Effects of Delay: Output Generated Multistage and Symmetric Topology w/ Single Hot Spot Scenarios

Bruce Kwan & Jin Ding

February 15, 2007

Overview

- Goals & System Parameters
- Output Generated Hotspot Multi-Stage
- Symmetric Topology Single HS – Non Bursty
- Conclusions & Next Steps

BROADCOM
Connecting everything®

Goals

- Characterize performance across the target range of control loop delay as specified in the IEEE 802.1Qau PAR

- March 14, 2006 PAR
  - "The bandwidth-delay product limit is expected to be in the region of 1-5Mbits (*100-500us* control loop delay for 10Gbps network) and simulation and analysis will verify performance characteristics up to the advertised bandwidth-delay product."

BROADCOM
Connecting everything

Base System Parameters

- No PAUSE
  - Goal is to characterize BCN behavior without PAUSE
- Switch Parameters
  - Buffer Size (B)
    - 600Kbytes/Port.
- 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
  - BCN(Max) Enabled
  - Other BCN Enhancements
    - No Oversampling
    - No BCN(0,0)
    - No Self Increase

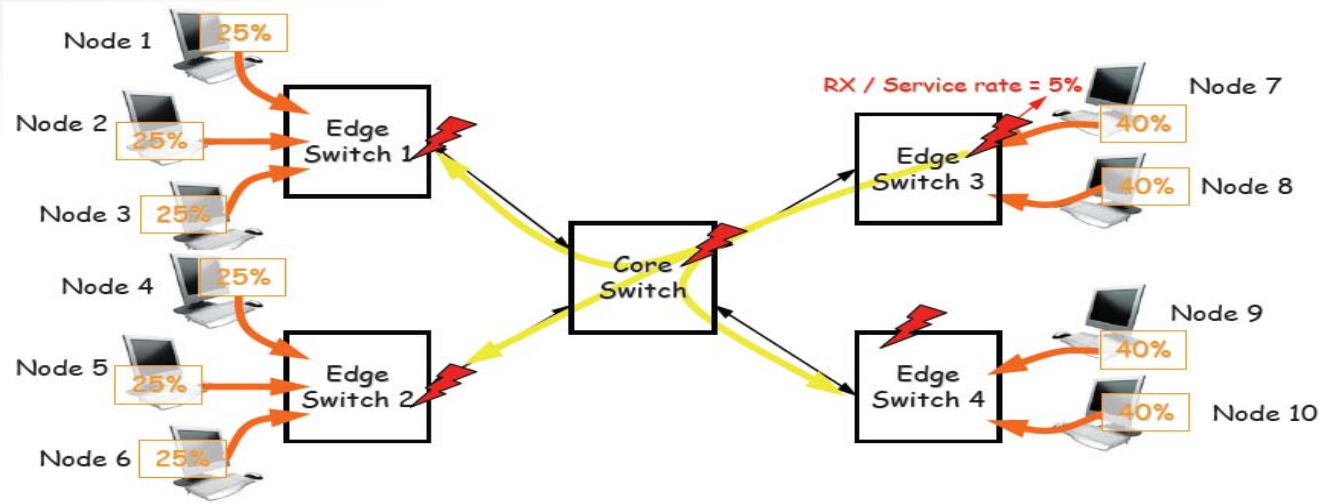
BROADCOM
Connecting everything

Overview

- System Parameters
- Output Generated Hotspot Multi-Stage
- Symmetric Topology Single HS – Non Bursty
- Conclusions & Next Steps

BROADCOM
Connecting everything®

OG HS Multi-Stage (Required Scenario #2)

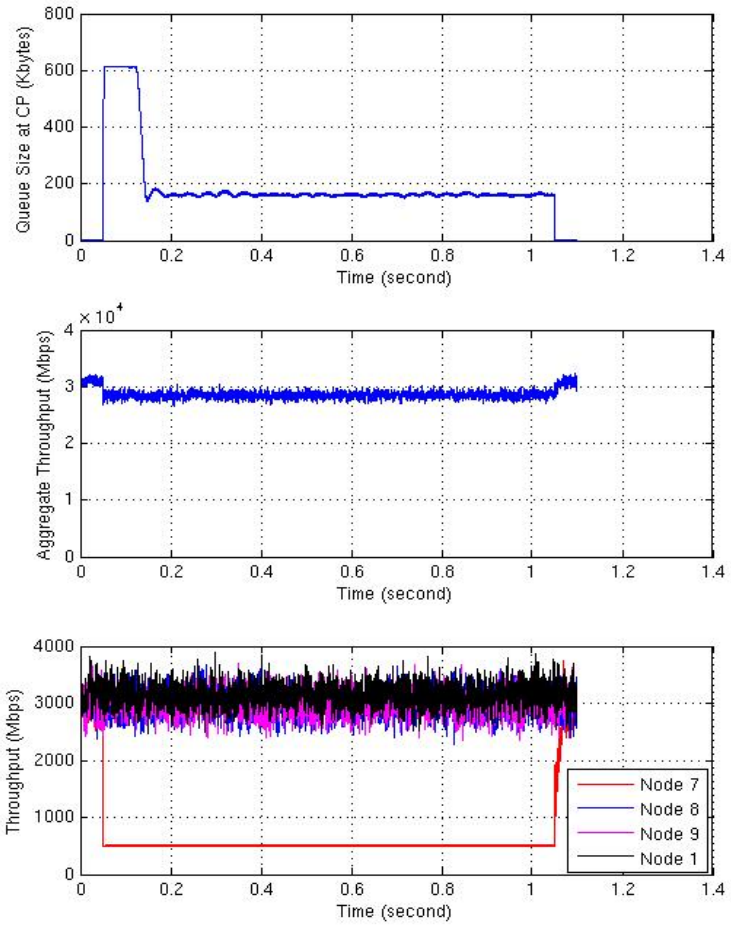


- Multi-stage Output-Generated Hotspot Scenario
  - Link Speed = 10Gbps for all links
- Traffic Pattern
  - 100% UDP (or Raw Ethernet) Traffic
  - Destination Distribution: Uniform distribution to all nodes (except self)
  - Frame Size Distribution: Fixed length (1500bytes) frames
  - Offered Load
    - 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.*

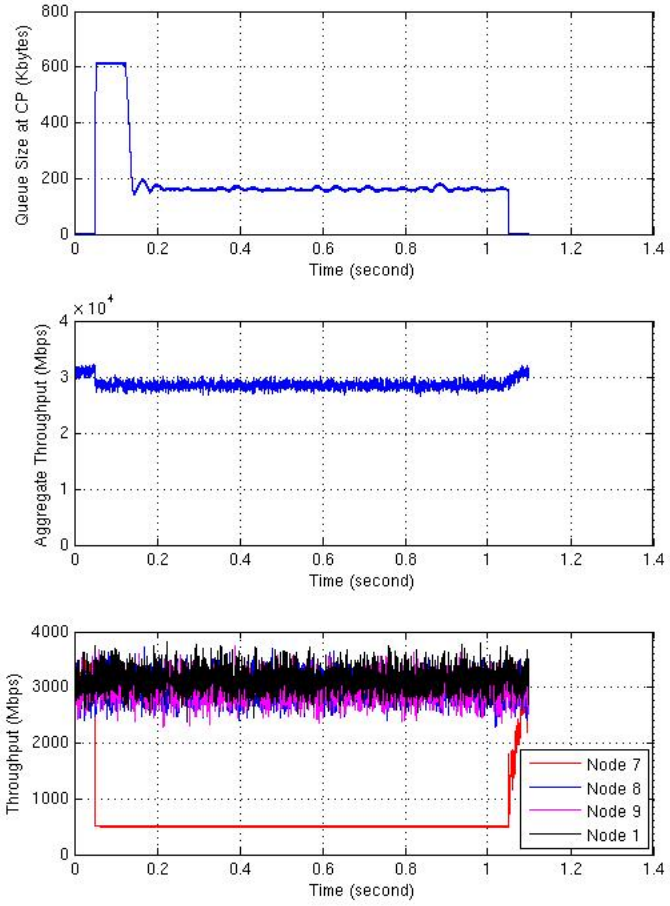
BROADCOM
Connecting everything

OG HS Multi-Stage Results
Delay = 12us



- Maximum control loop delay
  - 12us
- Observations
  - Number of drops = 9164

OG HS Multi-Stage Results
Delay = 600us



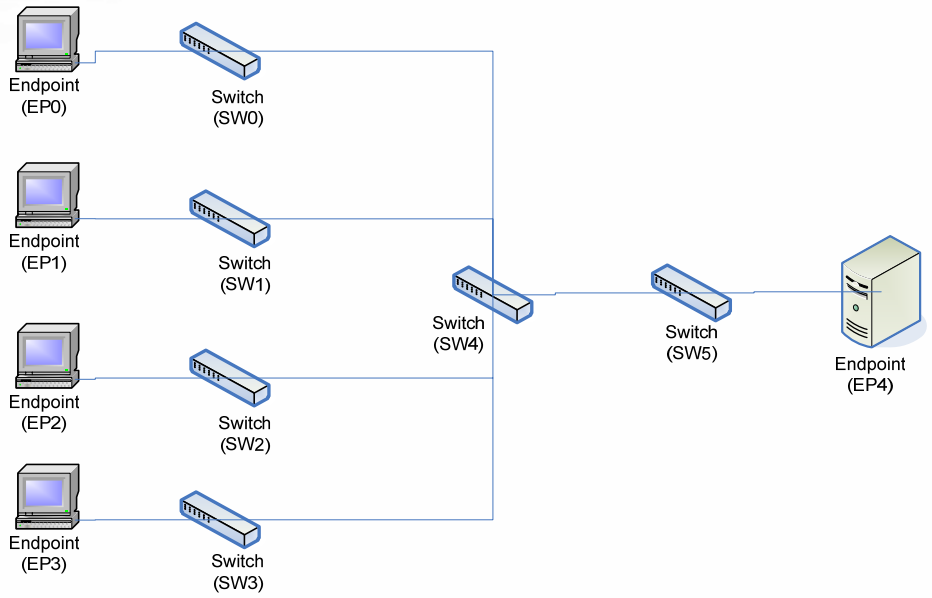
- Maximum control loop delay
  - 600us
- Observations
  - Number of drops = 9059
    - Requires further study to reduce this further if possible (assuming PAUSE is not active)
  - Degree of oversubscription is mild in this scenario
  - Impact on overall behavior is minimal despite increased loop delay

Overview

- System Parameters
- Output Generated Hotspot Multi-Stage
- Symmetric Topology Single HS – Non Bursty
- Conclusions & Next Steps

BROADCOM
Connecting everything®

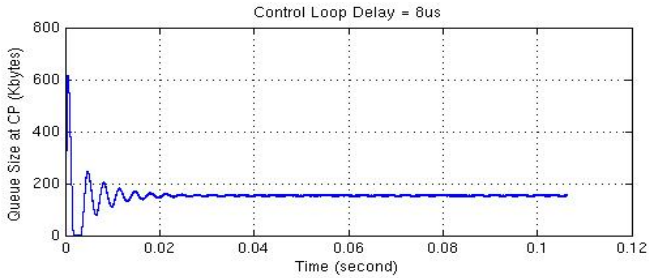
Symmetric Topology Single HS – Non Bursty (Similar to Required Scenario #5)

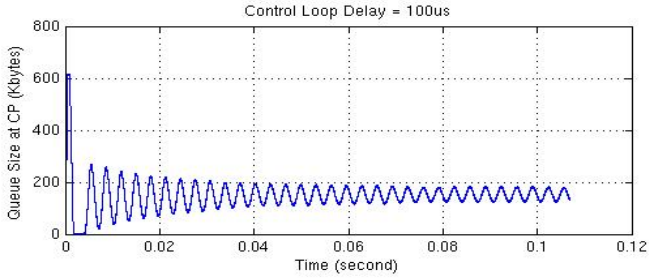


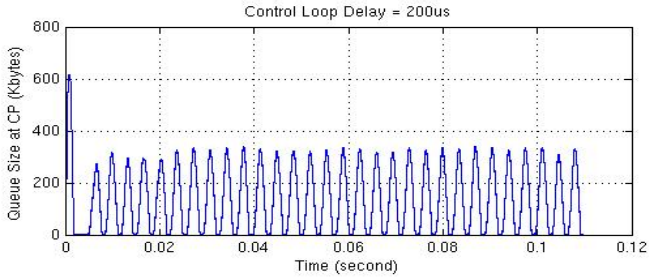
- Symmetric Topology Single HS
  - Link speed : 10Gbps for all links
- Traffic Pattern
  - Traffic Type: 100% UDP (or Raw Ethernet) Traffic
  - Destination Distribution: EP0-EP3 send to EP4
  - Frame Size Distribution: Fixed length (1500 bytes) frames
  - Arrival Distribution: Bernoulli temporal distribution
  - Offered Load/Endpoint = 50%
- Control Loop Delay is between the source Endpoints and the Congestion Point

BROADCOM
Connecting everything

Symmetric Topology Single HS – Non Bursty
Effects of Control Loop Delay (Queue Size @ CP)







Ctrl Loop Delay	# of drops	Throughput
8us	6	9.992Gbps
100us	106	9.982Gbps
200us	185	9.775Gbps

Mild throughput degradation and increase in frame drops as control loop delay increases. May result in increased transaction completion time variance. Requires further data collection to better characterize.

BROADCOM
Connecting everything

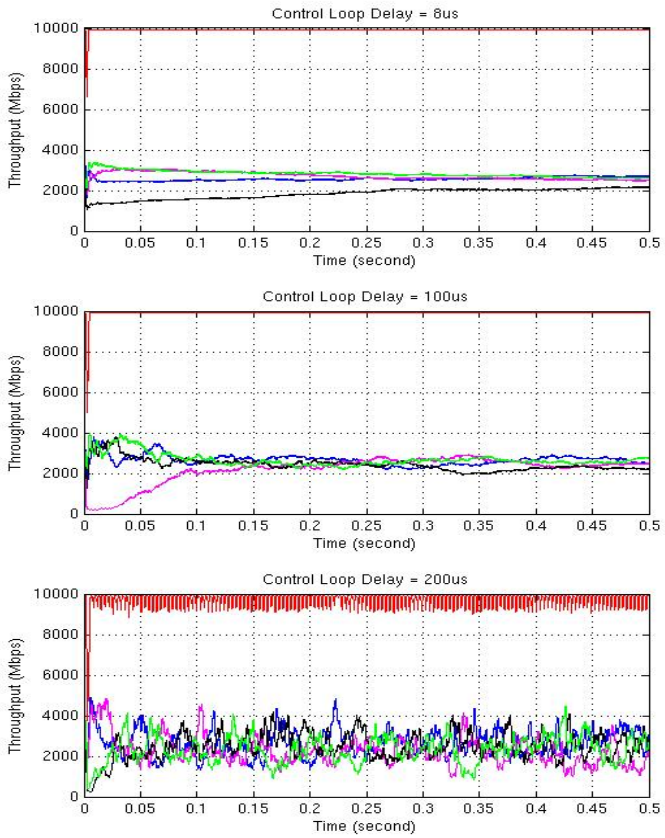
Symmetric Topology Single HS – Non Bursty Effects of Control Loop Delay (Throughput)

Ctrl Loop Delay

8us

100us

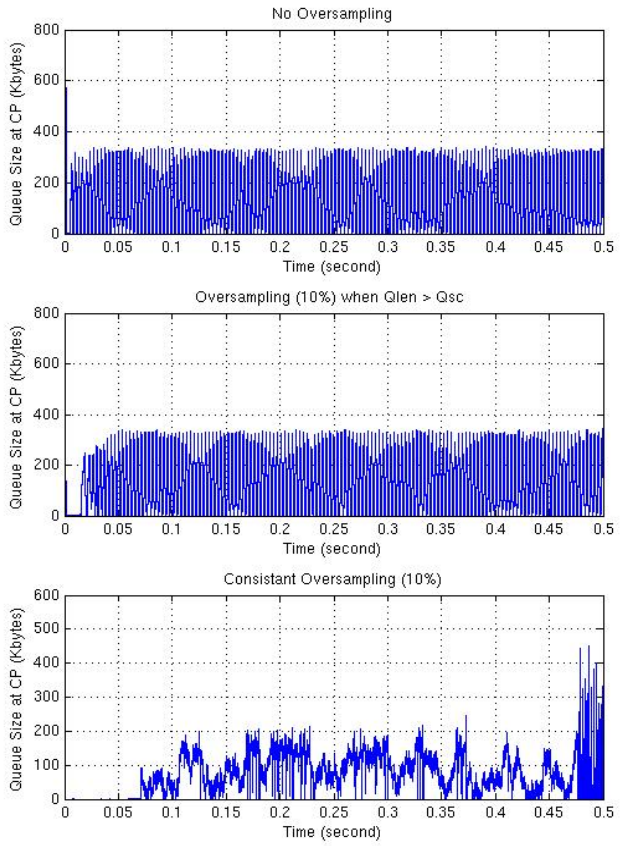
200us



Underutilization occurs as control loop delay increases to 200us.

BROADCOM.
Connecting everything®

Symmetric Topology Single HS – Non Bursty
Effects of Oversampling (Delay = 200us)
Queue Size

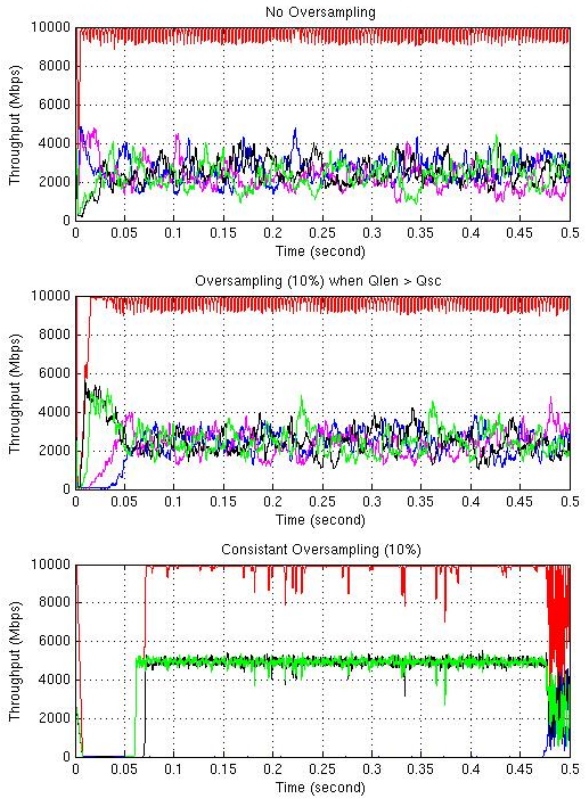


Oversampling	# of drops	Throughput
None	185	9.775Gbps
10% [Qlen > Qsc]	116	9.643Gbps
10% [always]	0	8.535Gbps

Increased the runtime.

Oversampling appears to be generating a stronger decrease which results in fewer packet drops but leads to further underutilization.

Symmetric Topology Single HS – Non Bursty Effects of Oversampling (Delay = 200us) Throughput



Oversampling

None

10% [Qlen > Qsc]

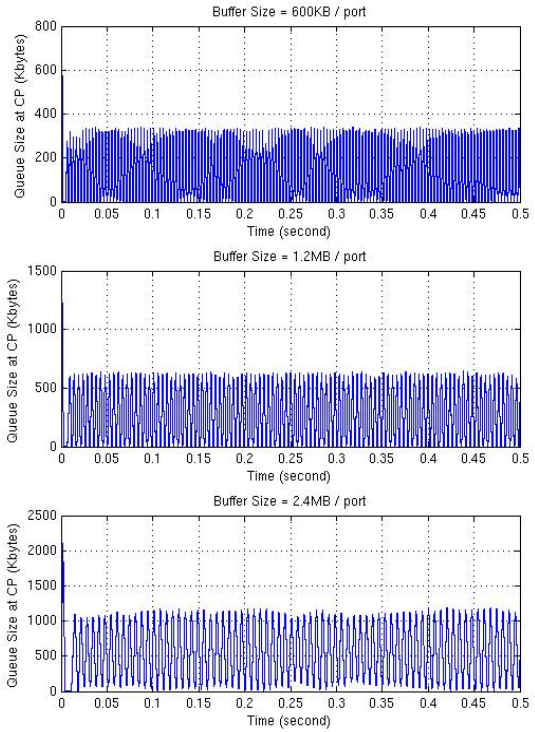
10% [always]

2 of the 4 flows have near 0 throughput in this specific run.

An abundance of negative BCN messages can result in unfair bandwidth distribution & underutilization.

BROADCOM
Connecting everything

Symmetric Topology Single HS – Non Bursty
Effects of Buffer Size (Delay = 200us)
Queue Size

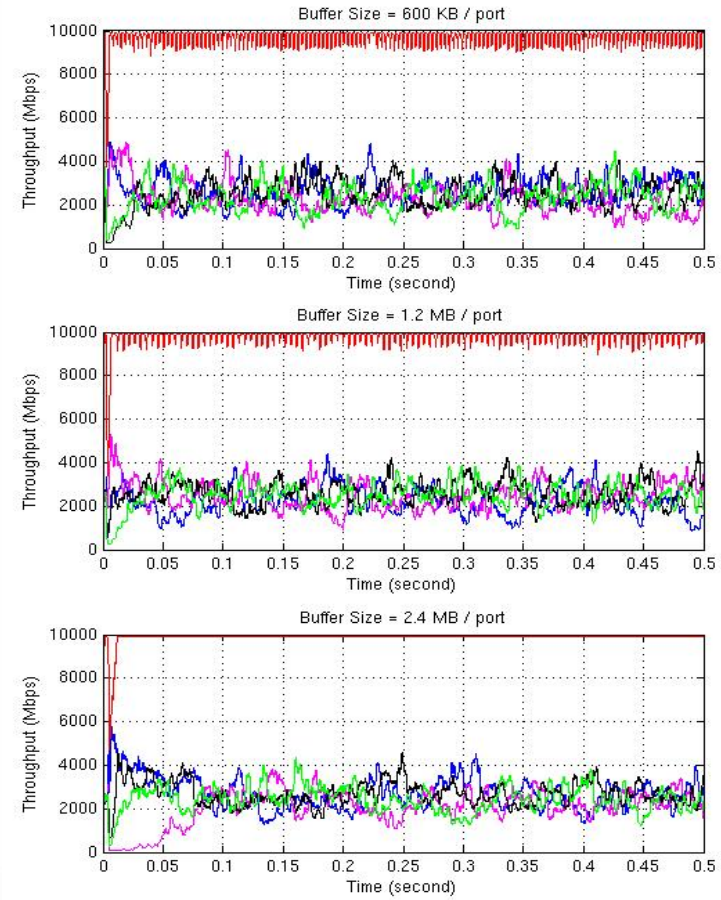


Buffer Size Per Queue (Kbytes)	# of drops	Throughput
600 Kbytes	185	9.775Gbps
1.2 Mbytes	43	9.855Gbps
2.4 Mbytes	0	9.955Gbps

2.4Mbytes per egress port required to avoid underutilization (i.e. queue hitting zero) for this scenario and baseline BCN parameters. Still challenging to control the queue. May worsen transaction completion time variation as buffer size is increased. To quantify effects, requires further characterization.

BROADCOM
Connecting everything

Symmetric Topology Single HS – Non Bursty Effects of Buffer Size (Delay = 200us) Throughput



Buffer Size Per Queue (Kbytes)

600 Kbytes

1.2 Mbytes

2.4 Mbytes

BROADCOM
Connecting everything

Overview

- System Parameters
- Output Generated Hotspot Multi-Stage
- Symmetric Topology Single HS – Non Bursty
- Conclusions & Next Steps

BROADCOM
Connecting everything®

Conclusions

- For the Output Generated Multi-Stage scenario, increased delay of up to 600us does not degrade BCN's ability to control the queue and maintain good utilization.
  - Transient duration requires further reduction using severe congestion enhancements (i.e. oversampling) or parameter tuning.
- For the Single Hop calibration scenario, increasing the delay up to 200us degrades BCN's ability to control the queue and results in underutilization and increased frame drops
  - Oversampling did improve packet drops but degraded utilization.
  - Increasing the buffer size did remove underutilization issues but did not lead to improved control over the queue (i.e. increased variation)
- Scenarios with a high degree of oversubscription appear to be more susceptible to queue variations and underutilization issues.

BROADCOM
Connecting everything

Next Steps

- Confirm the current target bandwidth-delay (control loop delay) as specified in the PAR (i.e. 100-500us).
- Characterize the case where control loop delay is 500us for single hop scenario.
- Characterize the impact of the varying queue size when delay is long for single hop scenario in terms of the variance of the transaction completion time.
- Test other scenarios with respect to the target control loop delay.