Multi-hop Output Generated Hotspot Scenarios

Jin Ding & Bruce Kwan

January 4, 2007

Parameters

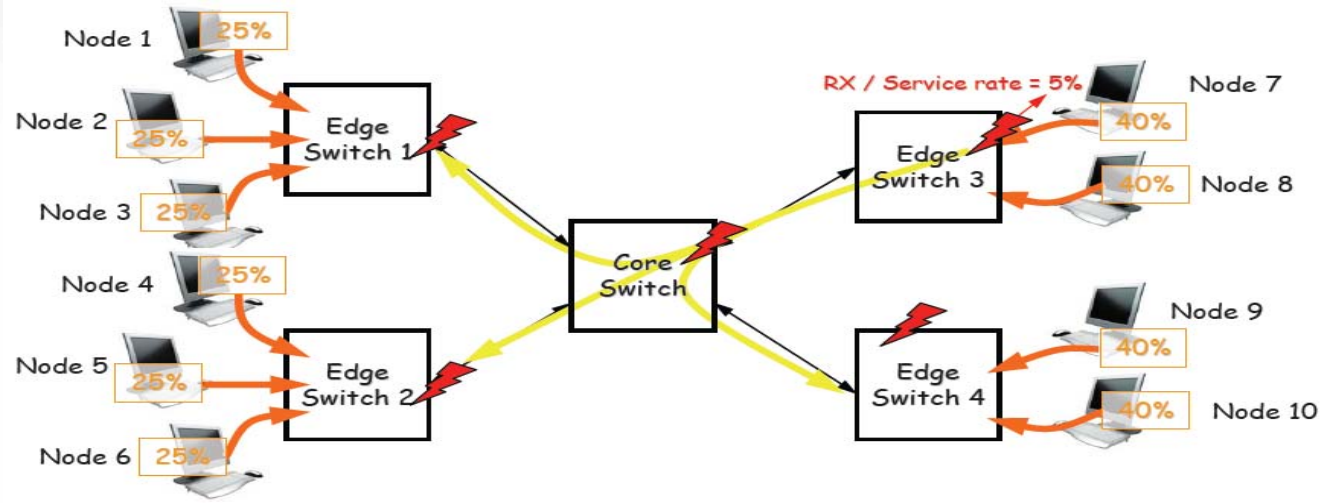
- Switch Parameters
  - Core switch and edge switches are all 4 port switches
  - Buffer Size (B) = 600Kbytes/Port
  - Shared Memory Switch Devices, total switch memory size = 4 * B = 2.4Mbytes
  - PAUSE Flow Control Settings
    - Applied per ingress port basis based on XON/XOFF thresholds
    - XOFF 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
    - No BCN-MAX or BCN(0,0)
    - No Self Increase
    - No Over-sampling during severe congestion

BROADCOM.
Connecting everything

Overview

- Experiment #1
- Experiment #2

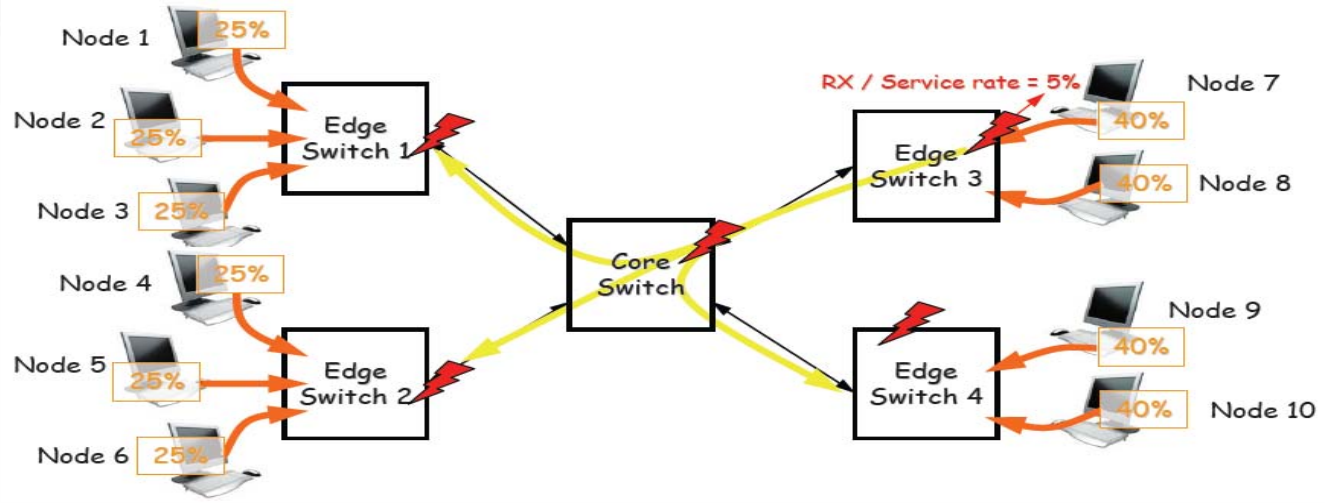
Example 1: Topology and Workload



- Multi-stage Output-Generated Hotspot Scenario
  - Link Speed = 10Gbps for all links
  - Loop Latency = 8us
- 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]

BROADCOM. Connecting everything

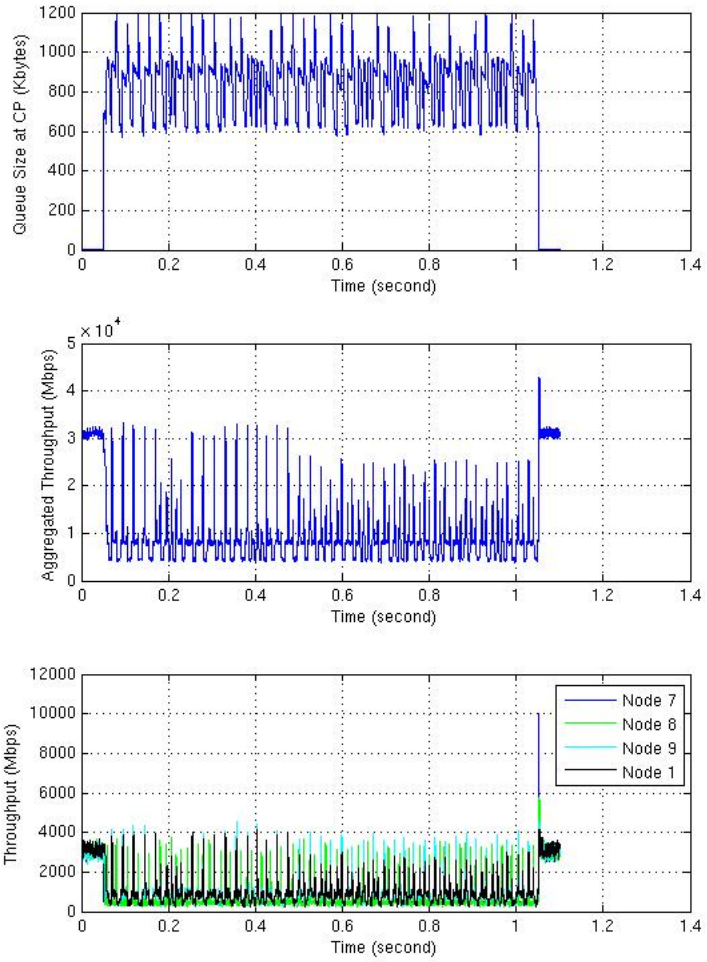
Experiment #1: Desired Throughput Performance



- Without hotspot, expected egress port throughput
  - @ Node 1-6: 3.167Gbps
  - @ Node 7-10: 3Gbps
  - Total aggregate throughput = 31Gbps
- With hotspot, desired egress port throughput during congestion period
  - @ Node 1-6: 3.167Gbps
  - @ Node 7: 500Mbps
  - @ Node 8-10: 3Gbps
  - Total aggregate throughput: 3.167*6+3*3+0.5 = 28.5Gbps

BROADCOM. Connecting everything

Experiment #1
(No BCN, PAUSE)



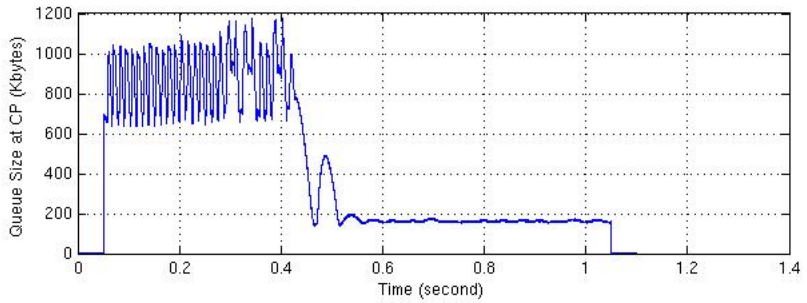
- Observations
  - PAUSE leads to congestion spread
    - All the flows are affected during congestion period
  - Packet Drops (in switch devices): 0
  - Total aggregate throughput (during congestion period)
    - 8.55 Gbps (Ideal = 28.5Gbps)

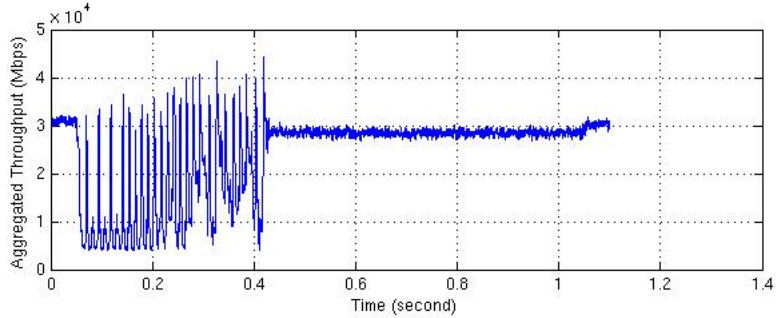
Egress Port Throughput (Mbps) during [50ms, 1050ms]

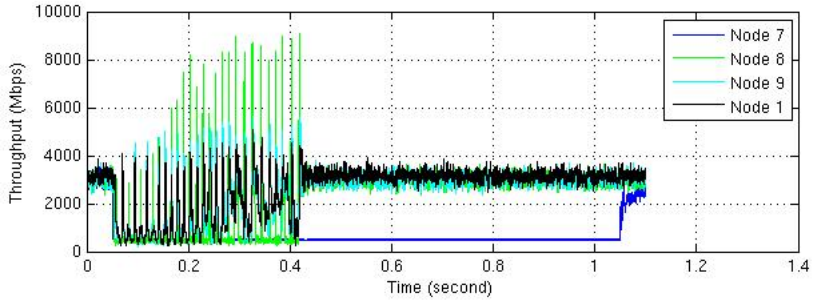
	Node 7	Node 8	Node 9	Node 1
Desired	*500*	*3000*	*3000*	*3167*
Observed	500	672.64	909.50	924.13
% Difference	0%	77%	69%	70%

BROADCOM.
Connecting everything

Experiment #1
(With BCN, PAUSE)







- Observations
  - PAUSE leads to congestion spread and results in multiple congestion points managed by BCN
  - All flows affected while PAUSE is active
  - BCN enhances aggregate throughput over PAUSE only scenario
  - Packet Drops (in switch devices): 0
  - Total aggregate throughput (during congestion period)
    - 23.587Gbps (Ideal = 28.5Gbps)

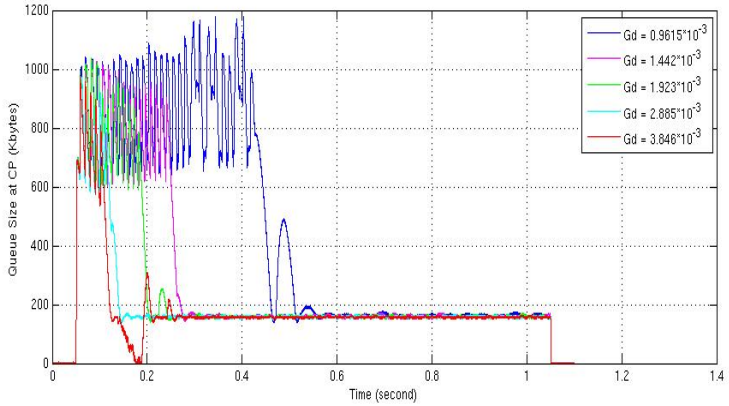
Egress Throughput (Mbps) during [50ms, 1050ms]

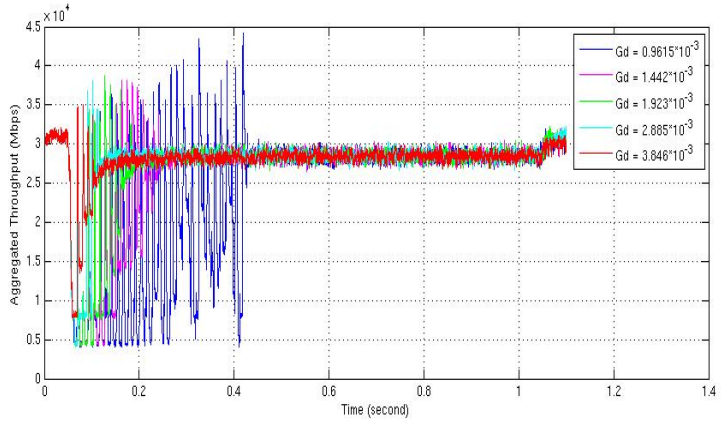
	Node 7	Node 8	Node 9	Node 1
Desired	*500*	*3000*	*3000*	*3167*
Observed	499.58	2296.22	2514.31	2617.56
% Difference	0.08%	23%	16.2%	17.3%

Experiment #1
Effects of Gd

- Setup
  - Qeq = 104 (1500-byte frames)
  - Gi = (Linerate/10) * [1/((1+2*W)*Q_eq)]
    - 1.923
  - Gd = Gd_factor * 1/((1+2*W)*Q_eq)
    - 0.5*1/((1+2*W)*Q_eq) = 0.9615 * 10^{-3}
    - 0.75*1/((1+2*W)*Q_eq) = 1.442 * 10^{-3}
    - 1.0*1/((1+2*W)*Q_eq) = 1.923 * 10^{-3}
    - 1.5*1/((1+2*W)*Q_eq) = 2.885 * 10^{-3}
    - 2.0*1/((1+2*W)*Q_eq) = 3.846 * 10^{-3}

Experiment #1
Effects of Gd





Egress Throughput (Mbps) during [50ms, 1050ms]

Gd	Node 7	Node 8	Node 9	Node 1
$0.9615 * 10^{-3}$	499.58	2296.22	2514.31	2617.56
$1.442 * 10^{-3}$	499.91	2590.89	2773.13	2939.31
$1.923 * 10^{-3}$	499.92	2681.68	2843.68	2979.96
$2.885 * 10^{-3}$	499.92	2847.34	2919.16	3077.30
$3.846 * 10^{-3}$	499.50	2715.04	2939.41	3102.10
Desired	*500*	*3000*	*3000*	*3167*

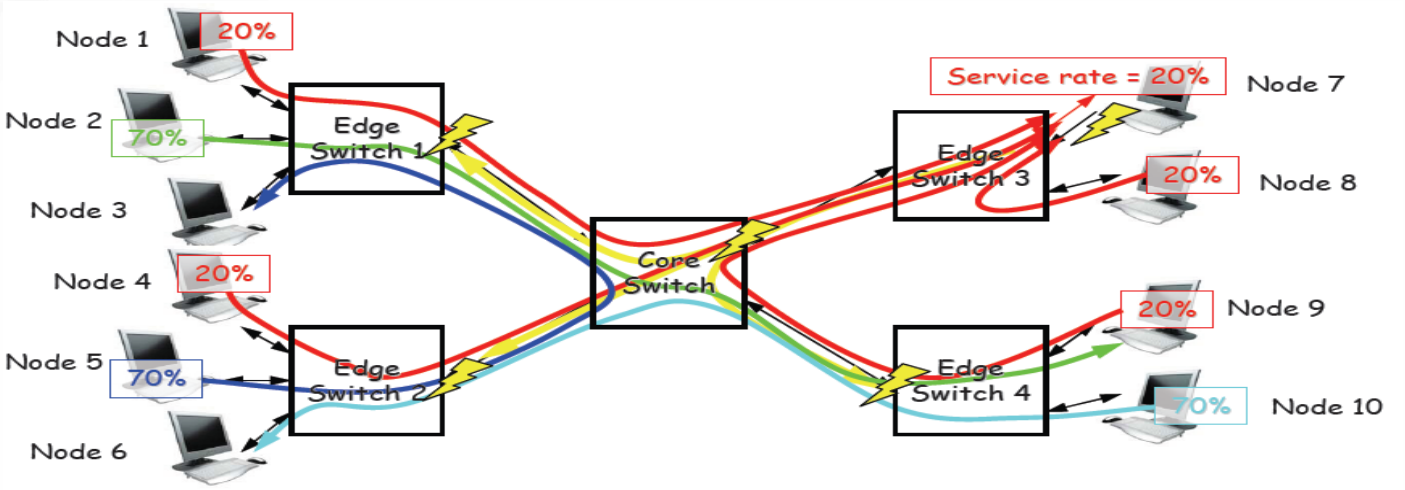
- As the strength of the Gd increases, the time spent with PAUSE active diminishes. However, underutilization issues also arise.
- With a weaker Gd, the time spent with PAUSE active increases.

BROADCOM.
Connecting everything

Overview

- Experiment #1
- Experiment #2

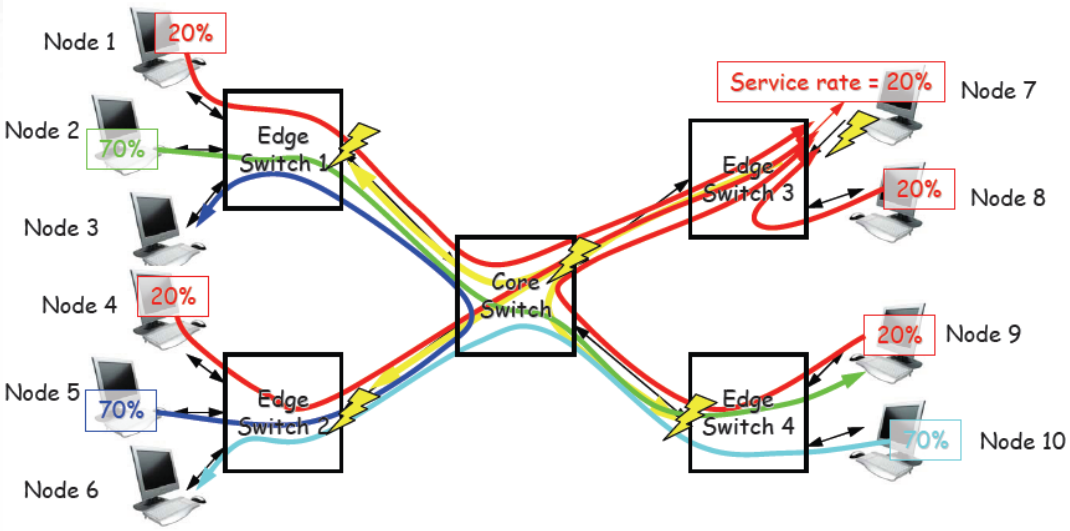
Example 2: Topology & Workload



- Multi-stage Output-Generated Hotspot Scenario
  - Link Speed = 10Gbps for all links
  - Loop Latency = 8us
- Traffic Pattern
  - 100% UDP (or Raw Ethernet) Traffic
  - Frame Size Distribution: Fixed length (1500bytes) frames
  - Four culprit flows of 2Gbps each from node 1, 4, 8, 9 to node 7
  - Three victim flows of 7Gbps each: node 2 to 9, node 5 to 3, node 10 to 6
- Congestion Scenario
  - Node 7 temporary reduce its service rate from 10Gbps to 2Gbps between [50ms, 1050ms]

BROADCOM. Connecting everything

Experiment #2
Desired Throughput Performance

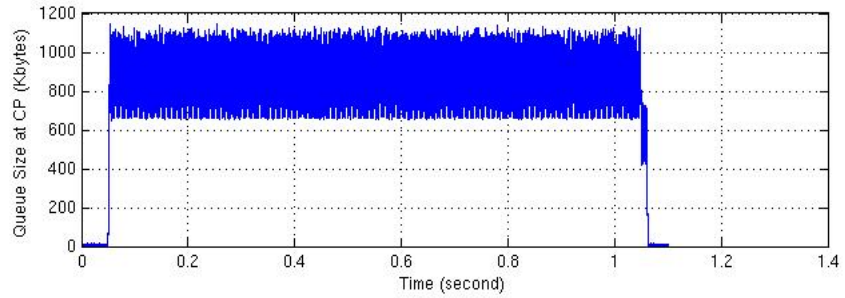


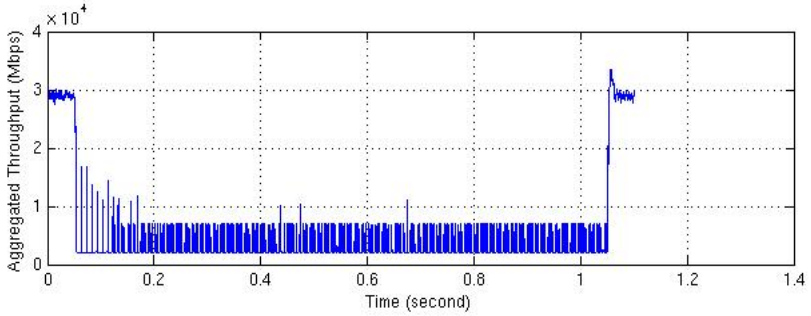
- Without hotspot, expected throughput
  - @ Node 7: 8Gbps
  - @ Nodes 3, 6, & 9: 7Gbps
  - Other nodes are 0
  - Total aggregated throughput
    - 10 * 20% * 4 + 10 * 70% * 3 = 29Gbps

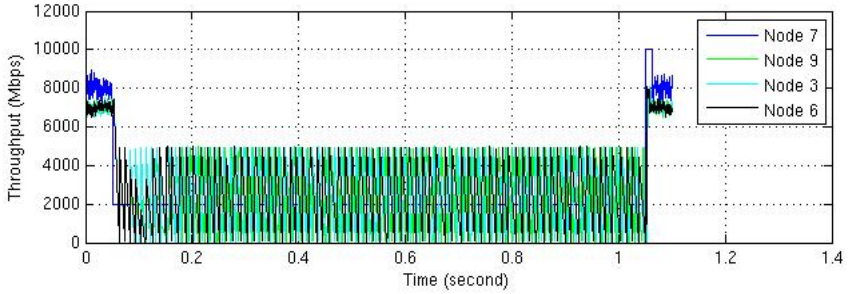
- With hotspot, desired throughput during congestion period
  - @ Node 7: 2Gbps
  - @ Nodes 3, 6, & 9: 7Gbps
  - Other nodes are 0.
  - Total aggregated throughput: 23Gbps
  - Fairness Attribute
    - Throughput to node 7 is fairly distributed among source nodes 1, 4, 8, & 9
    - Each with 500 Mbps

BROADCOM.
Connecting everything®

Experiment #2
(No BCN, PAUSE)







- PAUSE leads to congestion spread
  - All flows affected leading to degraded throughput
- Bandwidth at congestion point is spread between node 8 and the set of flows arriving from nodes 1, 4, 9.
- Total aggregate throughput
  - 3.21 Gbps (Ideal = 23 Gbps)
- RMS Fairness Index = 0.687 (Ideal = 0)

Throughput distribution at Node 7(Mbps) among incoming flows during [50ms, 1050ms]
(All should be 500Mbps)

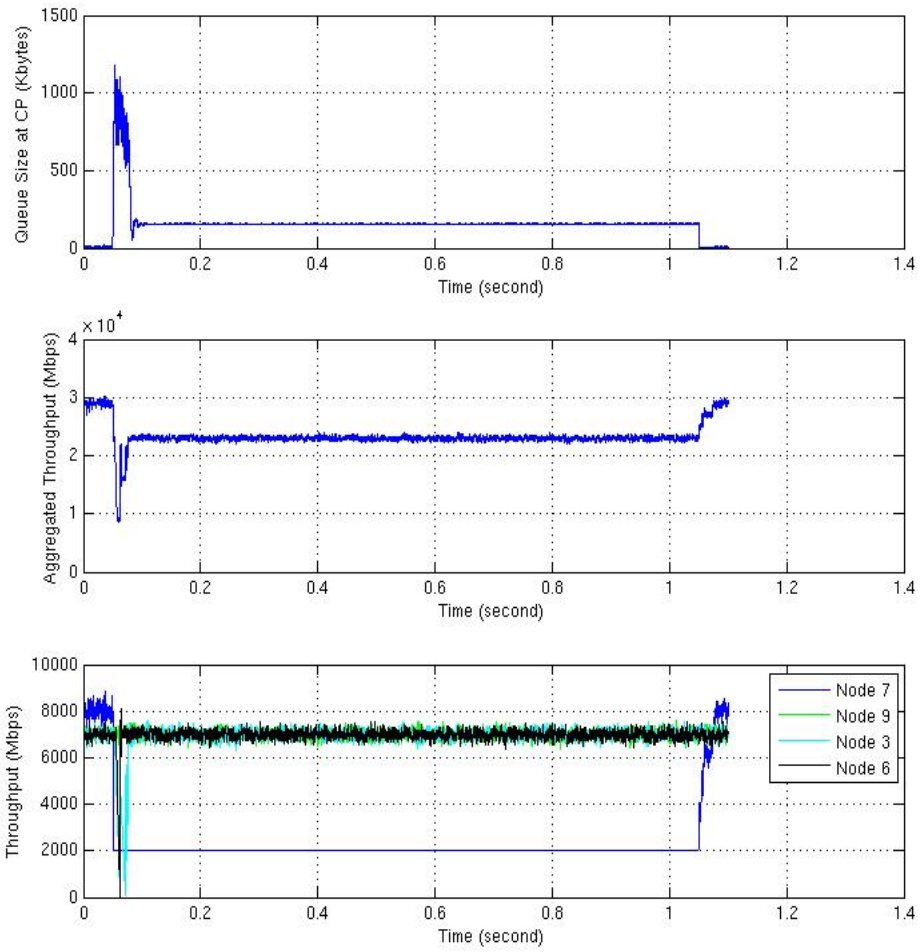
Node 8	Node 9	Node 1	Node 4
1094.99	303.12	301.27	300.65

Egress Port Throughput (Mbps) during [50ms, 1050ms]

Node 7	Node 9	Node 3	Node 6
2000	398.39	413.12	398.21

BROADCOM.
Connecting everything

Experiment #2
(With BCN, PAUSE)



- PAUSE leads to congestion spread and results in multiple congestion points managed by BCN
- Total aggregate throughput
  - 22.811Gbps (Ideal = 23Gbps)
- RMS Fairness Index = 1.417
  - Poor fairness due to multiple congestion points existing and leading to more BCN messages being sent to nodes 1,4, & 9.
  - When receiving multiple BCN messages with different CPID's, increase signals are ignored which exacerbate the issue.

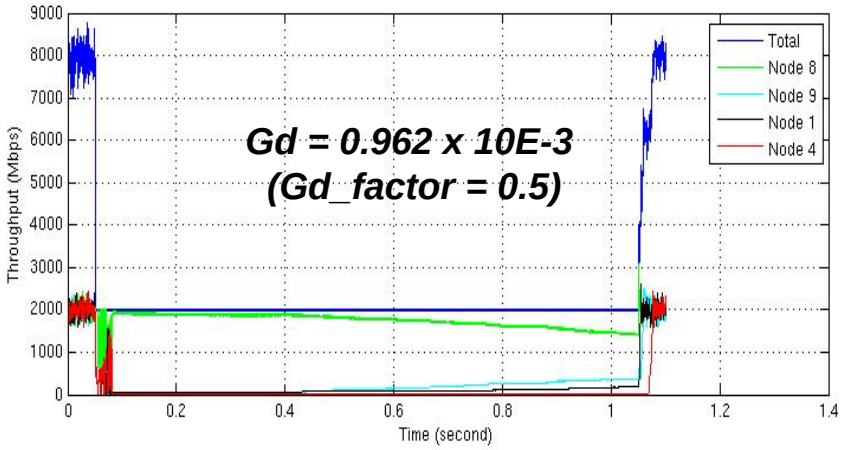
Throughput distribution at Node 7(Mbps) among incoming flows during [50ms, 1050ms] (All should be 500Mbps)

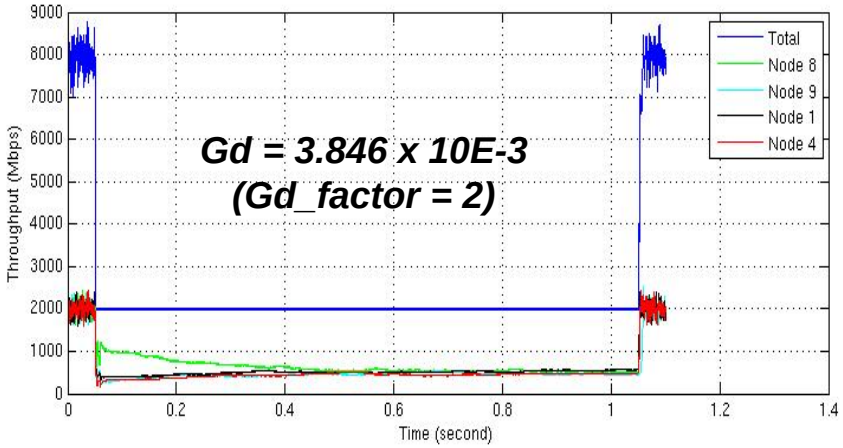
Node 8	Node 9	Node 1	Node 4
1723.91	159.35	93.34	23.39

Egress Port Throughput (Mbps) during [50ms, 1050ms]

Node 7	Node 9	Node 3	Node 6
2000	6996.68	6865.82	6949.16

Experiment #2: Fairness Issue (With BCN, PAUSE)





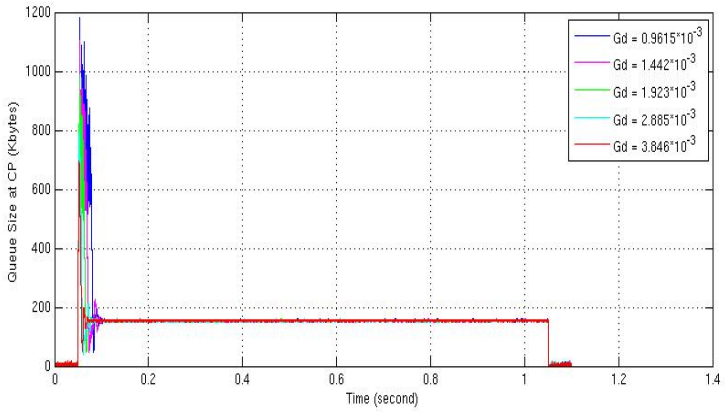
- Increasing Gd leads to faster convergence to a fair distribution of bandwidth.

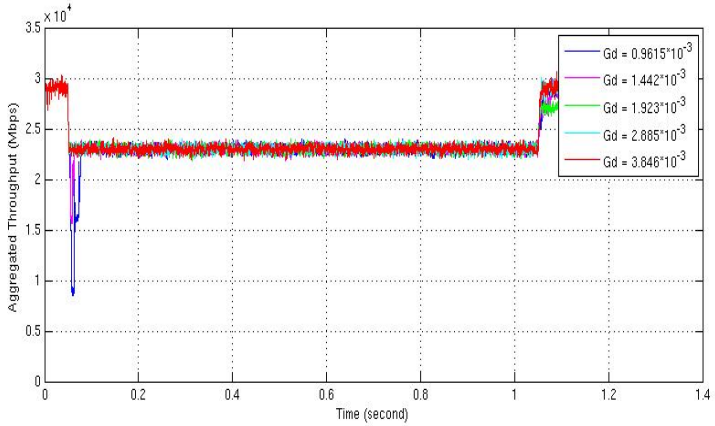
Experiment #2
Effects of Gd

- Setup
  - Qeq = 104 (1500-byte frames)
  - Gi = (Linerate/10) * [1/((1+2*W)*Q_eq)]
    - 1.923
  - Gd = Gd_factor * 1/((1+2*W)*Q_eq)
    - 0.5*1/((1+2*W)*Q_eq) = 0.9615 * 10^{-3}
    - 0.75*1/((1+2*W)*Q_eq) = 1.442 * 10^{-3}
    - 1.0*1/((1+2*W)*Q_eq) = 1.923 * 10^{-3}
    - 1.5*1/((1+2*W)*Q_eq) = 2.885 * 10^{-3}
    - 2.0*1/((1+2*W)*Q_eq) = 3.846 * 10^{-3}

BROADCOM.
Connecting everything®

Experiment #2
Effects of Gd





Throughput distribution at Node 7(Mbps) among incoming flows during [50ms, 1050ms] (All should be 500Mbps)

Gd (* 10^{-3})	Node 8	Node 9	Node 1	Node 4
0.9615	1723.91	159.35	93.34	23.39
1.442	1614.24	31.90	182.23	171.62
1.923	1095.18	12.19	372.86	519.75
2.885	1082.89	343.89	439.08	134.14
3.846	627.09	443.02	499.24	430.66

BROADCOM.
Connecting everything

Summary Observations

- General behavior observed is not too surprising.
- PAUSE leads to the need to manage multiple congestion points. This dynamic leads to unfair distribution of bandwidth at a congestion point unless further enhancements are considered to manage severe congestion events (i.e. BCN-MAX, Oversampling, etc).
- Buffer size assumptions need to be varied while also quantifying latency performance.
- When disabling PAUSE, need to specify assumptions on partitioning of buffering to avoid starvation issues.

BROADCOM.
Connecting everything®