Multi-hop Output Generated Hotspot Scenario: Preliminary Results

Bruce Kwan & Jin Ding

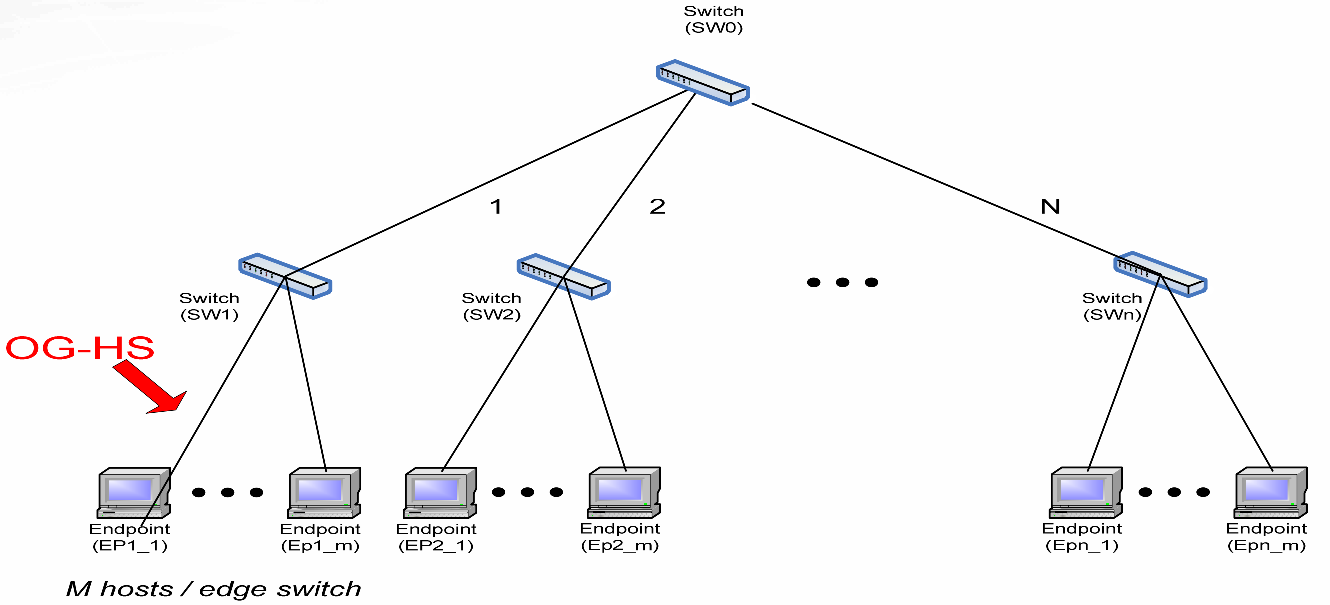
December 21, 2006

Overview

- Experiment Setup

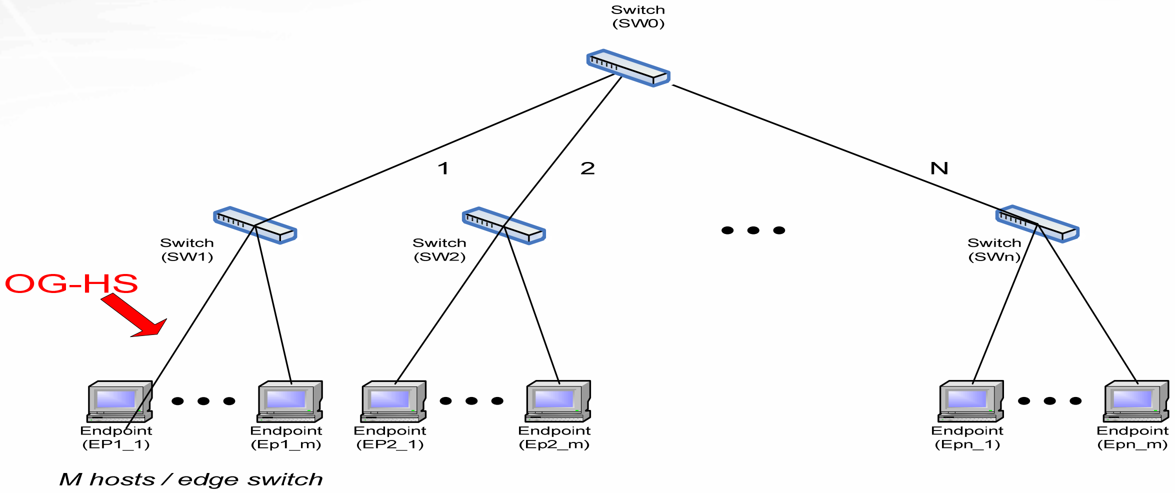
- Results

Topology



- Multi-stage Output-Generated Hotspot Scenario
- Link Speed = 10Gbps for all links
- Loop Latency = 8us
- N = Number of Edge Switches = 5
- M = Number of hosts per Edge Switch = 3

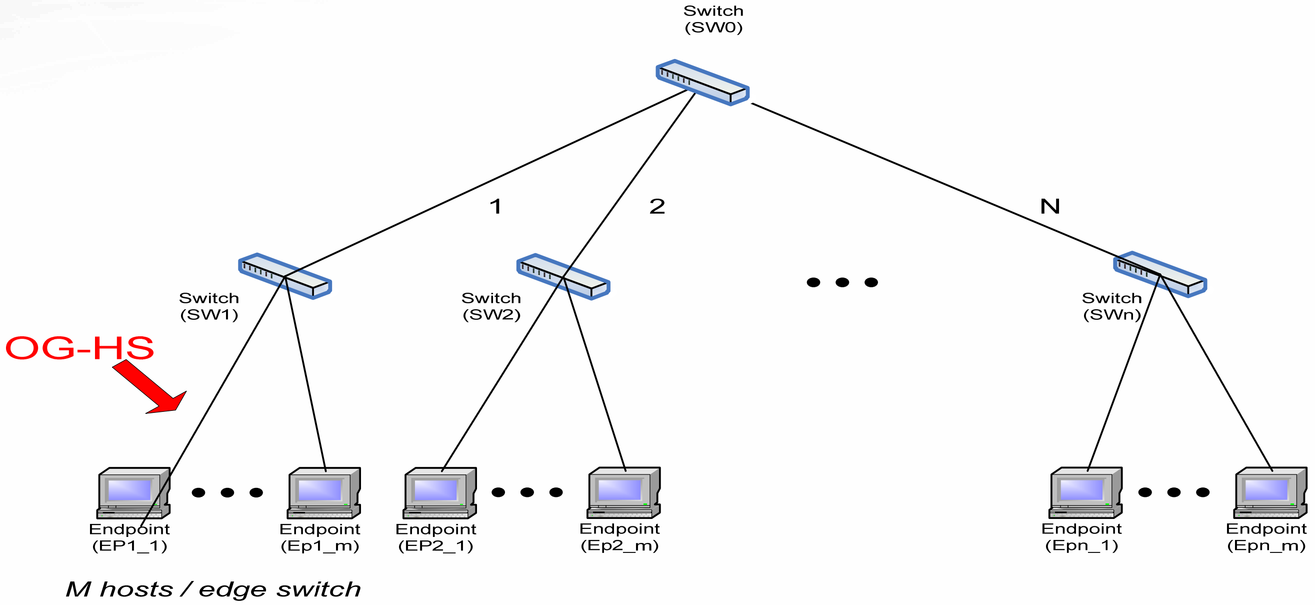
Workload



- 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 = 85% aggregate load across the hosts of an edge switch
    - For M=3, offered load per host is 85/3 = 28.3%
  - Arrival Distribution: Bernoulli temporal distribution

- Congestion Scenario
  - Output-generated hotspot (OG-HS)
  - Temporary reduction in service from 10Gbps to 2Gbps between [50ms, 450ms]
    - Mild congestion scenario

Switch Parameters



- Shared Memory Switch Devices
- Buffer Size (B) = 75Kbytes/Port, 150Kbytes/Port or 600Kbytes/Port
- PAUSE
  - Applied per ingress port basis based on XON/XOFF thresholds
  - XOFF Threshold = B – RTT*BW
  - XON Threshold = B/2

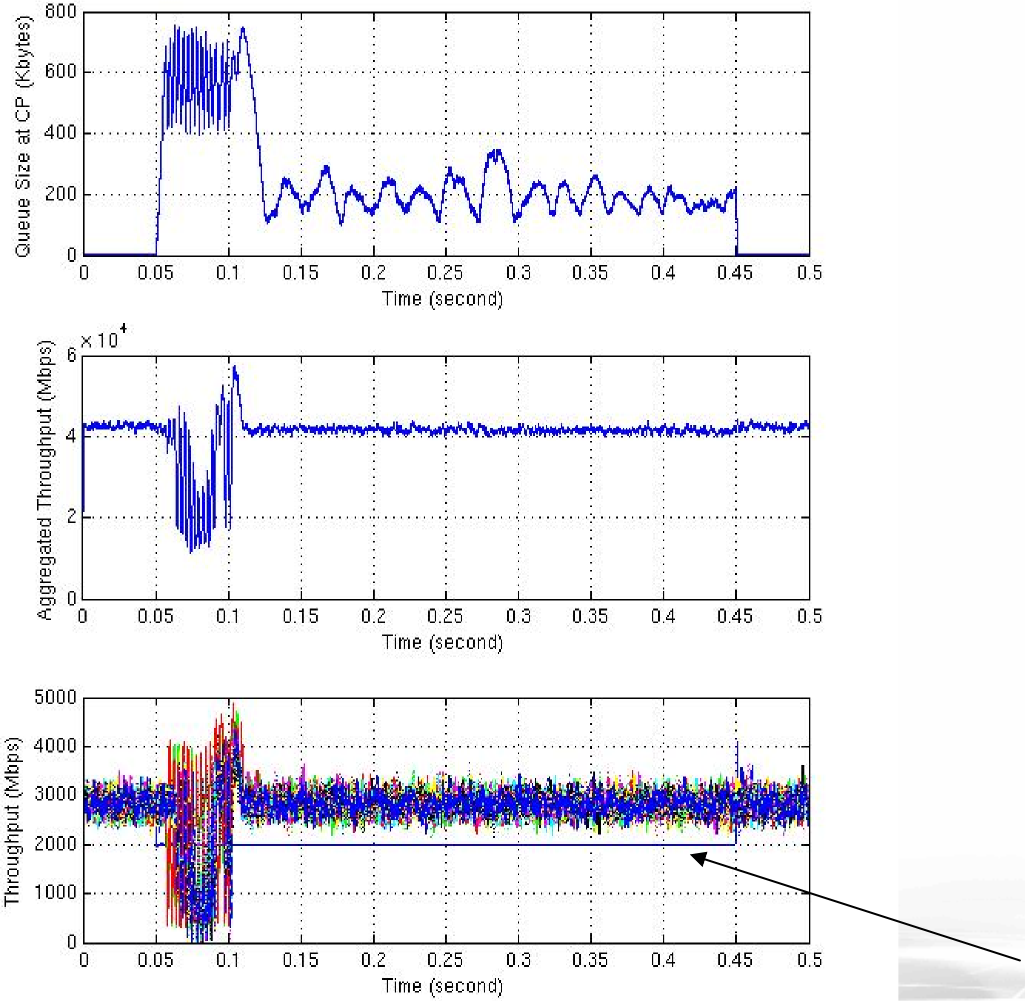
BCN Parameters

- Frame Sampling
  - Frames are periodically sampled (on avg) every 75KB (2%)
- W = 2
- Qeq = B/4
- Gi
  - Computed as Linerate/10 * (1/((1+2*W)*Q_eq
  - Same as in basline
    - Indicates maximum possible additive increase given constraints on Qoffset & Qdelta is 10% of linerate
- Gd
  - Computed as 0.5*1/((1+2*W)*Q_eq)
  - Same as in baseline
    - Maximum possible multiplicative decrease given constraints on Qoffset & Qdelta will be 50%
- Ru = 1Mbps
- Other BCN Enhancements
  - No BCN-MAX or BCN(0,0)
  - No Self Increase
  - No Oversampling during severe congestion

Overview

- Experiment Setup
- Results

With BCN and PAUSE B = 600Kbytes

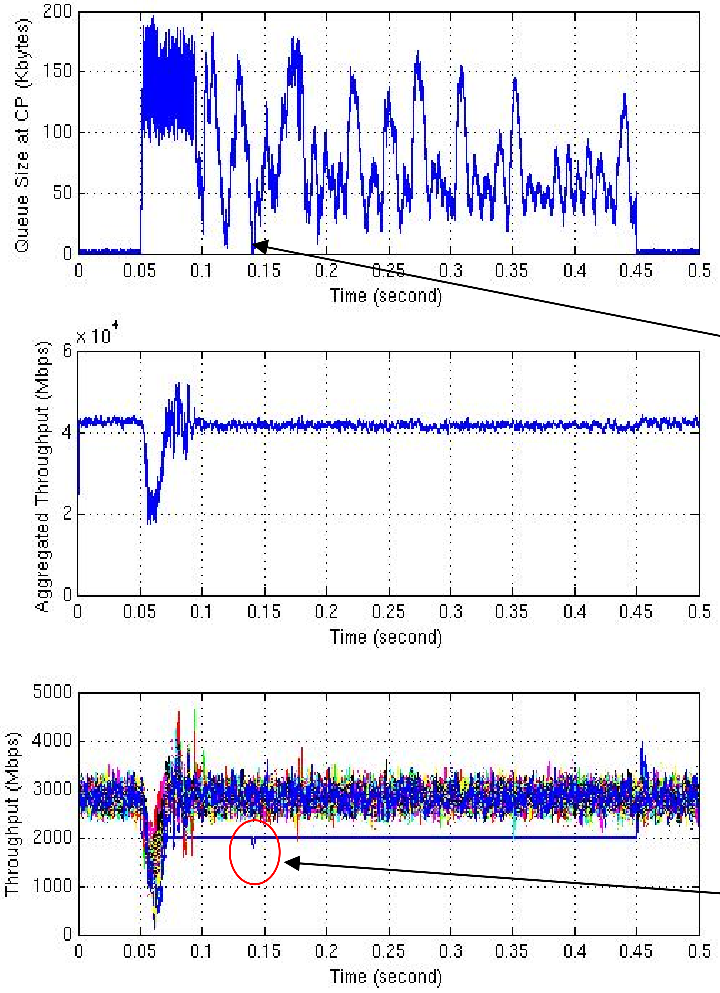


Throughput at every Edge Switch egress port

- Egress port memory size: 600kbytes
- Qeq = 150kbytes
- N = 5
- M = 3
- Gi = Linerate/10 * (1/((1+2*W)*Q_eq

Throughput at egress of congested switch

With BCN and PAUSE
B = 150Kbytes

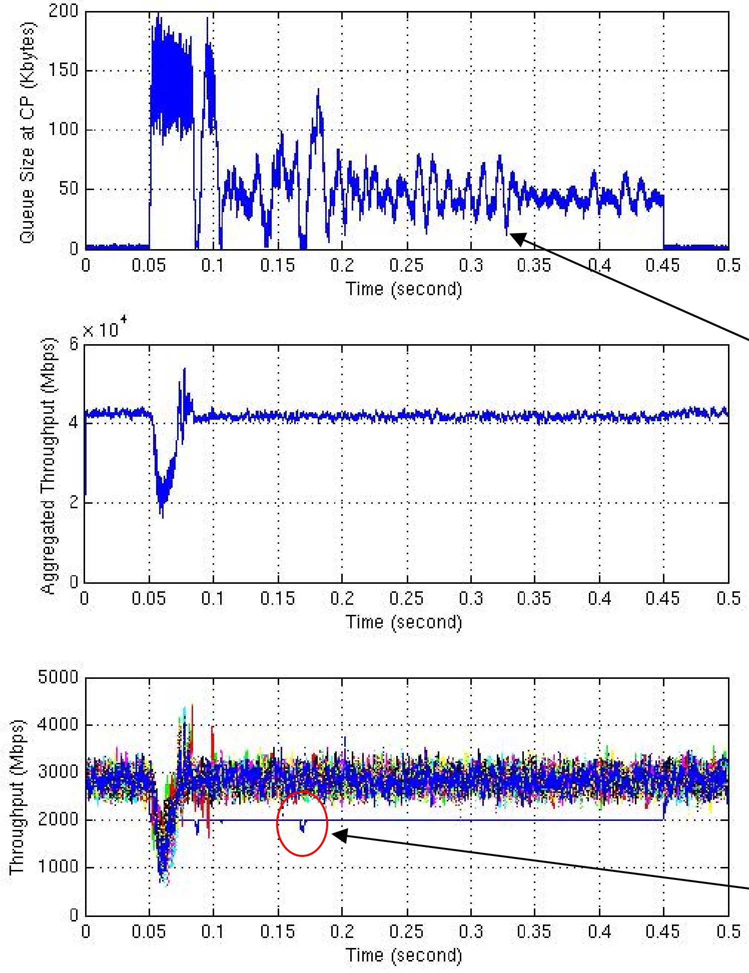


- Egress port memory size: 150Kbytes
- Qeq = 37.5Kbytes
- N = 5
- M = 3
- Gi = Linerate/10 * (1/((1+2*W)*Q_eq

Queue oscillations, some hitting zero at the start.

Moments of underutilization at the start.

With BCN and PAUSE
B = 150Kbytes (Reduced Gi)



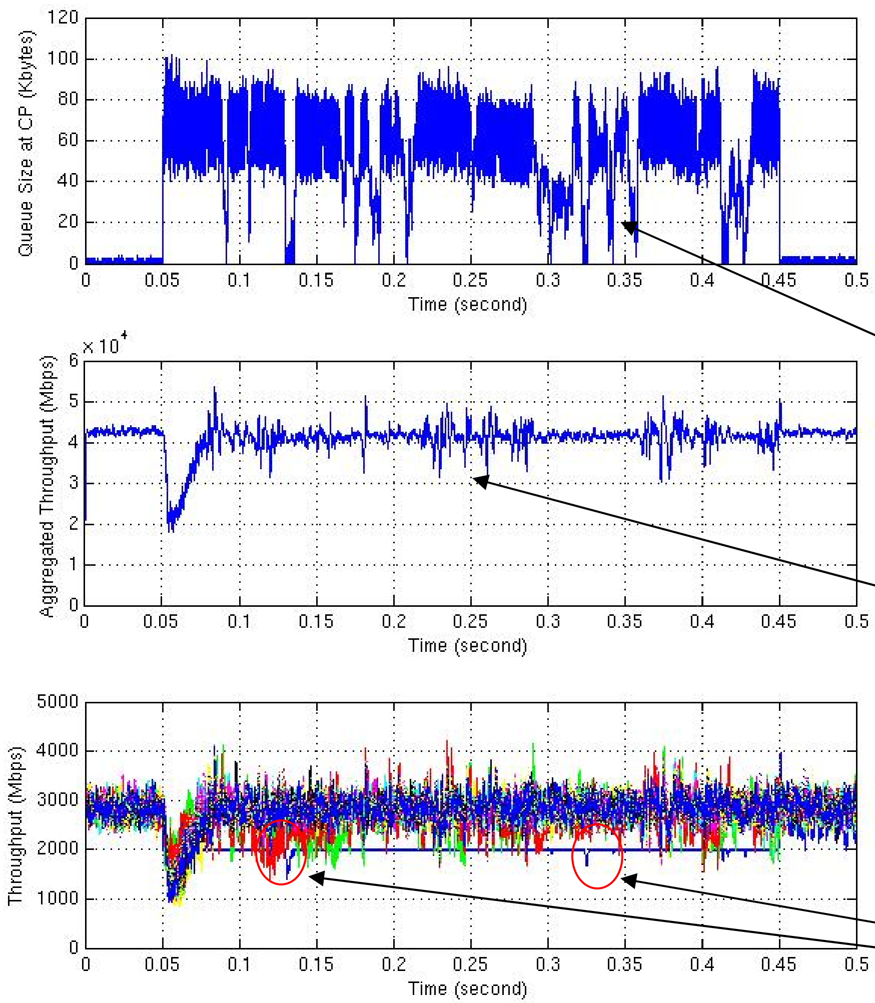
- Egress port memory size: 150K
- Qeq = 37.5Kbytes
- N = 5
- M = 3
- Gi = Linerate/40 * (1/((1+2*W)*Q_eq

Decreasing the increase gain in this case aids in dampening the queue oscillations.

Underutilization issues remain at the start.

With BCN and PAUSE
B = 75Kbytes

- Egress port memory size: 75K
- Qeq = 18.75Kbytes
- N = 5
- M = 3
- Gi = Linerate/10 * (1/((1+2*W)*Q_eq

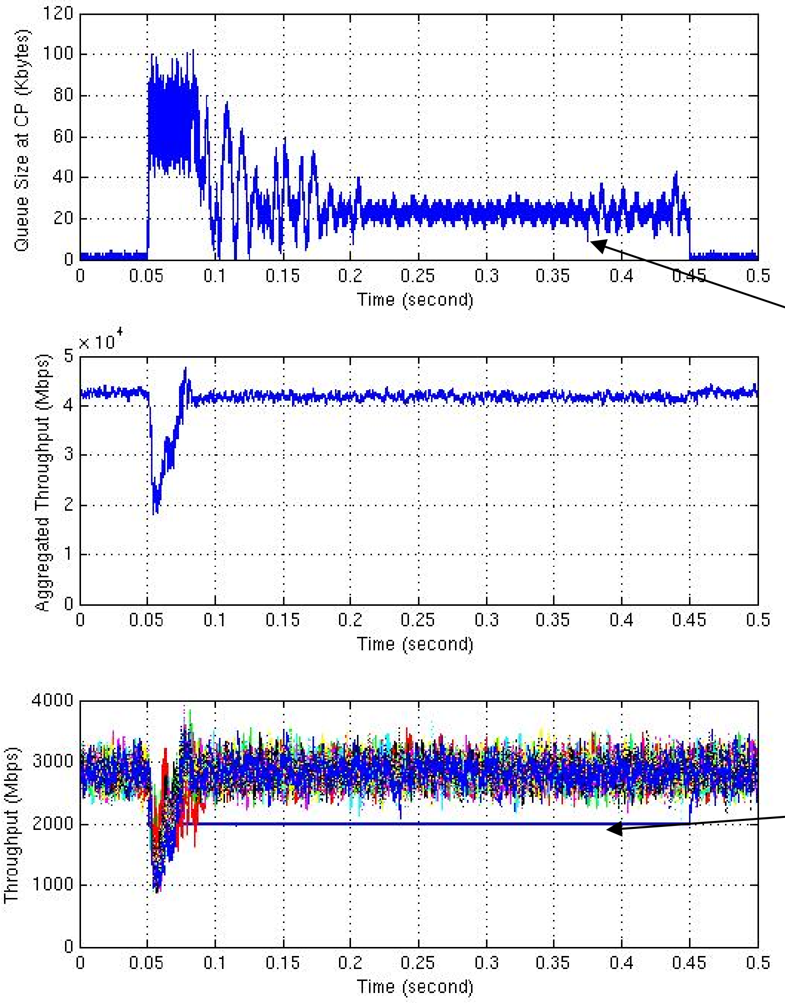


With reduced buffer size, BCN parameters require further adjustments to better control the queue size.

Aggregate throughput suffers in this scenario with the given BCN settings.

Underutilization issues.

With BCN and PAUSE
B = 75Kbytes (Reduced Gi)



- Egress port memory size: 75K
- Qeq = 18.75Kbytes
- N = 5
- M = 3
- Gi = Linerate/40 * (1/((1+2*W)*Q_eq

Decreasing the increase gain in this case aids in dampening the queue oscillations.

Utilization is good at congestion point.

Unclear that this set of settings is sufficient across other congestion scenarios.