congestion schemes are needed to improve the responsiveness?

Multi Hop Output Generated Scenario (BCN + PAUSE Only)



Overview

- Goals
- Observations
- Simulation Setup
- Results
- Summary & Next Steps

BROADCOM.
Connecting everything®

Observations

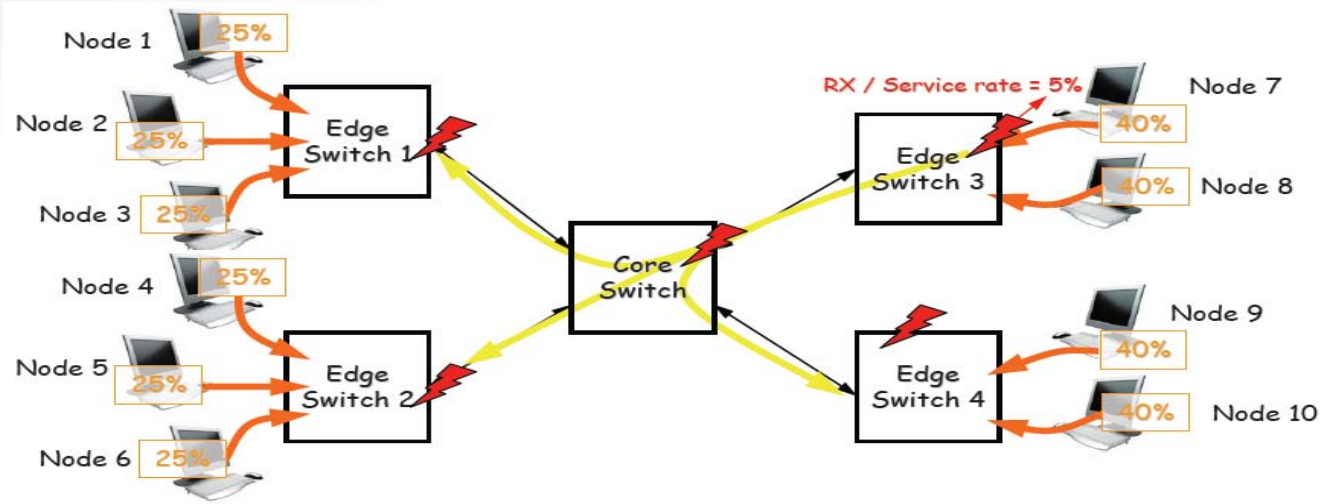
- Without additional severe congestion schemes, the transient duration for the multi-hop output generated scenario can be on the order of 462 ms (184 x RTT).
- With the use of one of the simple severe congestion management schemes (i.e. BCN-MAX), the transient duration for the multi-hop output generated scenario can be reduced to 52 ms (21 x RTT).
- Small buffer sizes can hamper transient duration (when measured in terms of RTT). Requires additional study.

BROADCOM.
Connecting everything®

Overview

- Goals
- Observations
- Simulation Setup
- Results
- Summary & Next Steps

Topology and Workload



- Multi-stage Output-Generated Hotspot Scenario
  - Link Speed = 10Gbps for all links
  - Loop Latency = 16us
- Traffic Pattern
  - 100% UDP (or Raw Ethernet) Traffic
  - 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 7-10 = 40% (4Gbps)
- Congestion Scenario
  - Node 7 temporary reduce its service rate from 10Gbps to 500Mbps between [50ms, 1050ms]

** Topology and Workload based on IBM Zurich's topology and workload as specified in Experiment #1 found: http://www.ieee802.org/1/files/public/docs2006/au-sim-Zurich-Hotspot-Benchmark-OG-MS-r2.pdf. Picture is from that presentation.*

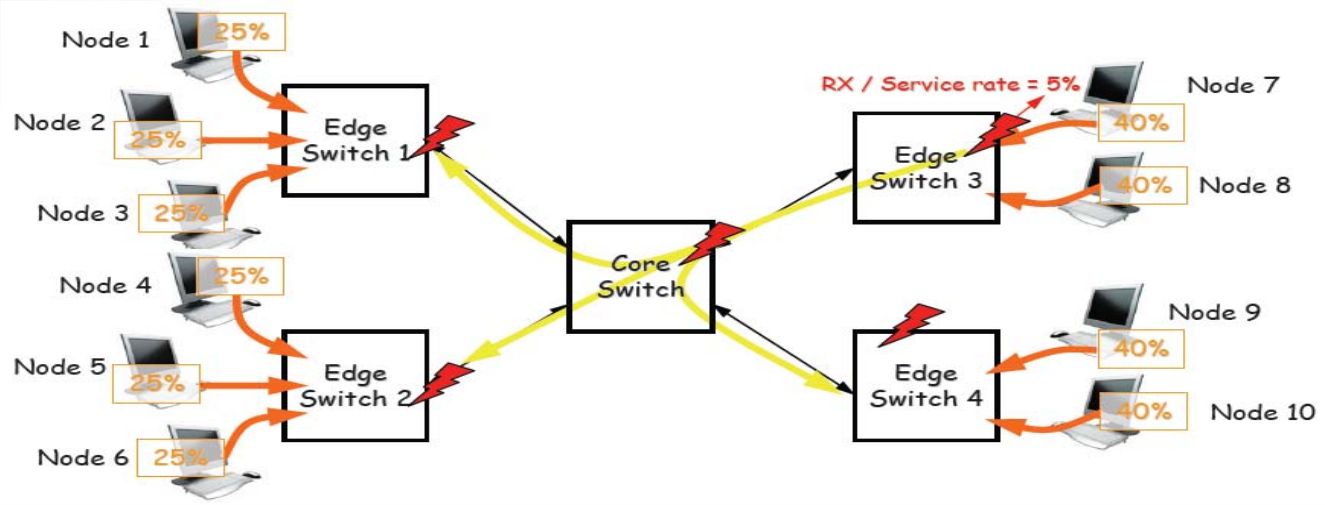
BROADCOM. Connecting everything

Parameters

- Switch Parameters
  - Core switch and edge switches are all 4 port switches
  - Buffer Size (B)
    - 600Kbytes/Port
    - 75Kbytes/Port
  - Shared Memory Switch Devices, total switch memory size = 4 * B
  - PAUSE Flow Control Settings
    - Applied per ingress port basis based on XON/XOFF thresholds
    - XOFF Threshold = B – RTT*BW
    - XON Threshold = B/2
- BCN Parameters
  - Frame Sampling
    - Frames are periodically sampled (on avg) every 75KB (2%)
  - W = 2
  - Qeq = B/4
  - Ru = 1Mbps
  - Gi (Initial)
    - Computed as (Linerate/10) * [1/((1+2*W)*Q_eq)]
    - Same as in baseline
  - Gd (Initial)
    - Computed as 0.5*1/((1+2*W)*Q_eq)
    - Same as in baseline
  - Other BCN Enhancements
    - BCN(Max)
    - OverSampling during severe congestion or consistent oversampling
    - No BCN(0,0)
    - No Self Increase

BROADCOM.
Connecting everything®

Desired Performance

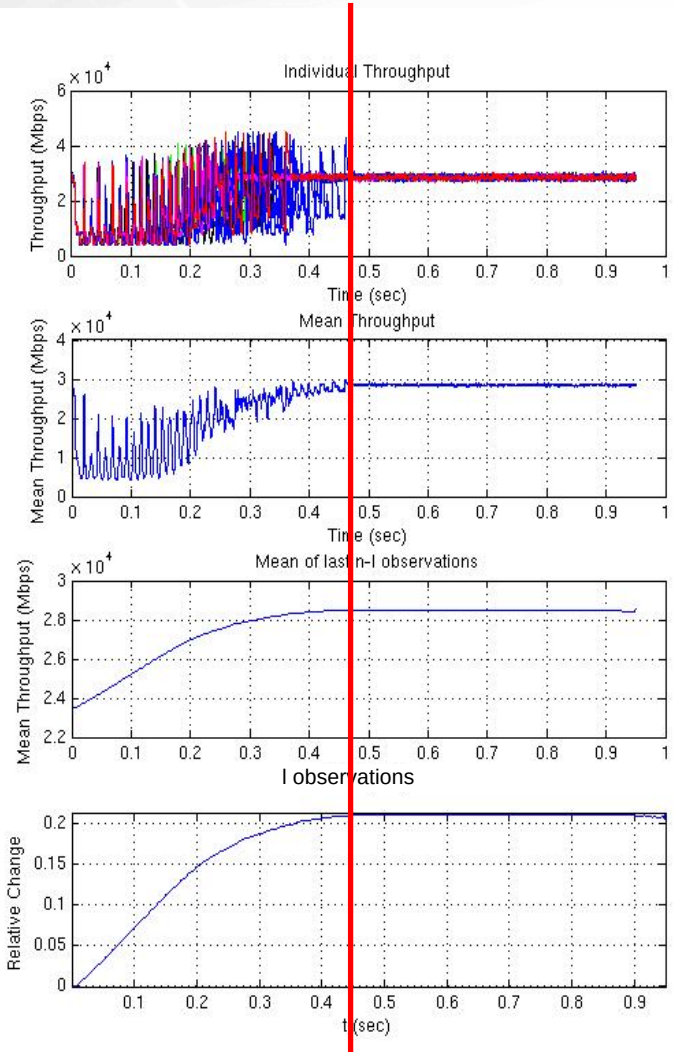


- Throughput performance characterization simulations may be found:
  - http://www.ieee802.org/1/files/public/docs2007/au-sim-ding-kwan-multihop-output-generated-010407v1.pdf
- Observations based on *aggregate* throughput (in the system)
  - Aimed to use a single metric to also take into account the innocent flows settling into a steady state.
  - Queue size at CP is not necessarily in steady state based on this metric
- Transient Duration Target
  - Collected aggregate throughput over time across 10 replications (w/ different seeds) using Initial Data Deletion* method
  - 10 x RTT
  - Qeq = 600Kbytes * ¼ = 150kbytes
  - Service Rate (@ congested port) = 500Mbps
  - RTT (Propagation Delay Only) = 16us
  - Total RTT = 16us + 2.5ms ~ 2.5ms
  - Target Transient Duration = 10 x RTT = 25ms

**Raj Jain "The Art of Computer Systems Performance Analysis", 1991, Pg 424*

BROADCOM. Connecting everything

Defining Transient Duration



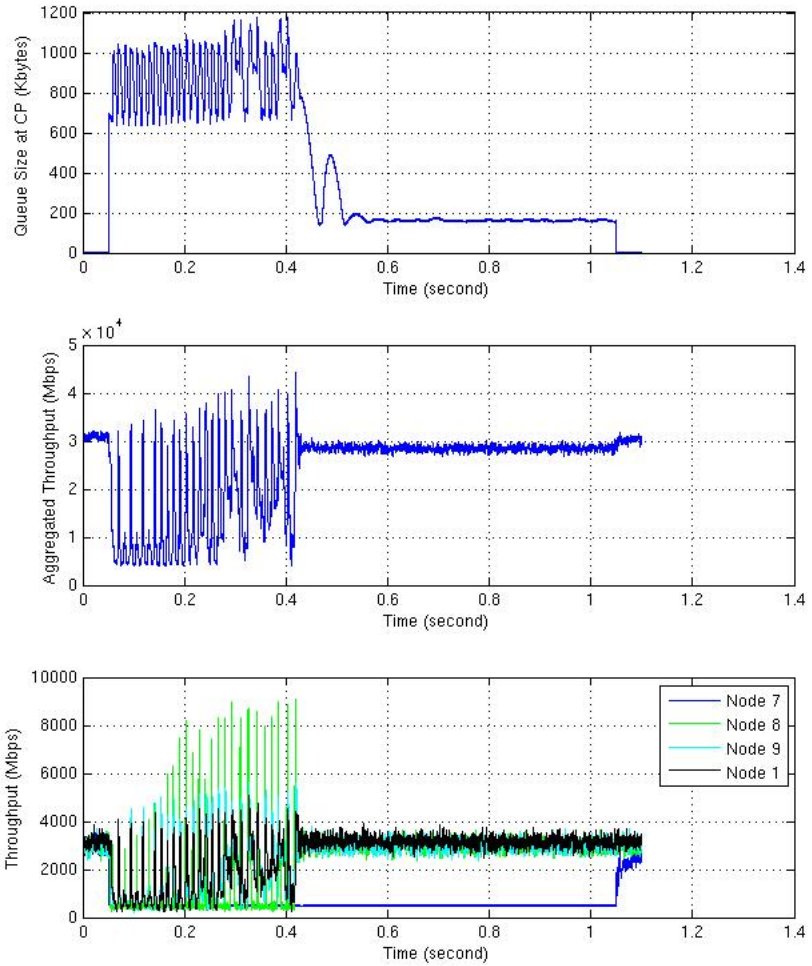
- Use *Initial Data Deletion** method to determine when steady state is achieved.
  - Study an averaged version across multiple replications.
- Computation of Transient Duration
  - Collect multiple simulation replications (i.e. 10) and compute the time series average across the samples.
  - Obtain the overall mean throughout the sample run
  - Obtain an overall mean from the remaining n-l values as a function of l
  - Compute the relative change in the overall mean. The knee of this curve is where the transient duration ends
- Target Transient Duration
  - 10 x RTT

**Raj Jain "The Art of Computer Systems Performance Analysis", 1991, Pg 424*

Overview

- Goals
- Observations
- Simulation Setup
- Results
- Summary & Next Steps

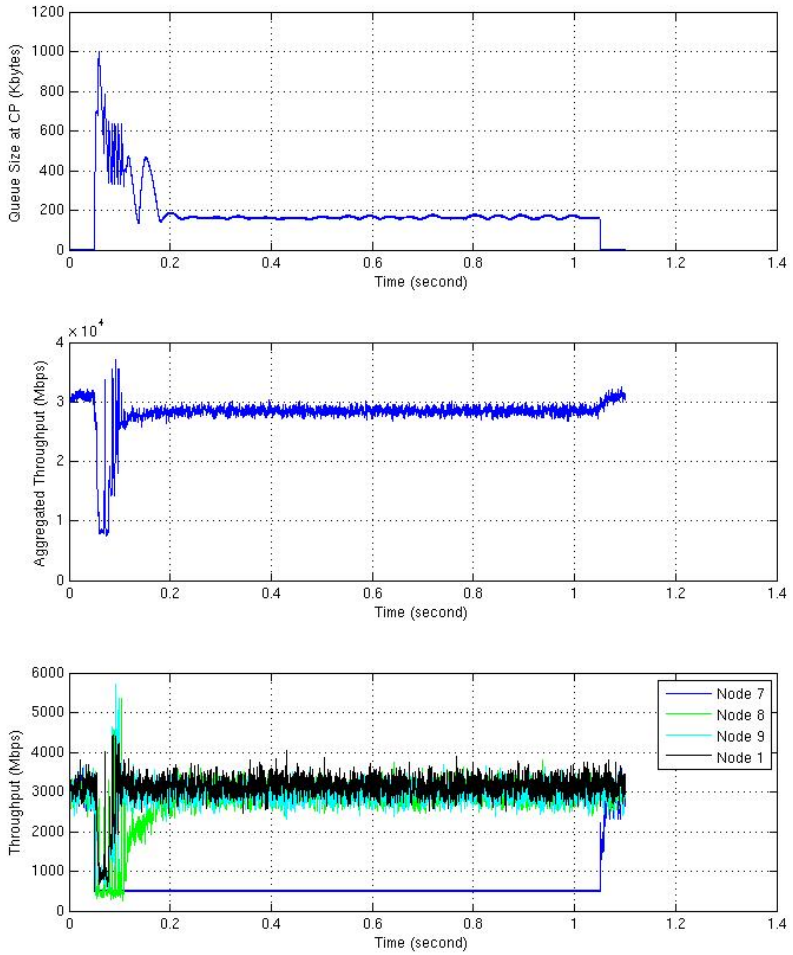
Basic BCN + PAUSE
(No BCN-Max, No Oversampling)



Multi Hop Output Generated Scenario (Single Replication)

- Transient Duration
  - 462 ms (184 x RTT)
- Next: Consider BCN-MAX

BCN w/BCN-MAX + PAUSE (No Oversampling)



Multi Hop Output Generated Scenario (Single Replication)

- Transient Duration
  - 52 ms (21 x RTT)
- Next: Consider additional severe congestion enhancements

Transient Duration Results (Effects of Qoffset/Qdelta Range)

- Setup
  - Buffer Size = 600KB/port
  - Sampling Rate = 2%

Case	BCN (Max)	Qoffset & Qdelta Range	Transient Duration
1	No	Qoffset: (-Qeq,+Qeq) Qdelta: (-2Qeq, +2Qeq)	462 ms (184 x RTT)
2	Yes		52ms (21 x RTT)
3	No	Qoffset: (-Qeq,+2Qeq) Qdelta: (-4Qeq, +4Qeq)	170 ms (68 x RTT)
4	Yes		274 ms (109 x RTT)
5	No	Qoffset: (-Qeq,+2Qeq) Qdelta: (-2Qeq, +2Qeq)	108ms(43 x RTT)
6	Yes		62ms(25 x RTT)

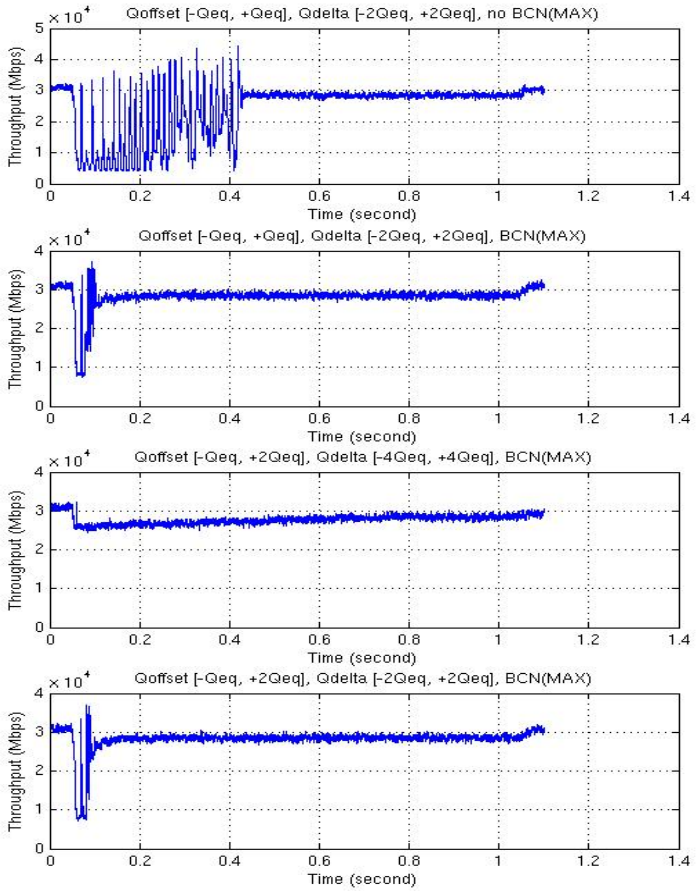
Transient duration is not further improved by increasing the range of Qoffset (nor Qdelta).

BROADCOM. Connecting everything®

Effects of Qoffset/Qdelta Range (Aggregate Throughput)

- Set up
  - Buffer Size 600KB/port
  - Sampling Rate = 2%

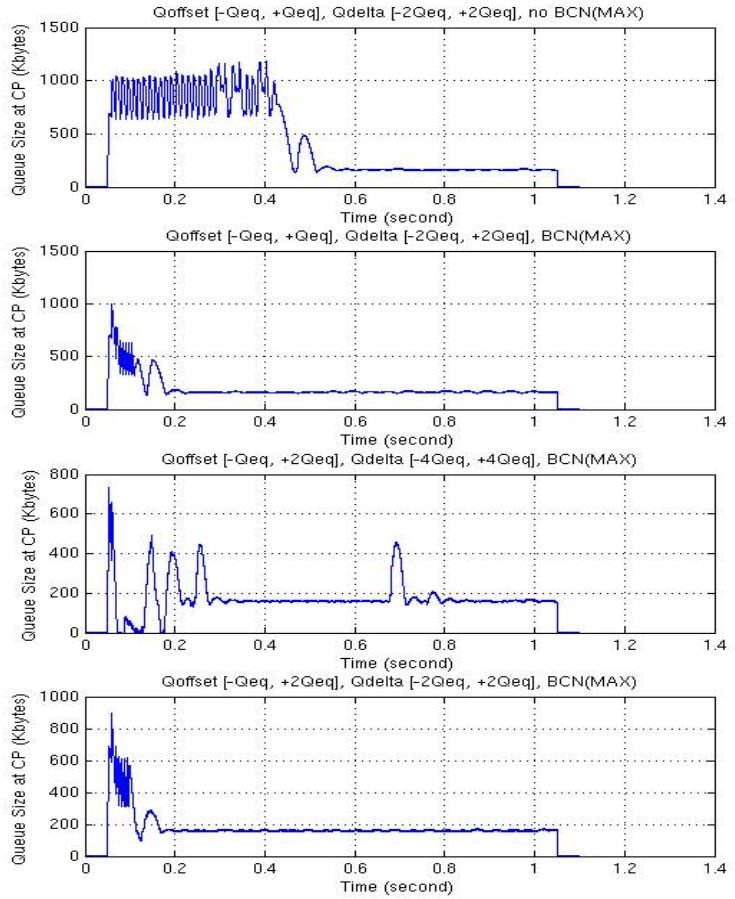
Case	BCN (Max)	Qoffset & Qdelta Range	Transient Duration
1	No	Qoffset: (-Qeq,+Qeq) Qdelta: (-2Qeq, +2Qeq)	462 ms (184 x RTT)
2	Yes	Qoffset: (-Qeq,+Qeq) Qdelta: (-2Qeq, +2Qeq)	52ms (21 x RTT)
4	Yes	Qoffset: (-Qeq,+2Qeq) Qdelta: (-4Qeq, +4Qeq)	274 ms (109 x RTT)
6	Yes	Qoffset: (-Qeq,+2Qeq) Qdelta: (-2Qeq, +2Qeq)	62ms(25 x RTT)



Effects of Qoffset/Qdelta Range (Queue Behavior @ CP)

- Set up
  - Buffer Size 600KB/port
  - Sampling Rate = 2%

Case	BCN (Max)	Qoffset & Qdelta Range	Transient Duration
1	No	Qoffset: (-Qeq,+Qeq) Qdelta: (-2Qeq, +2Qeq)	462 ms (184 x RTT)
2	Yes		52ms (21 x RTT)
4	Yes	Qoffset: (-Qeq,+2Qeq) Qdelta: (-4Qeq, +4Qeq)	274 ms (109 x RTT)
6	Yes	Qoffset: (-Qeq,+2Qeq) Qdelta: (-2Qeq, +2Qeq)	62ms(25 x RTT)



Transient Duration Results (Effects of Sampling Behavior)

- Setup
  - Buffer Size = 600KB/port
  - Qoffset: (-Qeq, +Qeq)
  - Qdelta: (-2Qeq, +2Qeq)

Case	BCN (Max)	Sampling	Transient Duration
1	No	2%	462 ms (184 x RTT)
2	Yes	2%	52ms (or 21 x RTT)
3	Yes	2% (Qlen < Qsc) *10% (Qlen >= Qsc)*	86 ms (34 x RTT)
4	Yes	2% (Qlen < Qsc) 20% (Qlen >= Qsc)	144 ms (58 x RTT)
5	Yes	2% (Qlen < Qsc) 30% (Qlen >= Qsc)	96 ms (38 x RTT)
6	Yes	5%	52ms (21 x RTT)
7	Yes	10%	38 ms (15 x RTT)

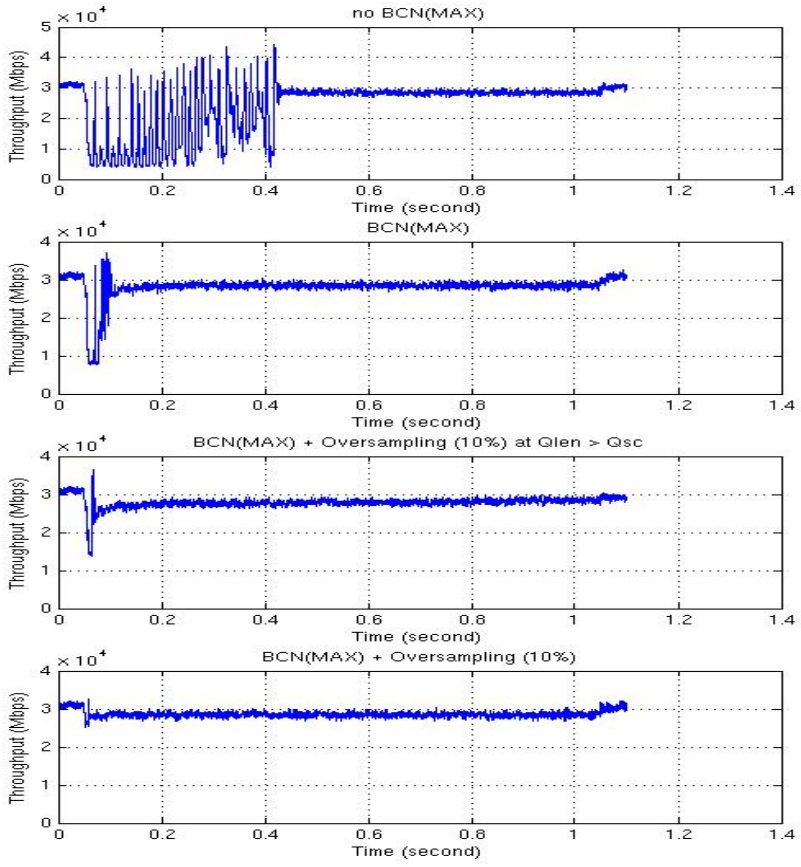
An overall increase in the sampling rate aids in reducing the transient duration down to 38ms. Need to consider also oversampling when Qlen = 0 to improve the adaptive oversampling scheme.

BROADCOM.
Connecting everything®

Effects of Sampling Behavior (Aggregate Throughput)

- Set up
  - Buffer Size = 600KB/Port
  - Qoffset = (-Qeq, +Qeq)
  - Qdelta = (-2Qeq, +2Qeq)

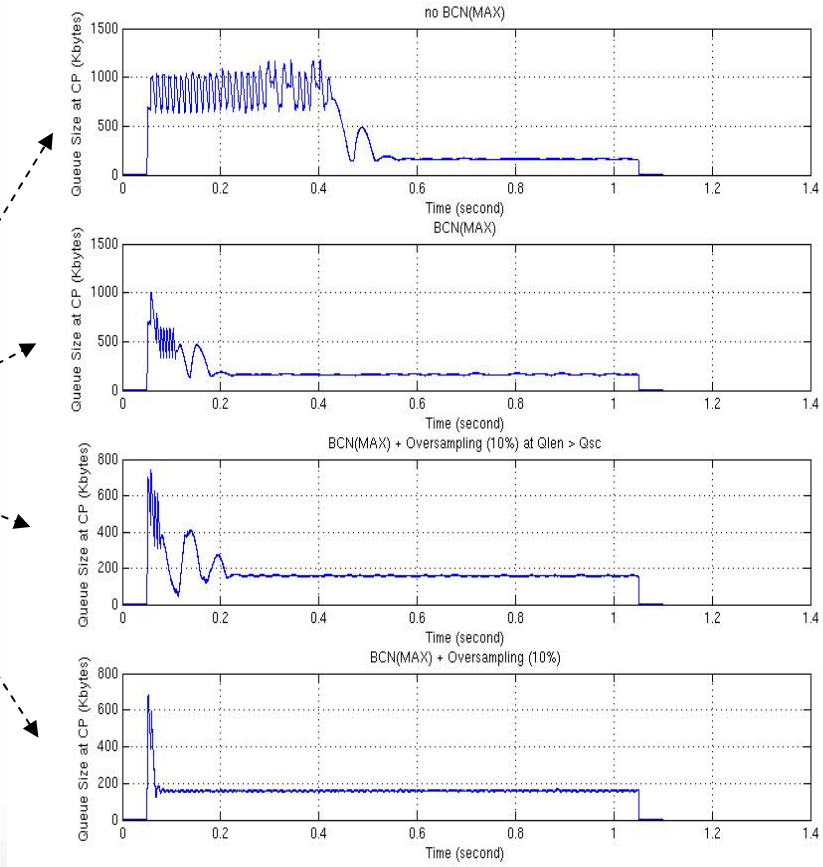
Case	BCN (Max)	Sampling	Transient Duration
1	No	2%	462 ms (184 x RTT)
2	Yes	2%	52ms (or 21 x RTT)
3	Yes	2% (Qlen < Qsc) *10% (Qlen >= Qsc)*	86 ms (34 x RTT)
7	Yes	10%	38 ms (15 x RTT)



Effects of Sampling Behavior (Queue Behavior @ CP)

- Set up
  - Buffer Size = 600KB/Port
  - Qoffset = (-Qeq, +Qeq)
  - Qdelta = (-2Qeq, +2Qeq)

Case	BCN (Max)	Sampling	Transient Duration
1	No	2%	462 ms (184 x RTT)
2	Yes	2%	52ms (or 21 x RTT)
3	Yes	2% (Qlen < Qsc) *10% (Qlen >= Qsc)*	86 ms (34 x RTT)
7	Yes	10%	38 ms (15 x RTT)



Effects of Buffer Size

- Setup
  - Qeq = Queue Size * 1/4
  - Qoffset: (-Qeq, +Qeq)
  - Qdelta: (-2Qeq, +2Qeq)

Case	BCN (Max)	Sampling	Transient Duration	
			Queue Size = 600 KBytes (Qeq = 150KBytes)	Queue Size = 75 Kbytes (Qeq = 18.75KBytes)
1	No	2%	462 ms (184 x RTT)	264ms (880 x RTT)
2	Yes	2%	52ms (or 21 x RTT)	56ms (187 x RTT)
3	Yes	2% (Qlen < Qsc) *10% (Qlen >= Qsc)*	86 ms (34 x RTT)	42ms (140 x RTT)
4	Yes	2% (Qlen < Qsc) 20% (Qlen >= Qsc)	144 ms (58 x RTT)	52ms (173 x RTT)
5	Yes	2% (Qlen < Qsc) 30% (Qlen >= Qsc)	96 ms (38 x RTT)	92ms (306 x RTT)
6	Yes	5%	52ms (21 x RTT)	26 ms (87 x RTT)
7	Yes	10%	38 ms (15 x RTT)	20 ms (67 x RTT)

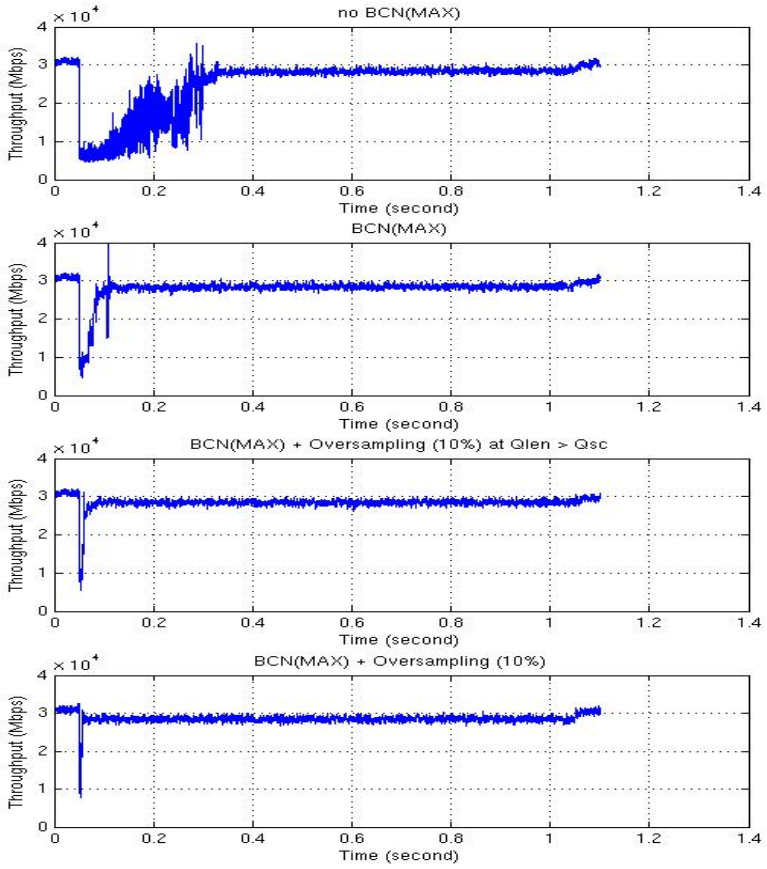
With small buffers, more challenging to minimize transient duration

BROADCOM. Connecting everything

Effects of Buffer Size (Aggregate Throughput)

- Set up
  - Buffer Size = 75KB
  - Qoffset: (-Qeq, +Qeq)
  - Qdelta: (-2Qeq, +2Qeq)

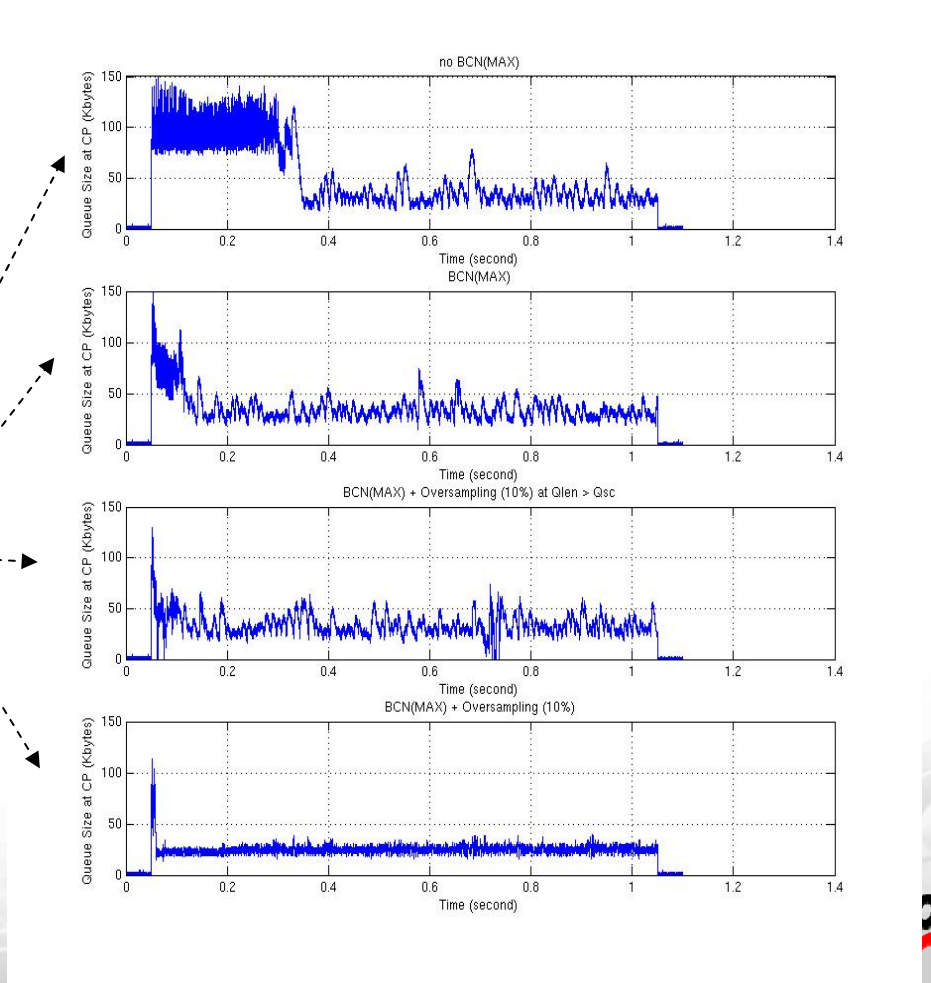
Case	BCN (Max)	Sampling	Transient Duration
1	No	2%	264ms (880 x RTT)
2	Yes	2%	56ms (187 x RTT)
3	Yes	2% (Qlen < Qsc) *10% (Qlen >= Qsc)*	42ms (140 x RTT)
7	Yes	10%	20 ms (67 x RTT)



Effects of Buffer Size (Queue Behavior @ CP)

- Set up
  - Buffer Size = 75KB
  - Qoffset: (-Qeq, +Qeq)
  - Qdelta: (-2Qeq, +2Qeq)

Case	BCN (Max)	Sampling	Transient Duration
1	No	2%	264ms (880 x RTT) *(Heavy Oscillations)*
2	Yes	2%	56ms (187 x RTT)
3	Yes	2% (Qlen < Qsc) *10% (Qlen >= Qsc)*	42ms (140 x RTT)
7	Yes	10%	20 ms (67 x RTT)



Overview

- Goals
- Observations
- Simulation Setup
- Results
- Summary & Next Steps

BROADCOM.
Connecting everything®

Summary

- Without additional severe congestion schemes, the transient duration for the multi-hop output generated scenario can be on the order of 462 ms (184 x RTT).
- With the use of one of the simple severe congestion management schemes (i.e. BCN-MAX), the transient duration for the multi-hop output generated scenario can be reduced to 52 ms (21 x RTT).
- Small buffer sizes can hamper transient duration (when measured in terms of RTT). Requires additional study.

BROADCOM.
Connecting everything®

Next Steps

- Refine metric to quantify the response time of the CN mechanism.
- Collect input from the group on identifying a target response time for the traffic of interest.
- Quantify benefits of oversampling when Qlen = 0. Measure actual amount of BCN messaging control bandwidth.
- Additional study required of how BCN operates under small buffer conditions at the switch.

BROADCOM.
Connecting everything®