Zurich Hotspot Benchmark: BCN Sensitivity Analysis II

Cyriel Minkenberg and Mitch Gusat
IBM Zurich Research Lab

Jan. 2007

Overview

- This presentation documents ZRL simulation results of applying Zurich Hotspot Benchmark (ZHB) on .1au BCN
- Progress since Dallas Plenary
- Next ZHB simulations

Simulation Setup Overview

- Use algorithmic and parametrical sensitivity analysis to answer
  - Which parameters?
  - To what traffic?

Algorithmical and Parametrical Sensitivity Analysis

BCN stability model equations:

1. Conservation: $dq/dt = HSD*\lambda(t) - \mu_{HS}$ =>
2. $q(s) = HSD* \lambda(s) / s$
3. Feedback: $Fb(t) = -(q(t) - Q_{eq}) + w*(dq/dt) / (\mu_{HS}* p_s)$ =>
4. $Fb(s) \approx G * [1 + w*s / (\mu_{HS}* p_s)]$

5. AI: $d\lambda(t)/dt = G_i*\lambda(t)* p_s*Fb(t-\tau)$
6. $\delta AI(t)/\delta Fb(t-\tau) = G_i*p_s*\mu_{HS}/HSD$ =>
7. AP sensitivity of $G_i = \delta AI(t)/\delta Fb(t-\tau) * HSD/(p_s*\mu_{HS})$

8. MD: $d\lambda(t)/dt = G_d*\lambda(t)*\lambda(t-\tau)* p_s*Fb(t-\tau)$
9. $\delta MD(t)/\delta Fb(t-\tau) \approx G_d*p_s*(\mu_{HS}/HSD)^2$ =>
10. AP sensitivity of $G_d = \delta MD(t)/\delta Fb(t-\tau) * (\mu_{HS}/HSD)^{-2} / p_s$.

q(t) =queue occupancy; HSD=no. of hot flows, each with rate $\lambda(t)$, at hotspot served w/ rate μ_{HS}

Algorithmical and Parametrical Sensitivity Analysis, ctd

Eq. (7,10) =>

a) p_s *directly* impacts G_i and G_d

- 1st order sensitivity on p_s

b) G_i and G_d depend on the HSD/μ_{HS} ratio

- congestion w/ low μ_{HS} or / and high HSD stresses stability

(10) =>

c) G_d has quadratic sensitivity to the HSD/μ_{HS} ratio

(4,7,10) =>

if denominator ~ f (p_s*μ_{HS}), where p_s« 1 and μ_{HS}≤1

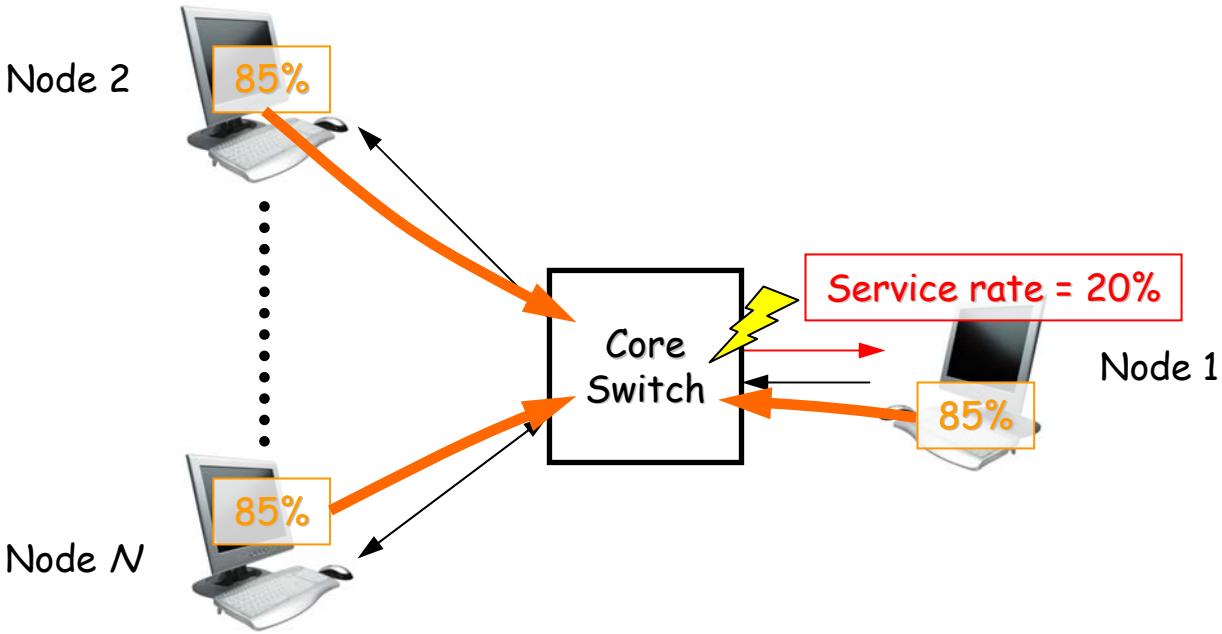
=> the hotspot drain rate *further* increases the sensitivity to p_s

d) everyting else being equal (e.g. hotspot severity and degree), *output-generated* (OG) congestion is more stressful for stability than input-generated (IG)

What to begin with?

- BCN params: p_s (most influential BCN parameter!) ; gains G_d and partially, G_i
- Traffic: Output-generated congestion w/ high HSD and low μ_{HS} .

Case 1: Output-Generated Single-Hop Hotspot

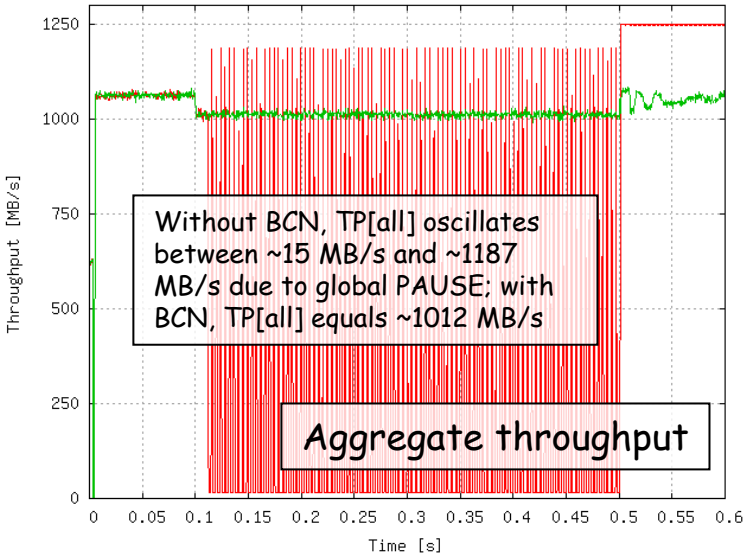


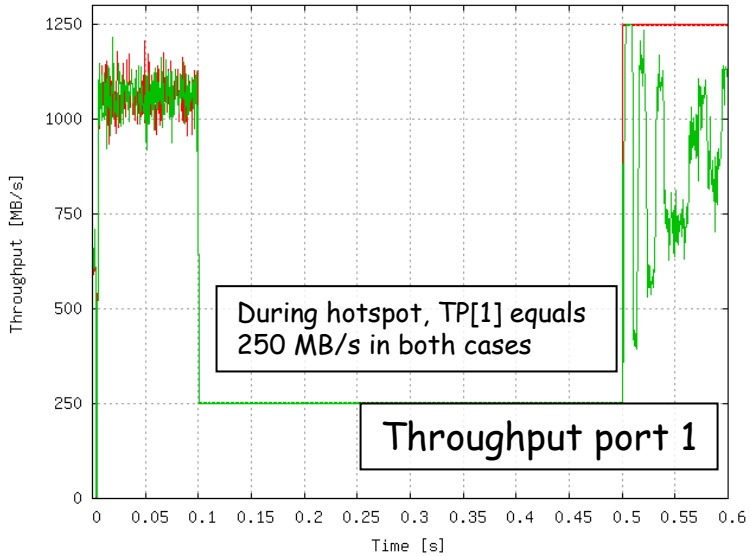
- All nodes: Uniform destination distribution, load = 85% (8.5 Gb/s)
- Node 1 service rate = 20%
- One congestion point
  - Hotspot degree = N-1
  - All flows affected

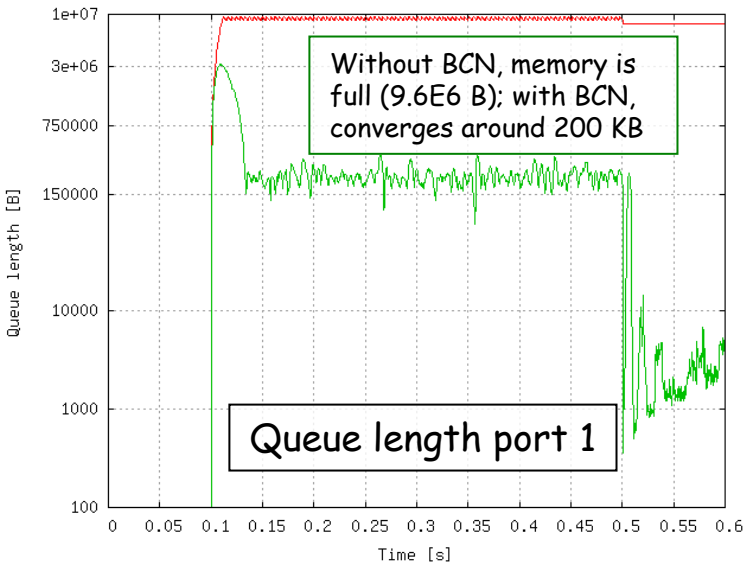
Simulation Setup

- Traffic
  - I.i.d. Bernoulli arrivals
  - Uniform destination distribution (to all nodes except self)
  - Fixed frame size = 1500 B
  - Load = 85%
- Network
  - Single-stage
  - N = 16
  - M = 600 KB/port
  - Shared memory
    - PAUSE applied to all ports simultaneously based on global high/low watermarks
    - $watermark_{high}$ = N*(M – rtt*bw)
    - $watermark_{low}$ = $watermark_{high}$ / 2
  - Partitioned memory per input
    - Deadlock prevention
    - PAUSE applied on a per input basis based on local high/low watermarks
    - $watermark_{high}$ = M – rtt*bw
    - $watermark_{low}$ = $watermark_{high}$ / 2
- BCN
  - W = 2.0
  - $G_i = 6.6667*10^{-4}$
  - $G_d = 1.6667*10^{-6}$
  - Q_{eq} = 150 KB (= M/4)
  - P_{sample} = 2% (on average 1 sample every 75 KB)
  - $R_u = R_{min}$ = 10 Mb/s
  - No BCN(0,0) or BCN_MAX, no self-increase

Results: Throughput & queue length - Shared memory



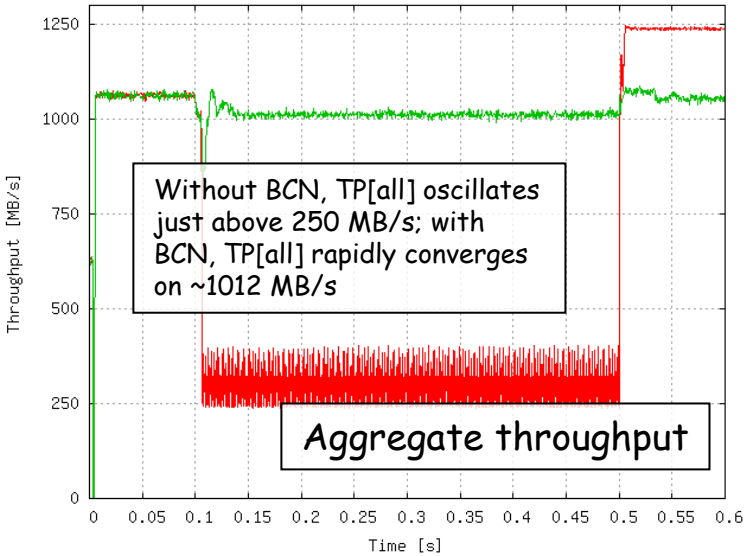


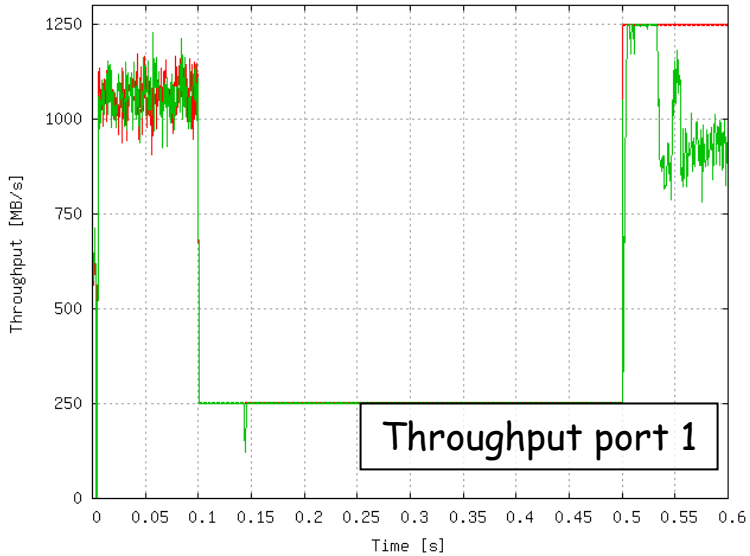


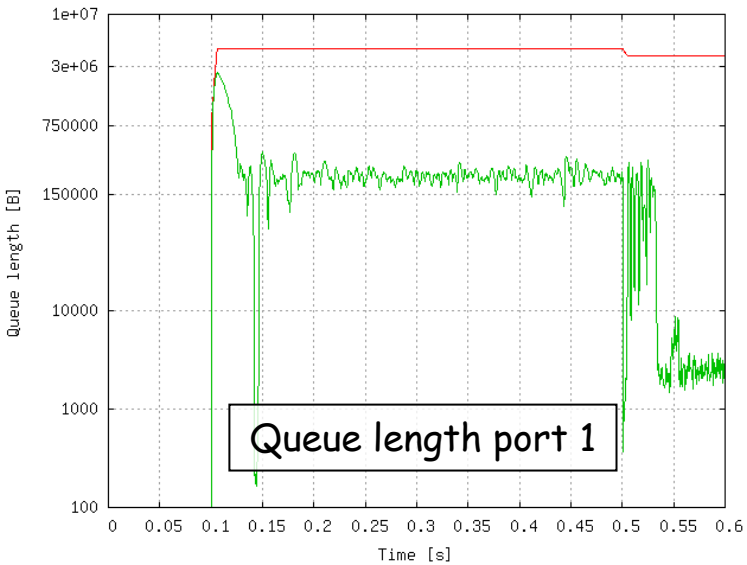
P_{sample} = 2%

No BCN

Results: Throughput & queue length - Partitioned memory



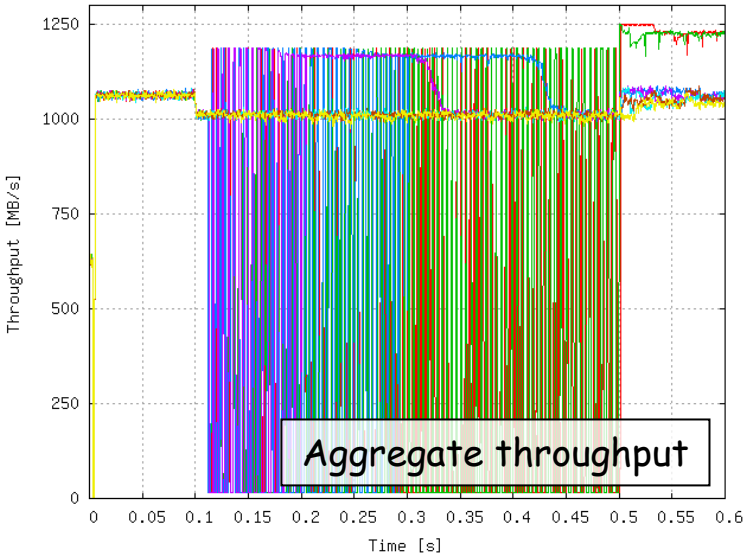


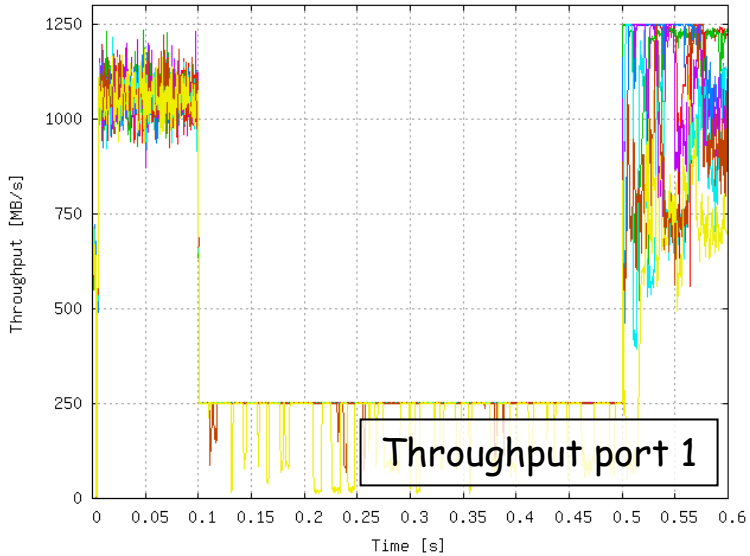


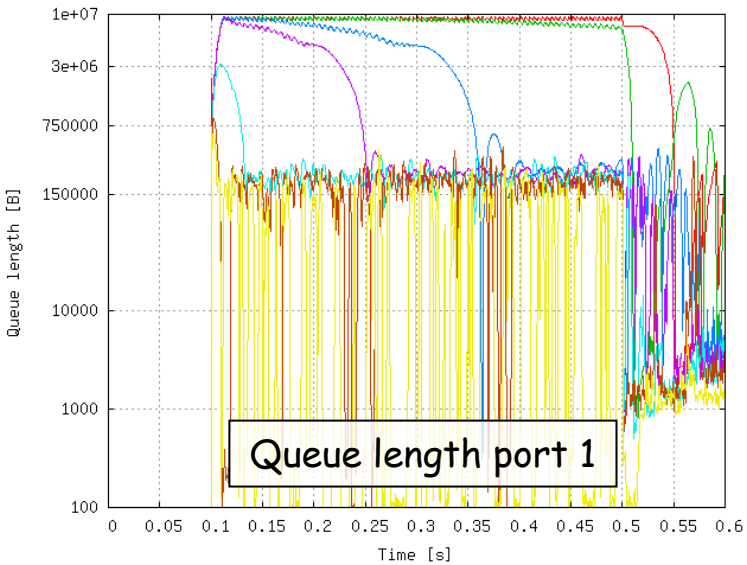
P_{sample} = 2%

No BCN

Results: G_d sensitivity – Shared memory







$G_{d0} = 6.6667*10^{-7}$

$G_d = 0.10*G_{d0}$

$G_d = 0.25*G_{d0}$

$G_d = 0.50*G_{d0}$

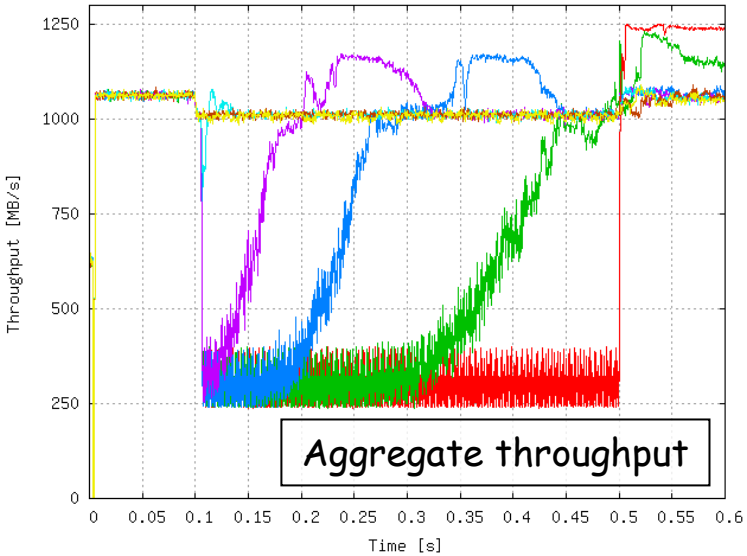
$G_d = 1.0*G_{d0}$

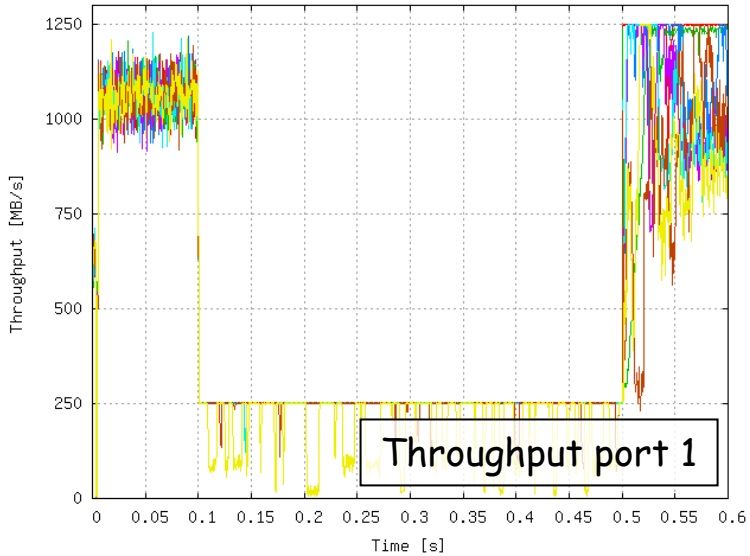
$G_d = 2.5*G_{d0}$

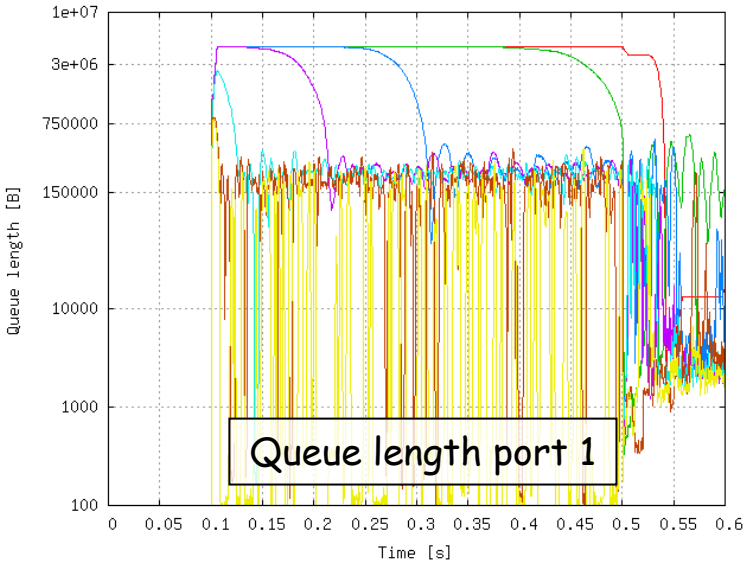
$G_d = 5.0*G_{d0}$

$G_d = 10.0*G_{d0}$

Results: G_d sensitivity – Partitioned memory





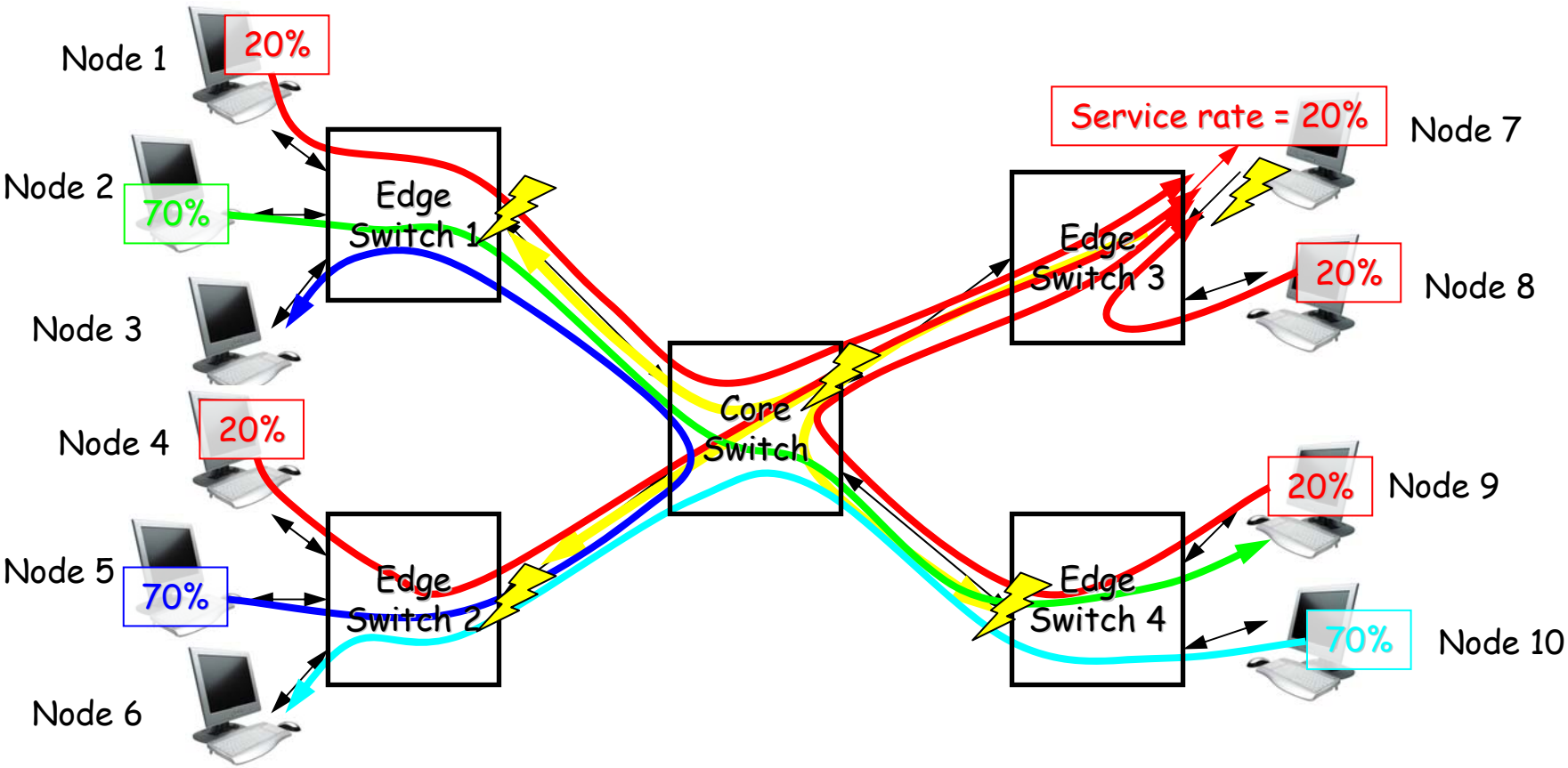


$G_{d0} = 6.6667*10^{-7}$
$G_d = 0.10*G_{d0}$
$G_d = 0.25*G_{d0}$
$G_d = 0.50*G_{d0}$
$G_d = 1.0*G_{d0}$
$G_d = 2.5*G_{d0}$
$G_d = 5.0*G_{d0}$
$G_d = 10.0*G_{d0}$

Single Hop OG Preliminary Conclusions

- Without BCN, overall performance is severely degraded
  - Hogging occurs with shared as well as partitioned memory
  - Mean aggregate throughput gated by hotspot throughput
- BCN is able to control the hotspot
  - OQ steady state length exceeds target
  - Quite sensitive to G_d setting
    - G_d too low: Slow reaction; overall throughput suffers because hogging not sufficiently reduced
    - G_d too high: Excessive throttling; hotspot throughput suffers, queue length oscillates strongly

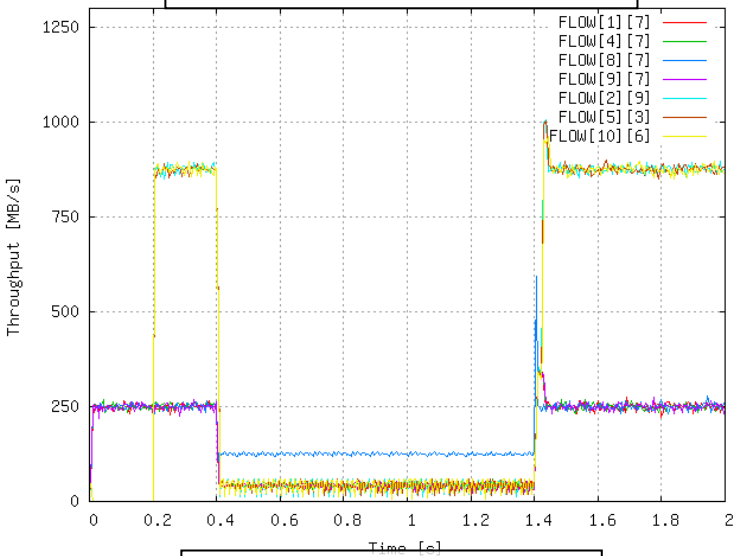
Case 2: Output-Generated Multi-Hop Hotspot



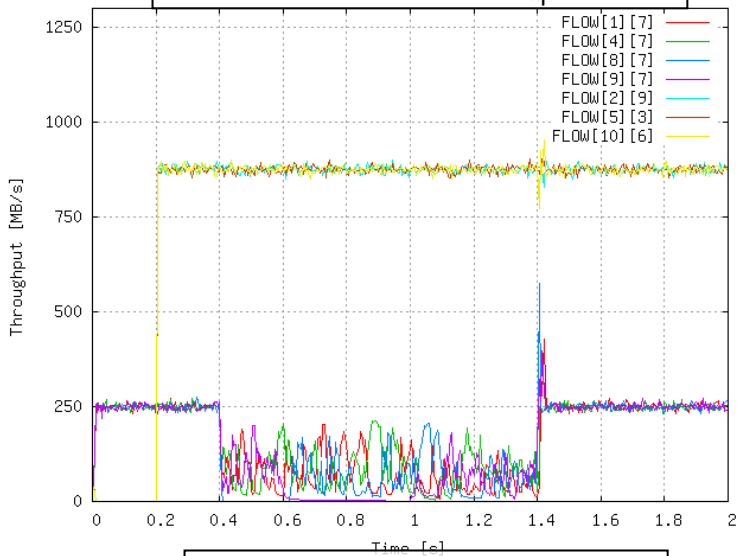
- Four culprit flows of 2 Gb/s each from nodes 1, 4, 8, 9 to node 7 (hotspot)
- Three victim flows of 7 Gb/s each: node 2 to 9, node 5 to 3, node 10 to 6
- Node 7 service rate = 20%
- Five congestion points
  - All switches and all flows affected
  - Fair allocation provides 0.5 Gb/s to all culprits and 7 Gb/s to all victims

Results: Without background traffic - Partitioned memory

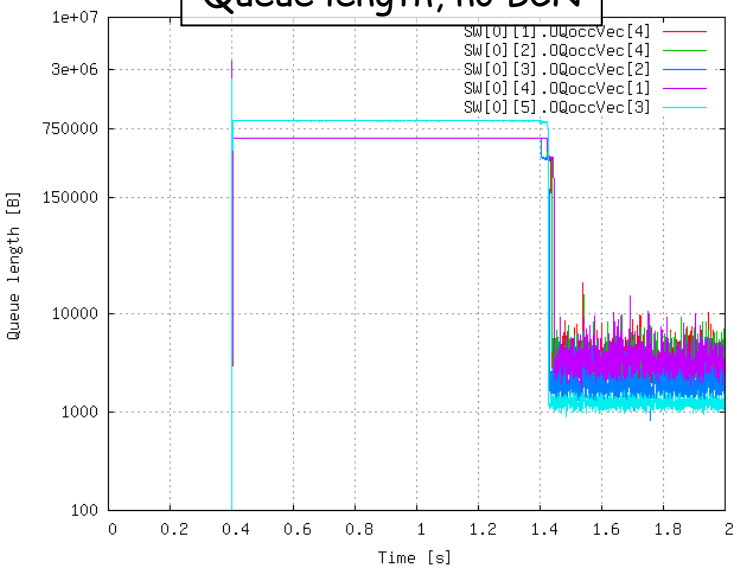
Flow throughput, no BCN



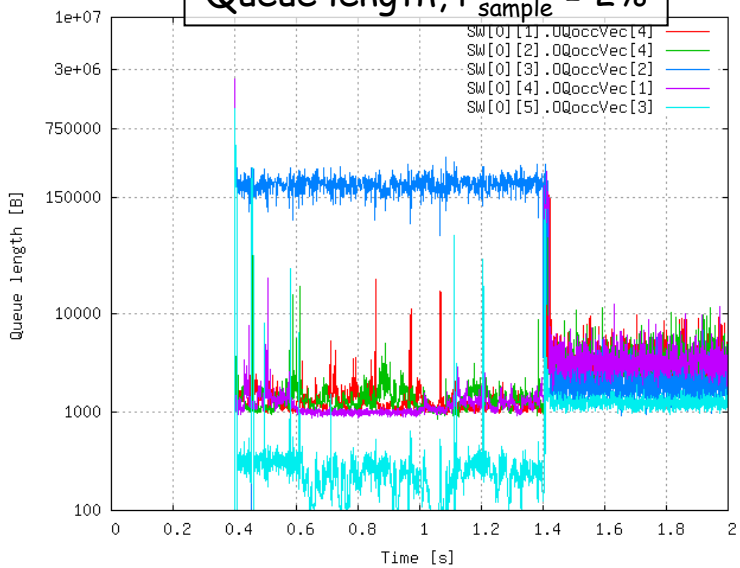
Flow throughput, P_{sample} = 2%



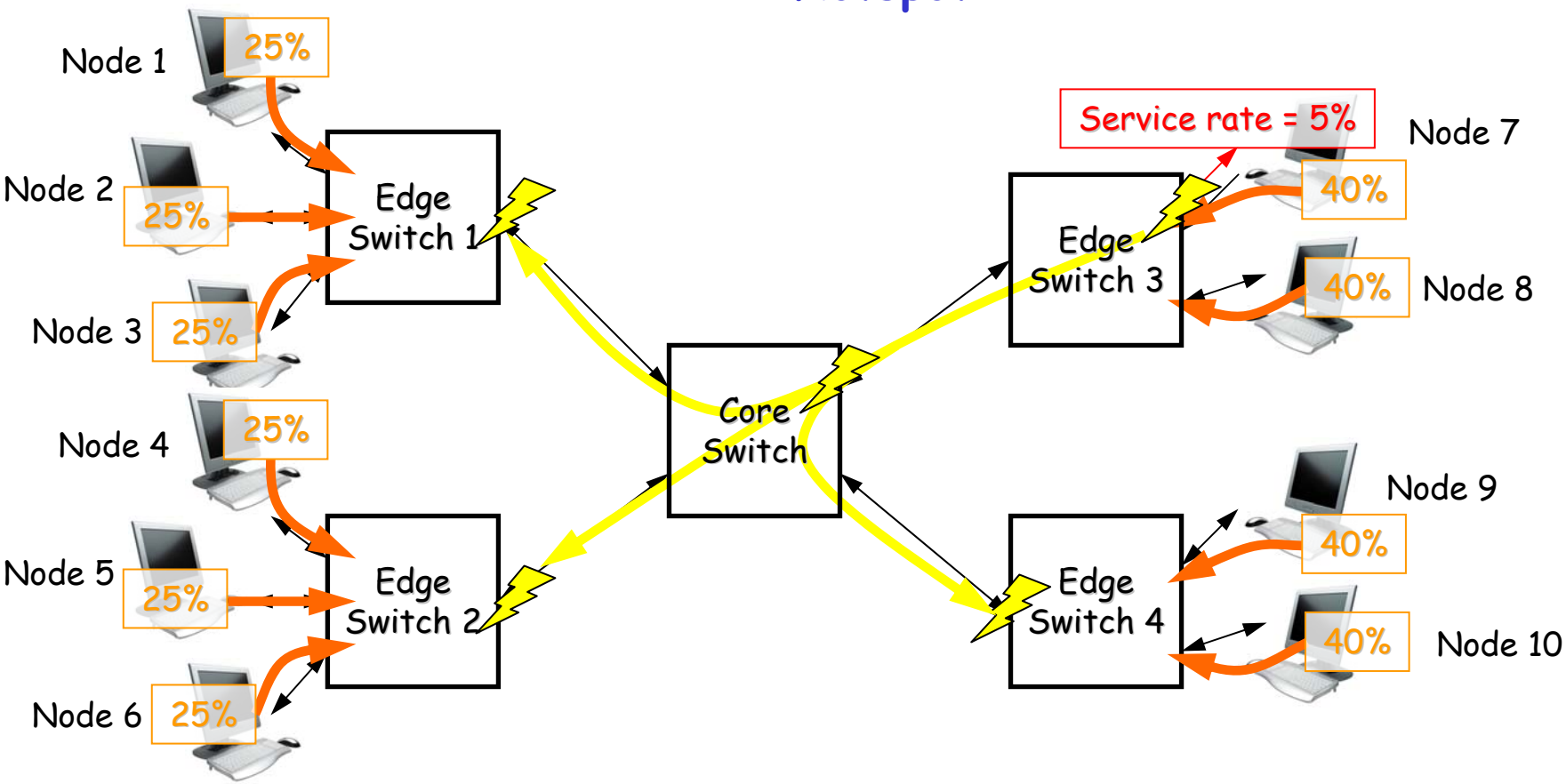
Queue length, no BCN



Queue length, P_{sample} = 2%



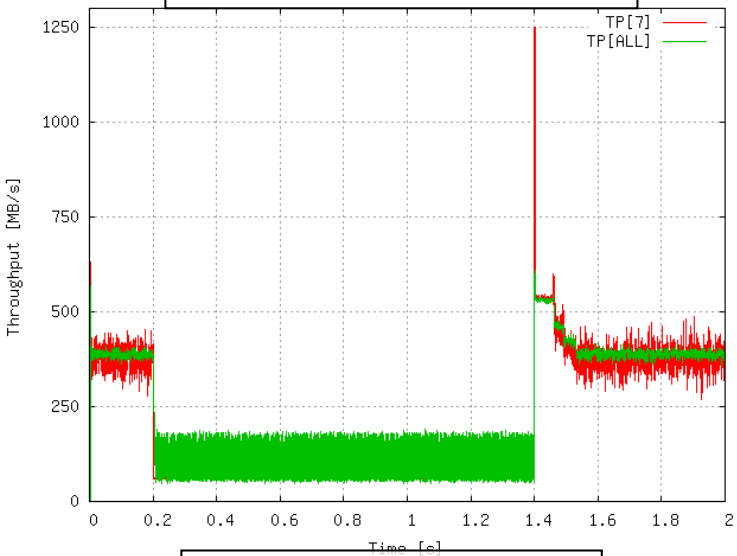
Case 3: Output-Generated Background Traffic Multi-Hop Hotspot



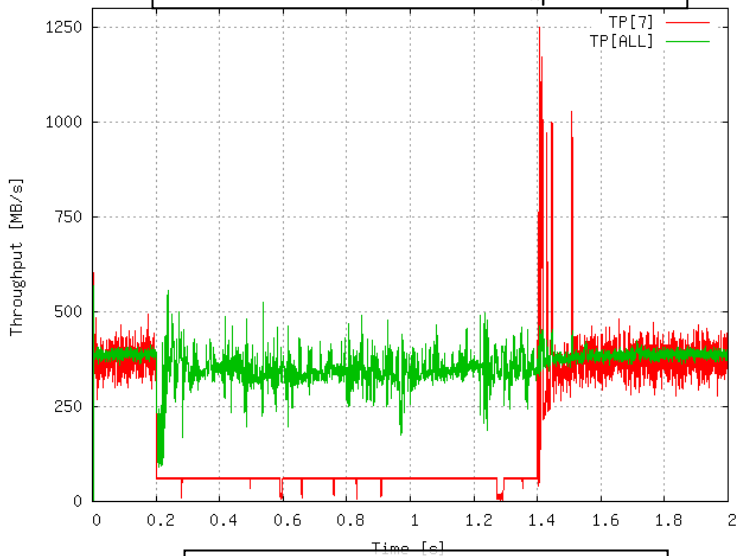
- All nodes: Uniform destination distribution
- Nodes 1-6 load = 25% (2.5 Gb/s), nodes 7-10 load = 40% (4 Gb/s)
  - Mean aggregate load = (6*.25+4*.4)/10 = 31% (3.1 Gb/s)
- Node 7 service rate = 5%
- Five congestion points
  - All switches and all flows affected

With background traffic - Partitioned memory

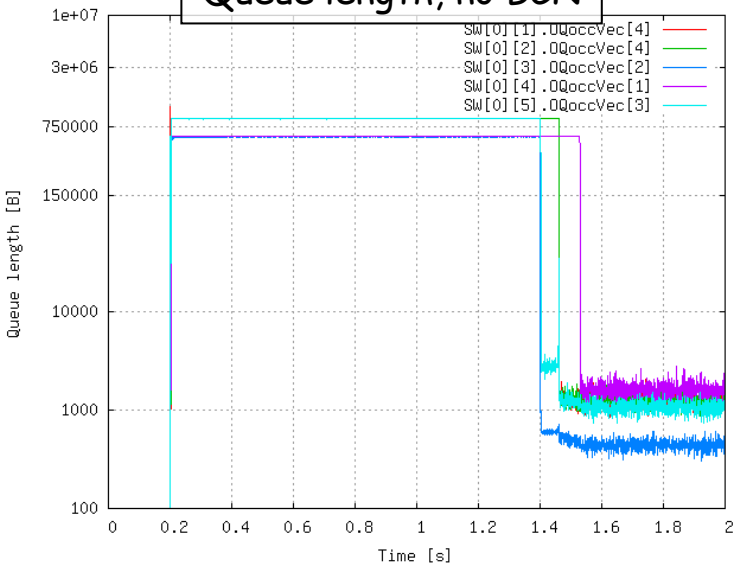
Flow throughput, no BCN



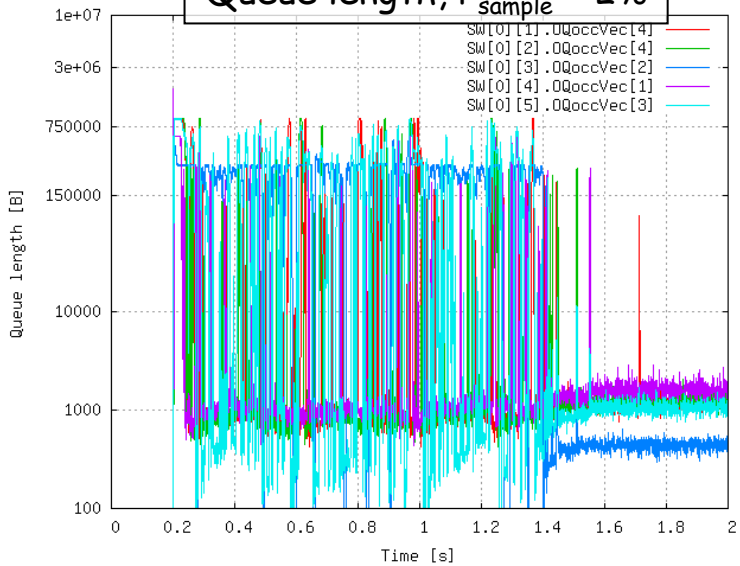
Flow throughput, P_{sample} = 2%



Queue length, no BCN

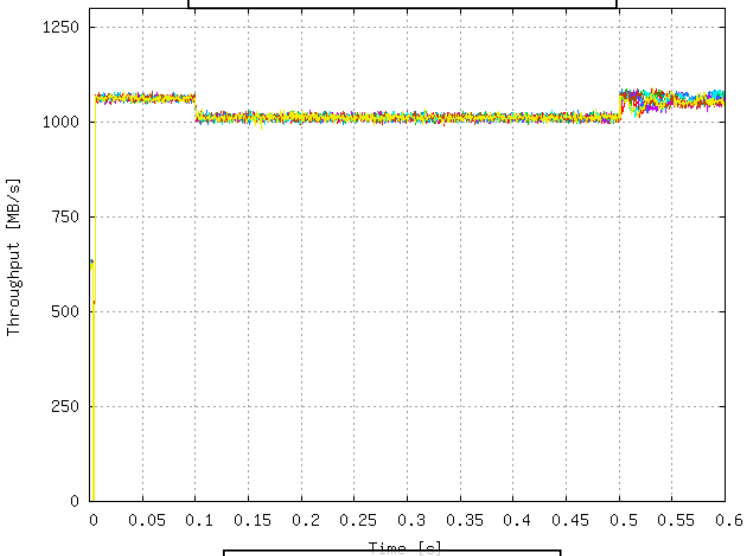


Queue length, P_{sample} = 2%

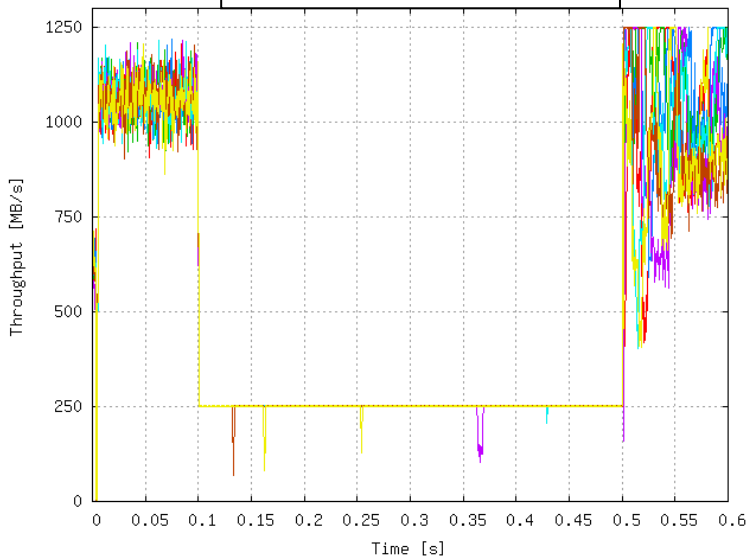


SS-OG: G_i sensitivity - fixed G_d & W

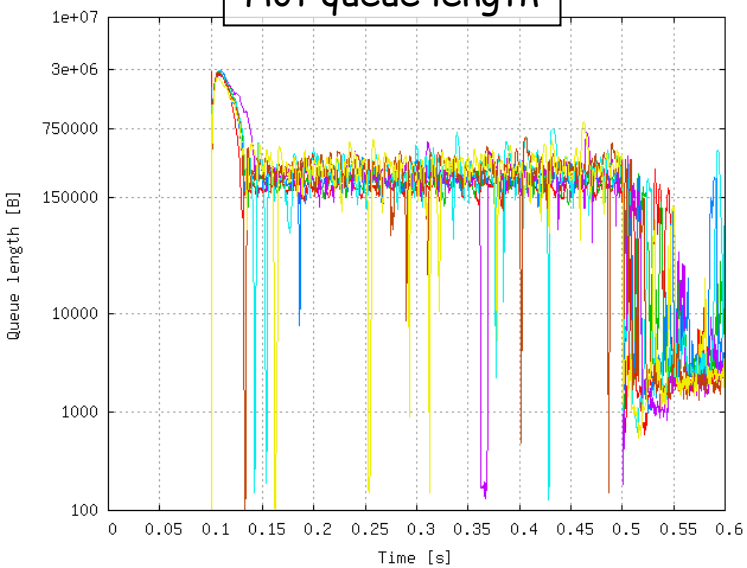
Aggregate throughput



Hot port throughput



Hot queue length

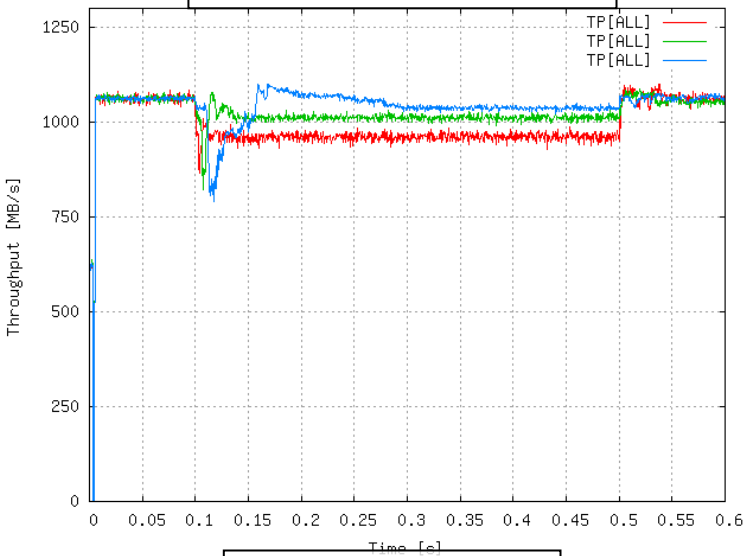


G_d = 1.6667*10^{-6}
G_i = 0.10*G_d
G_i = 0.25*G_d
G_i = 0.50*G_d
G_i = 1.0*G_d
G_i = 2.5*G_d
G_i = 5.0*G_d
G_i = 10.0*G_d

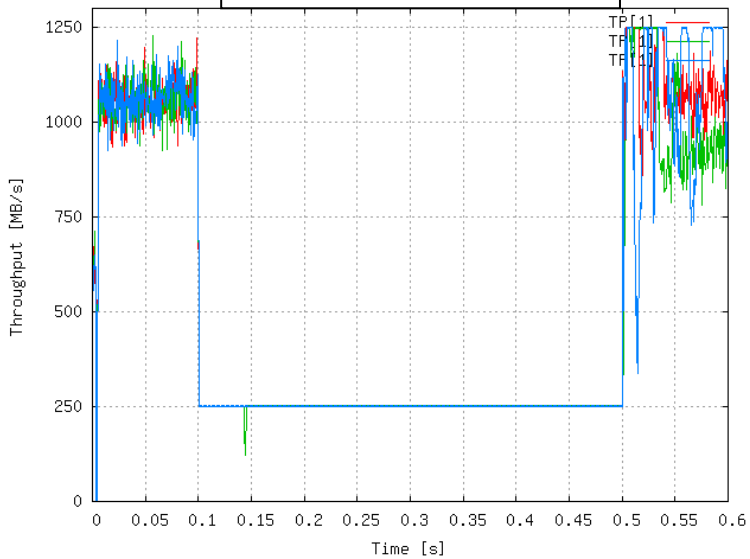
M = 600 KB/port, shared memory
N = 16
G_d = 1.6667*10^{-6}

SS-OG – N sensitivity (partitioned memory)

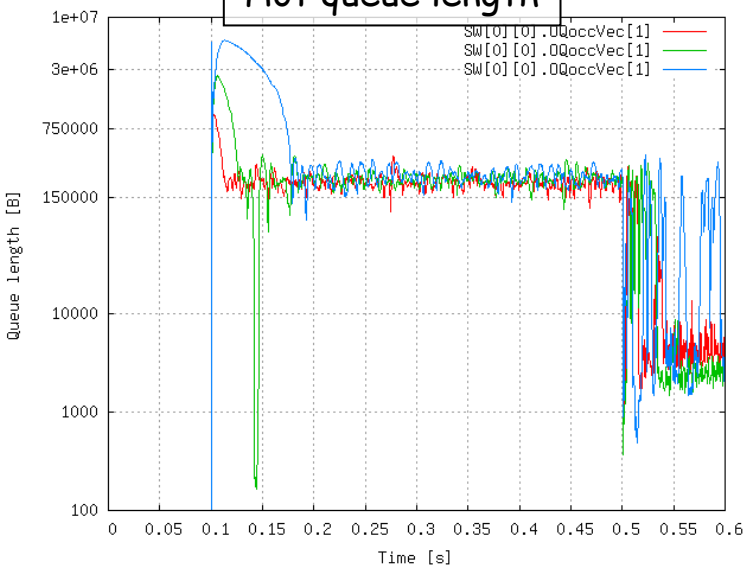
Aggregate throughput



Hot port throughput



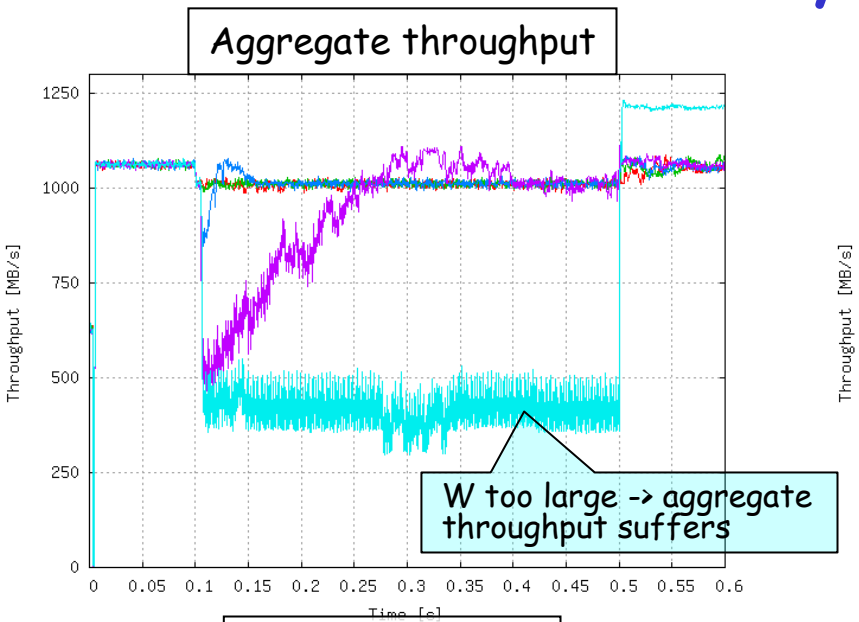
Hot queue length

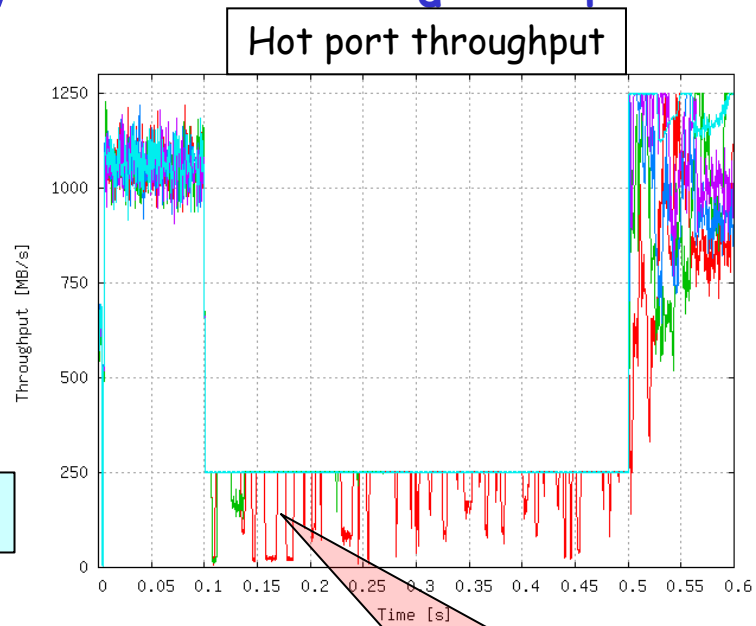


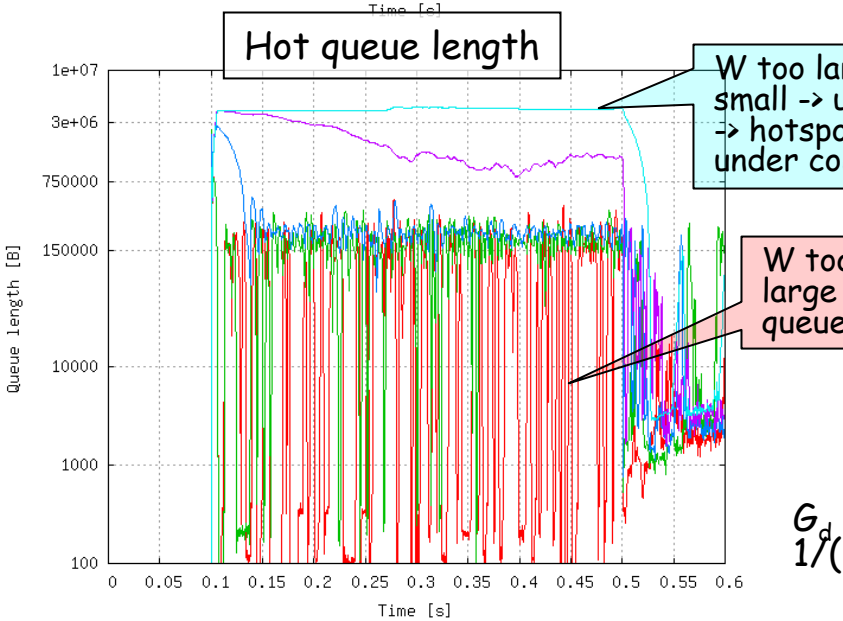
N = 8 | N = 16 | N = 32

W = 2.0
G_i = 6.6667*10-4
G_d = 1.6667*10-6
M = 600 KB/port
Q_{eq} = 150 KB (= M/4)
P_{sample} = 2%
R_u = R_{min} = 10 Mb/s
No BCN(0,0) or BCN_MAX

SS-OG: W sensitivity – variable G_d & G_i







W = 0.1
W = 0.5
W = 2
W = 10
W = 50

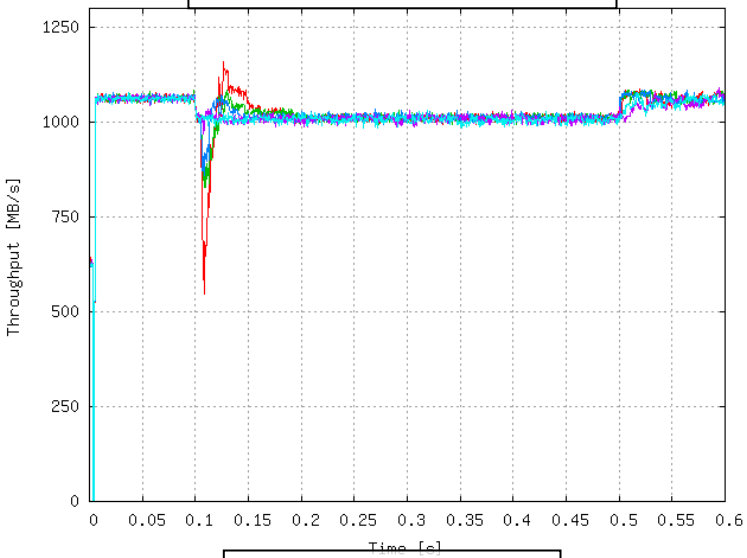
underdamped

overdamped

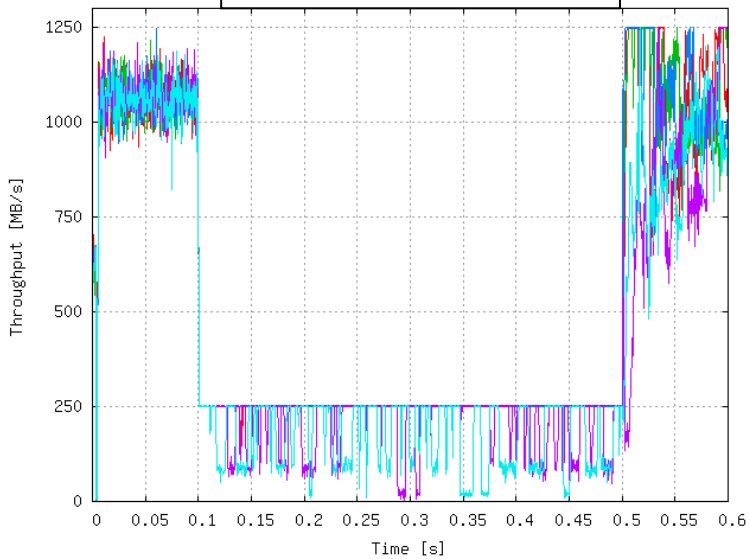
G_d and G_i are scaled proportional to $1/(2W+1)$, N = 16

SS-OG: W sensitivity - fixed G_d & G_i

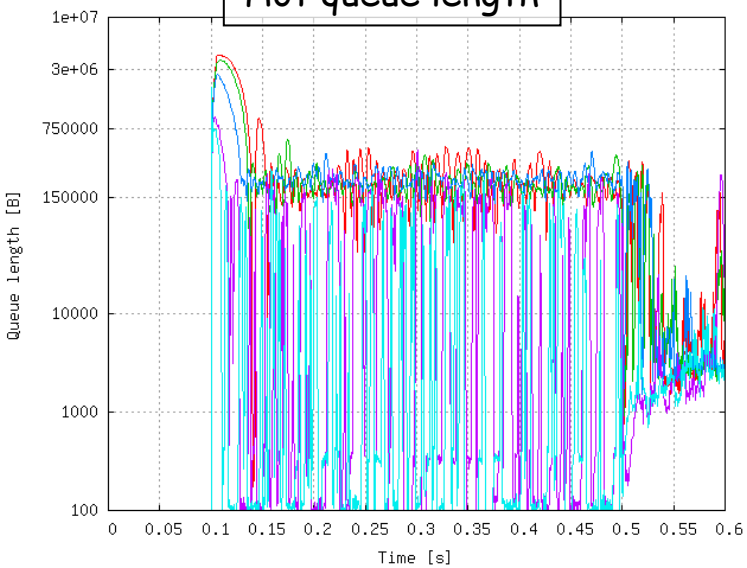
Aggregate throughput



Hot port throughput



Hot queue length

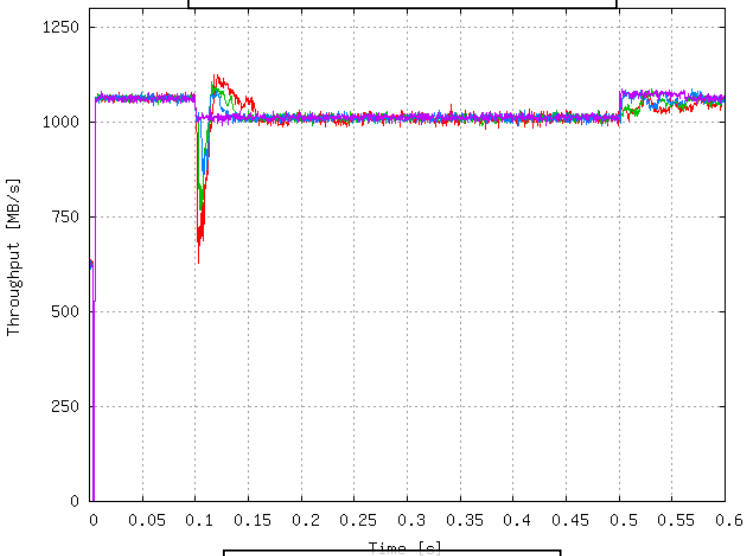


W = 0.1
W = 0.5
W = 2
W = 10
W = 50

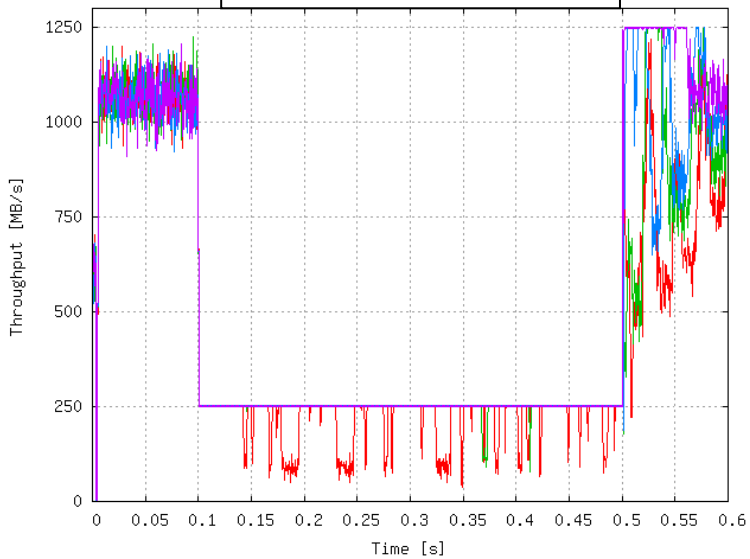
N = 16
G_d = 1.6667*10^{-6}
G_i = 6.6667*10^{-4}

SS-OG: Memory size sensitivity

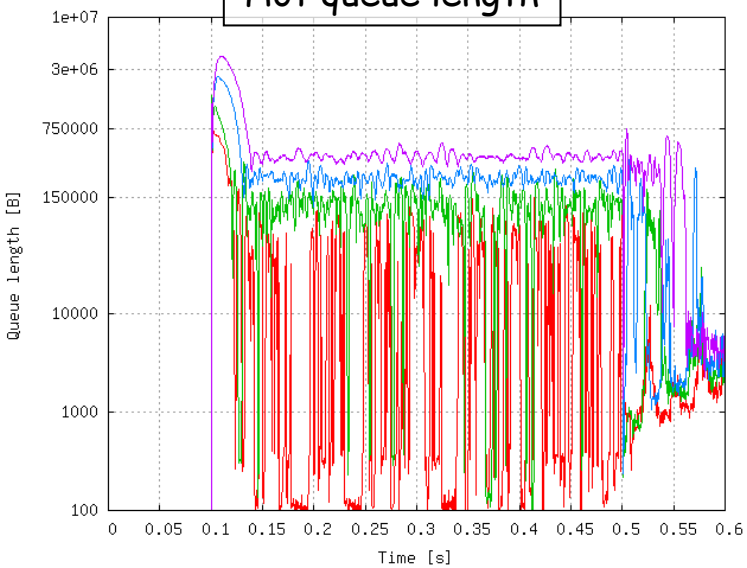
Aggregate throughput



Hot port throughput



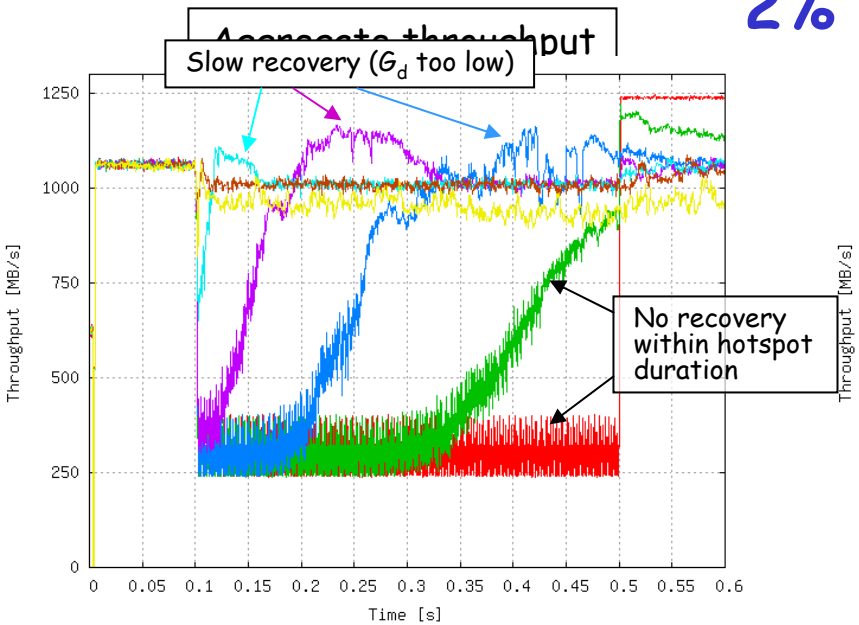
Hot queue length

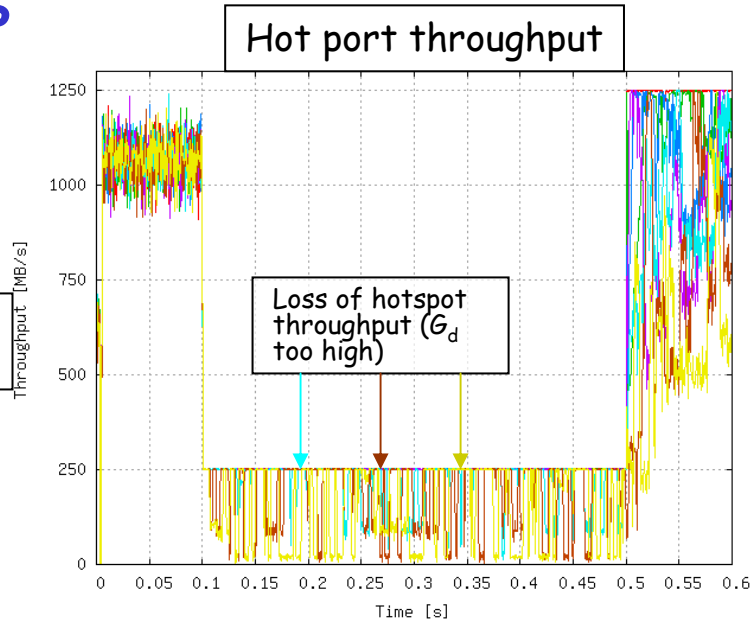


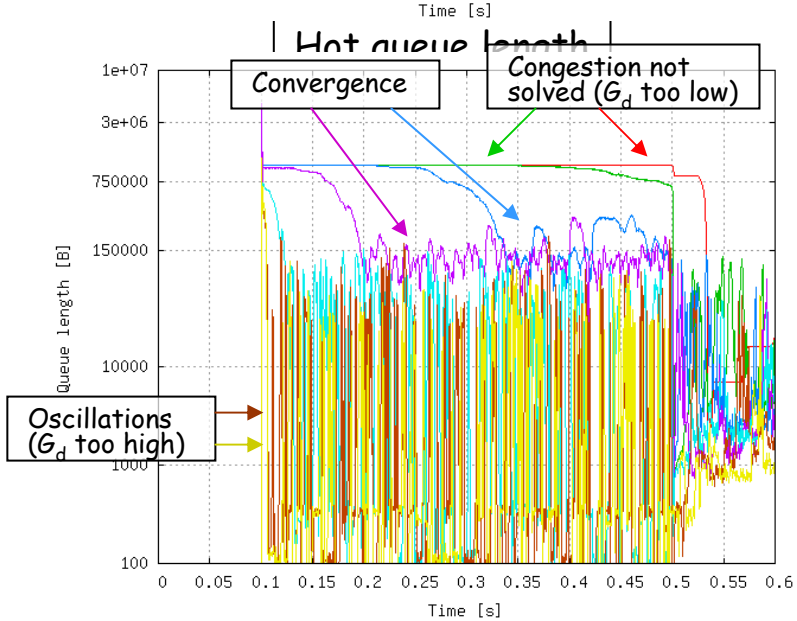
M = 150 KB/port
M = 300 KB/port
M = 600 KB/port
M = 1200 KB/port

N = 16
W = 2.0
$G_d = 1.6667*10^{-6}$
$G_i = 6.6667*10^{-4}$

SS-OG: G_d sensitivity, M = 150 KB/port, P_{sample} = 2%



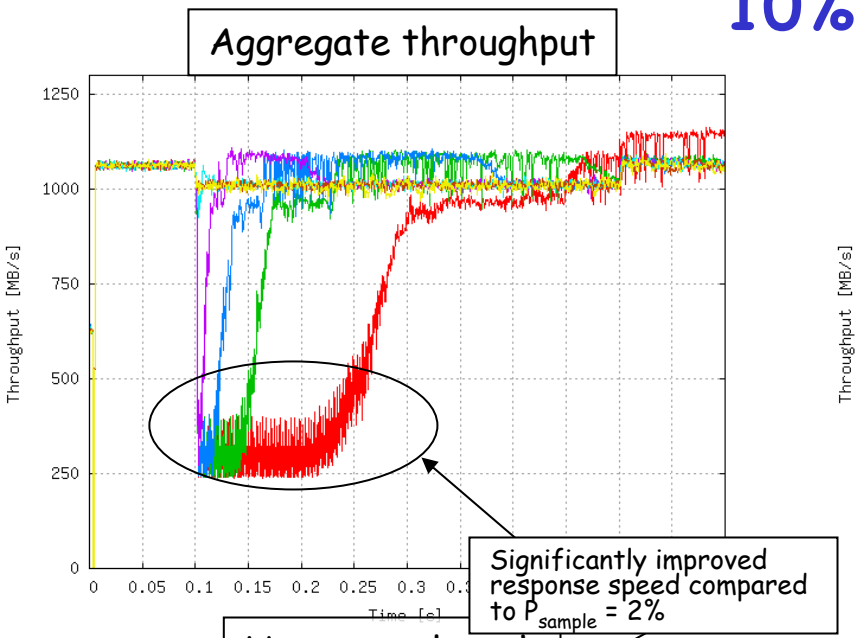


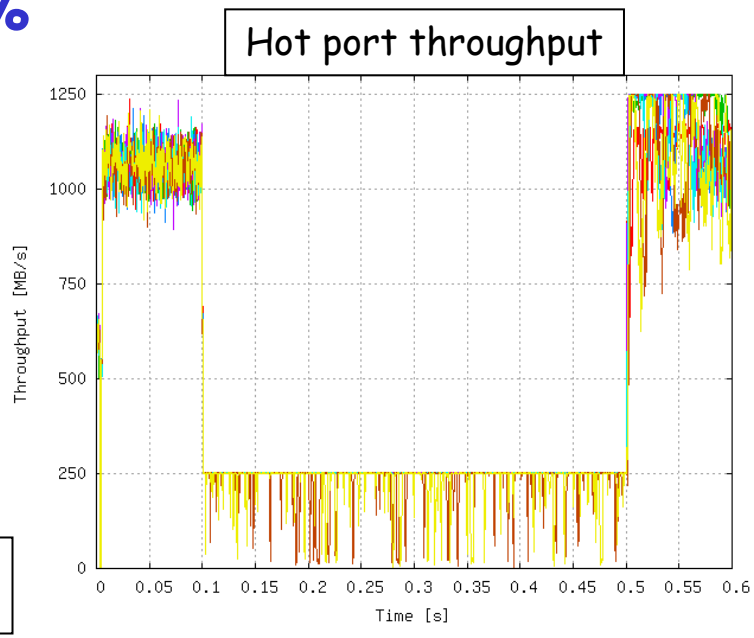


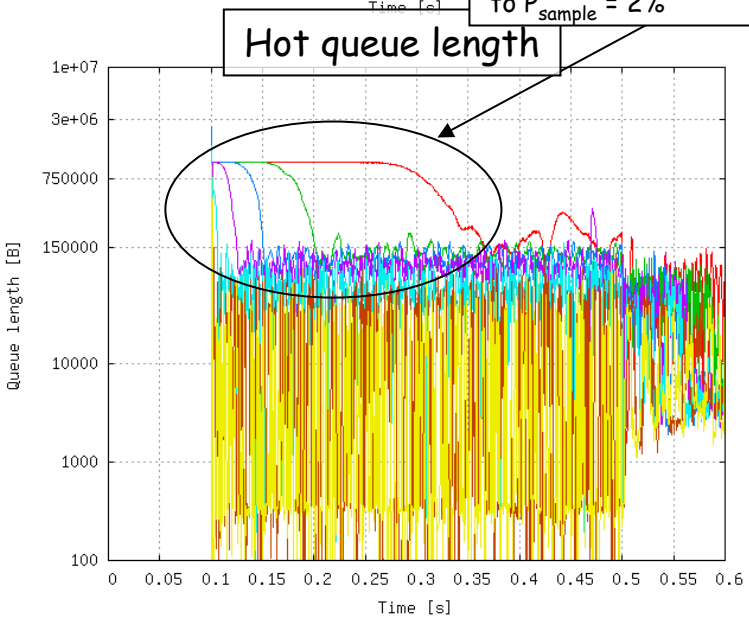
G_{d0} = 2.6667*10^{-6}
G_d = 0.10*G_{d0}
G_d = 0.25*G_{d0}
G_d = 0.50*G_{d0}
G_d = 1.0*G_{d0}
G_d = 2.5*G_{d0}
G_d = 5.0*G_{d0}
G_d = 10.0*G_{d0}

Q_{eq} = 37.5 KB
N = 16
W = 2.0

SS-OG: G_d sensitivity, M = 150 KB/port, P_{sample} = 10%



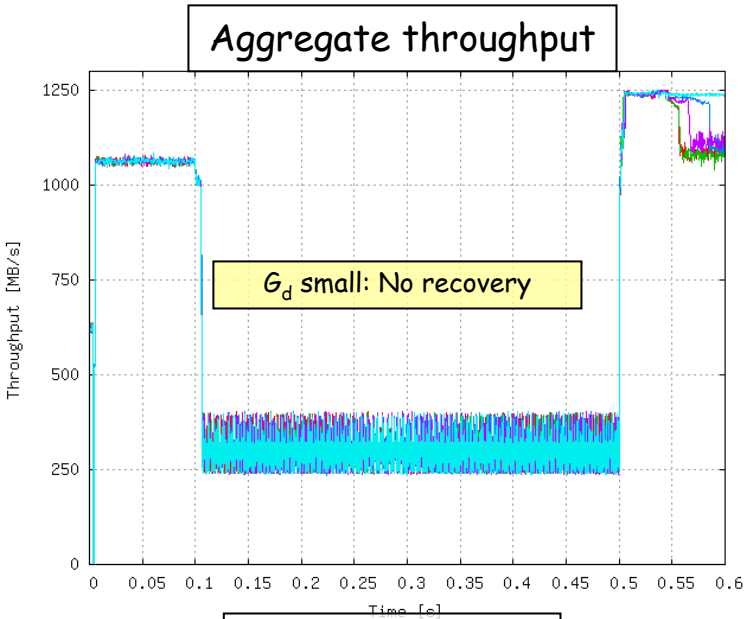


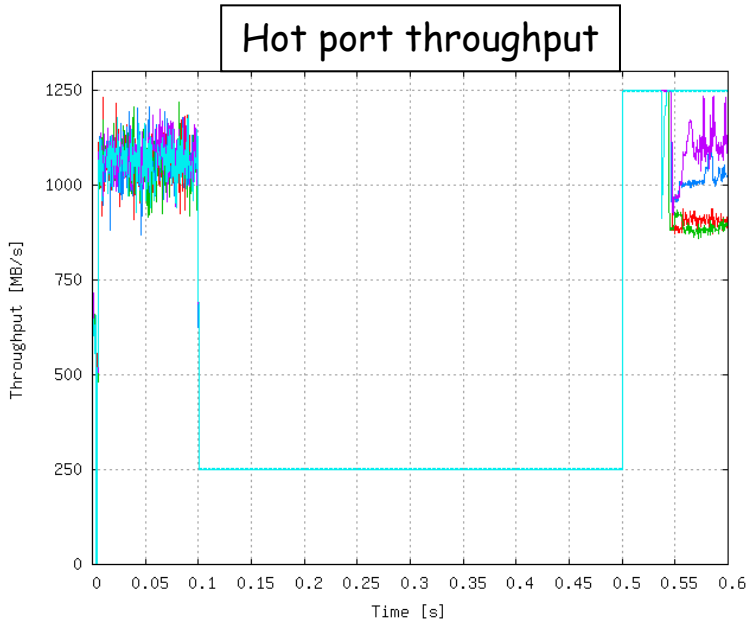


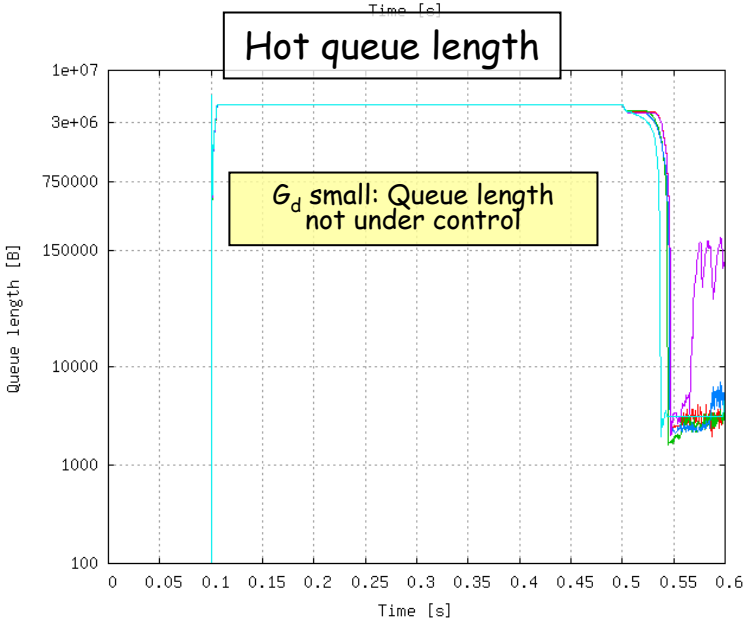
G_{d0} = 2.6667*10^{-6}
G_d = 0.10*G_{d0}
G_d = 0.25*G_{d0}
G_d = 0.50*G_{d0}
G_d = 1.0*G_{d0}
G_d = 2.5*G_{d0}
G_d = 5.0*G_{d0}
G_d = 10.0*G_{d0}

Q_{eq} = 37.5 KB
N = 16
W = 2.0

SS-OG: G_i sensitivity, M = 600 KB/port, G_d = 6.6667*10e^{-8}

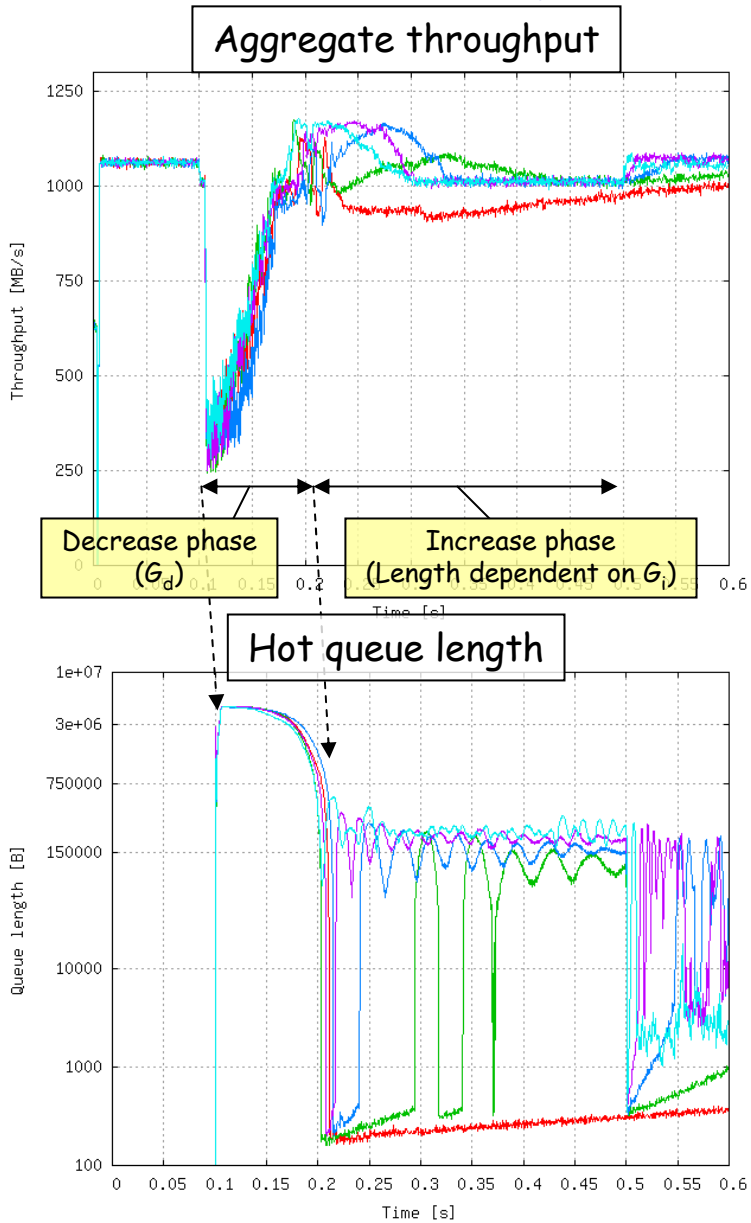


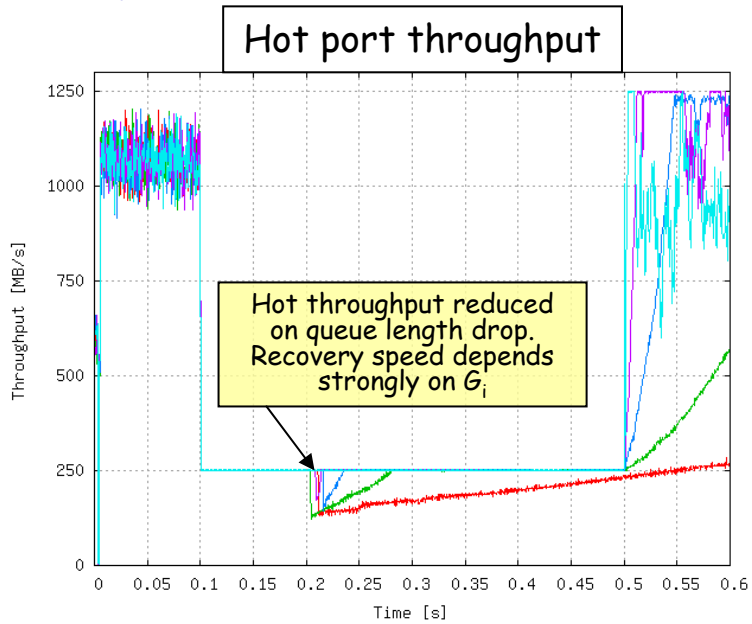




G_d = 6.6667*10^{-8}
G_i = 1*G_d
G_i = 5*G_d
G_i = 20*G_d
G_i = 100*G_d
G_i = 500*G_d

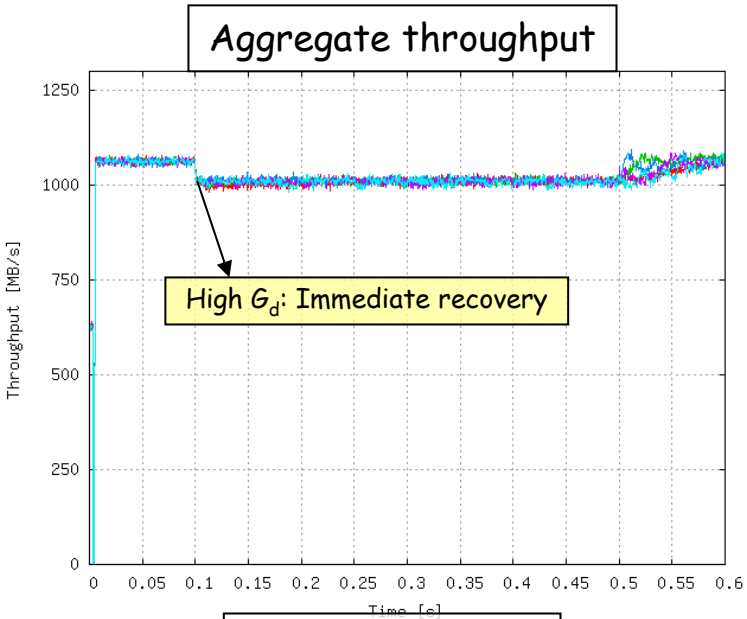
SS-OG: G_i sensitivity, M = 600 KB/port, G_d = 6.6667*10e^{-7}

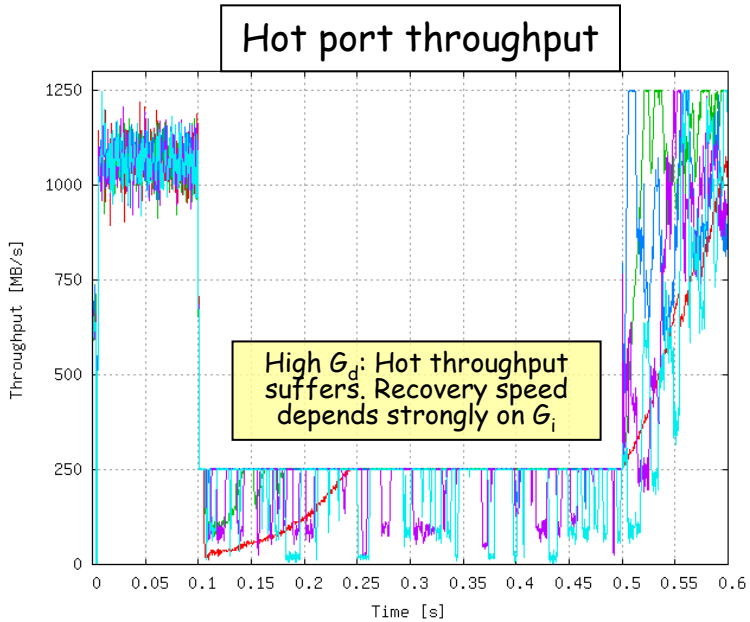


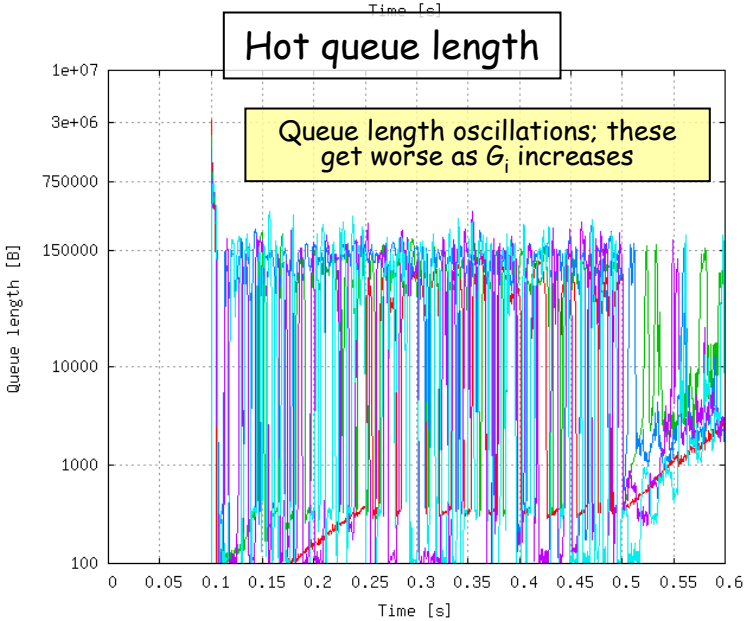


G_d = 6.6667*10^{-7}
G_i = 1*G_d
G_i = 5*G_d
G_i = 20*G_d
G_i = 100*G_d
G_i = 500*G_d

SS-OG: G_i sensitivity, M = 600 KB/port, G_d = 6.6667*10e^{-6}





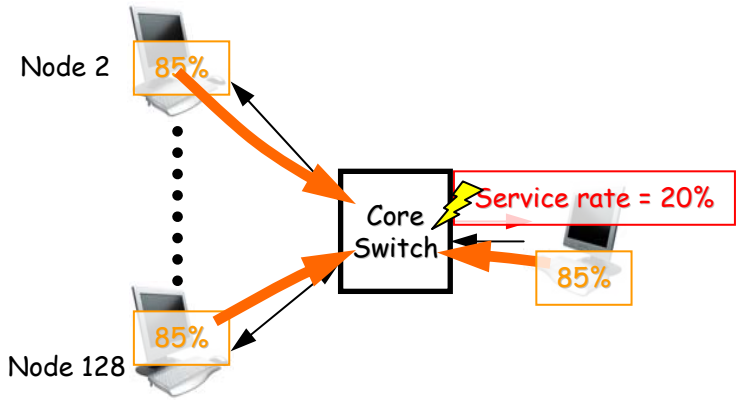


G_d = 6.6667*10^{-6}
G_i = 1*G_d
G_i = 5*G_d
G_i = 20*G_d
G_i = 100*G_d
G_i = 500*G_d

The Following Results will Focus on:
High-degree, Dual and Sweeping Hotspot Cases

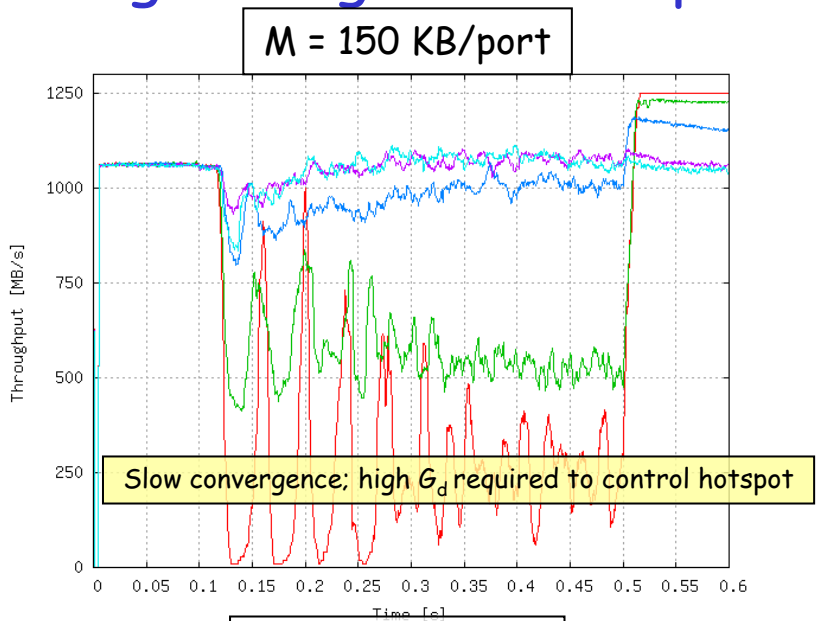
Case 4: High-degree single-stage OG hotspot

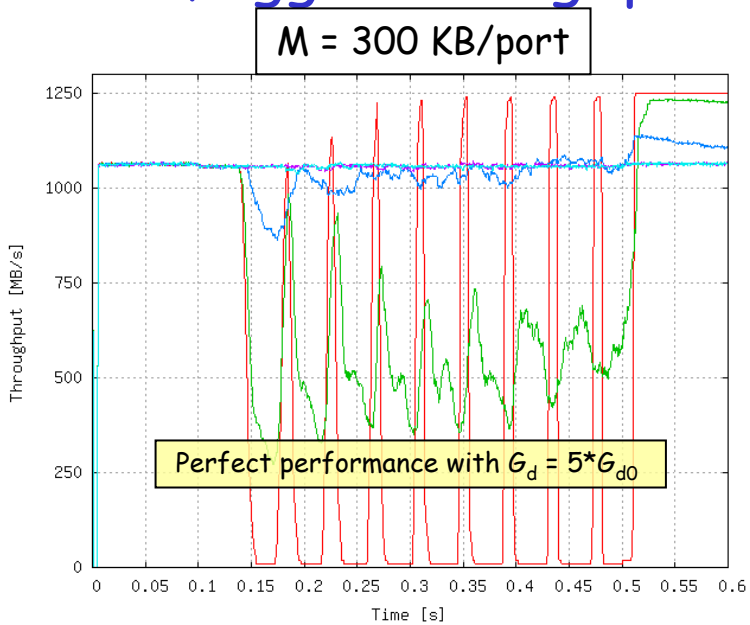
- 128 ports, single-stage
- Load = 85%
- Uniform destination distribution
- 1500 B frames
- Partitioned memory
- Lossless operation
  - PAUSE applied on a per input basis based on local high/low watermarks
  - $watermark_{high}$ = M – rtt*bw
  - $watermark_{low}$ = $watermark_{high}$ / 2

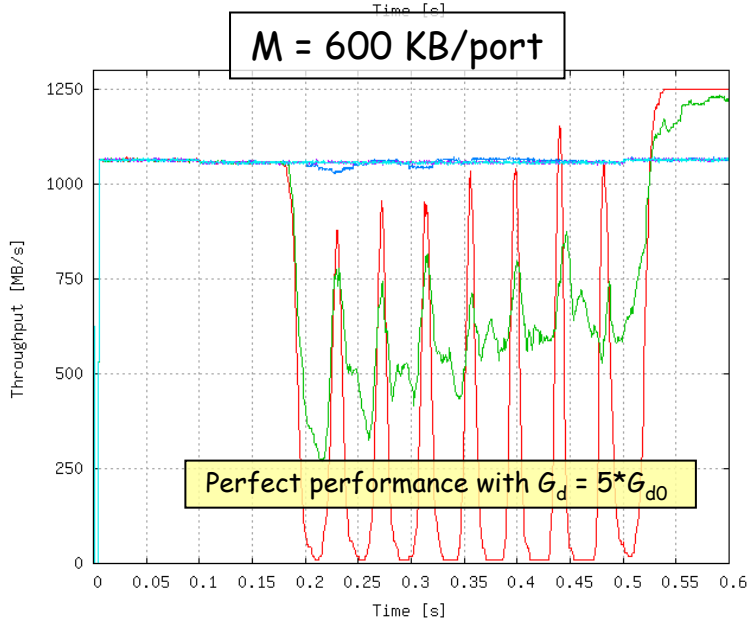


- W = 2.0
- Q_{eq} = M/4
- G_{d0} = 1 / ((2W+1) *Q_{eq}) = 4/(5*Q_{eq})
- G_d = [1, 2.5, 5, 10] * G_{do}
- G_i = 400*G_d
- R_u = R_{min} = 10 Mb/s
- No BCN(0,0), no BCN_MAX, no self-increase

Single-stage OG hotspot N = 128, aggr. throughput



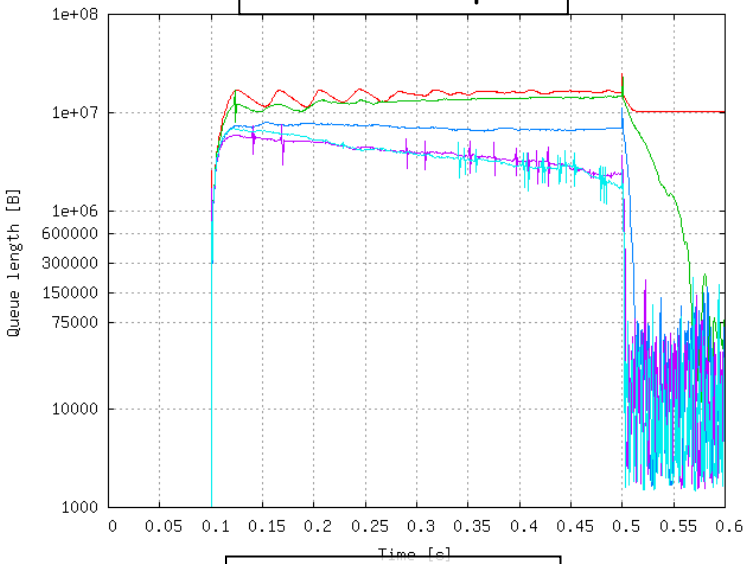




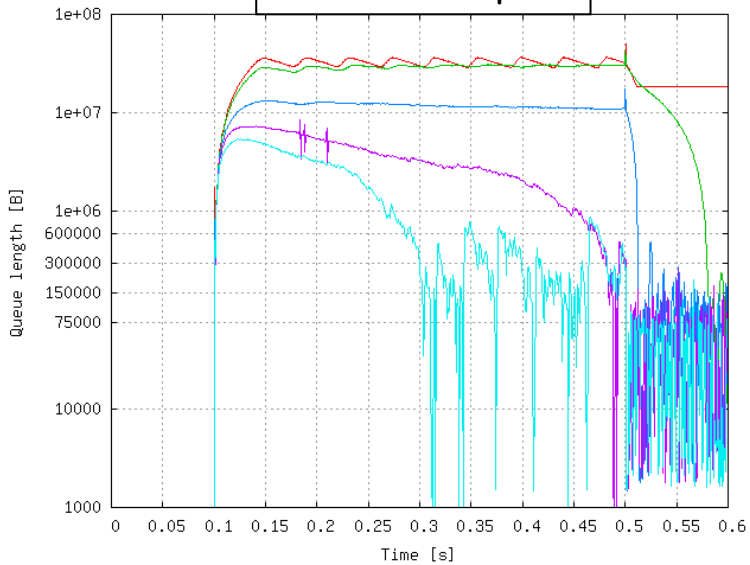
P_s = 0%
P_s = 2%, G_d = 1.0*G_{d0}
P_s = 2%, G_d = 2.5*G_{d0}
P_s = 2%, G_d = 5.0*G_{d0}
P_s = 2%, G_d = 10.0*G_{d0}

Single-stage OG hotspot N = 128, queue length

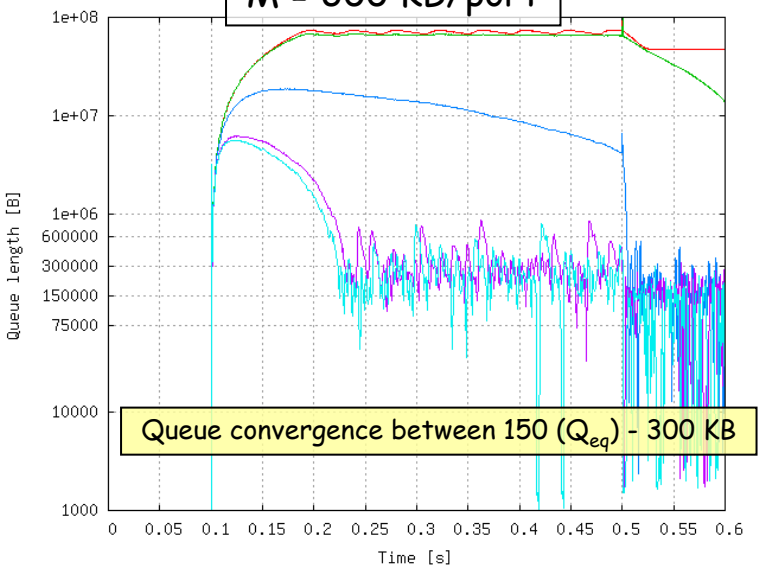
M = 150 KB/port



M = 300 KB/port



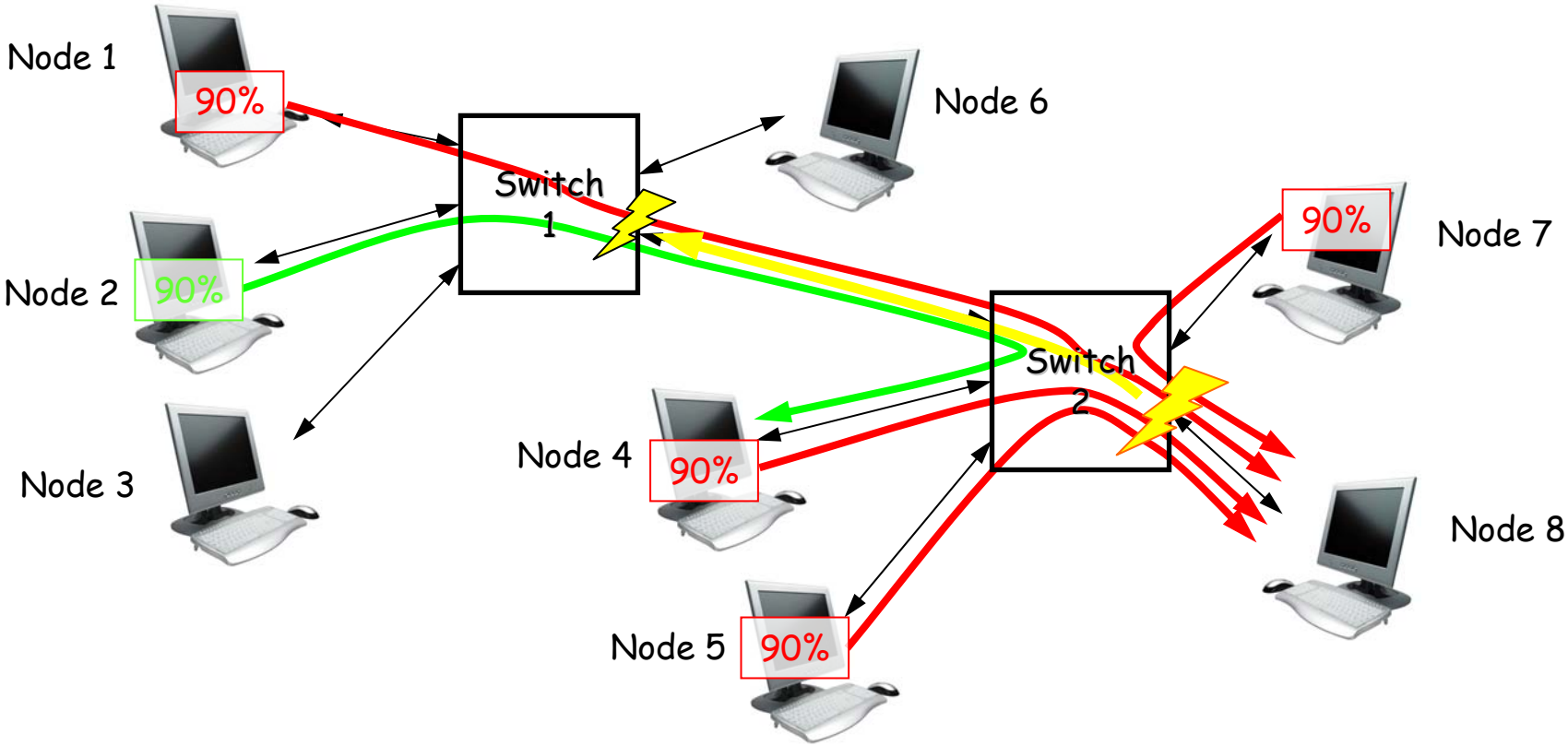
M = 600 KB/port



Queue convergence between 150 (Q_{eq}) - 300 KB

P_s = 0%
P_s = 2%, G_d = 1.0*G_{d0}
P_s = 2%, G_d = 2.5*G_{d0}
P_s = 2%, G_d = 5.0*G_{d0}
P_s = 2%, G_d = 10.0*G_{d0}

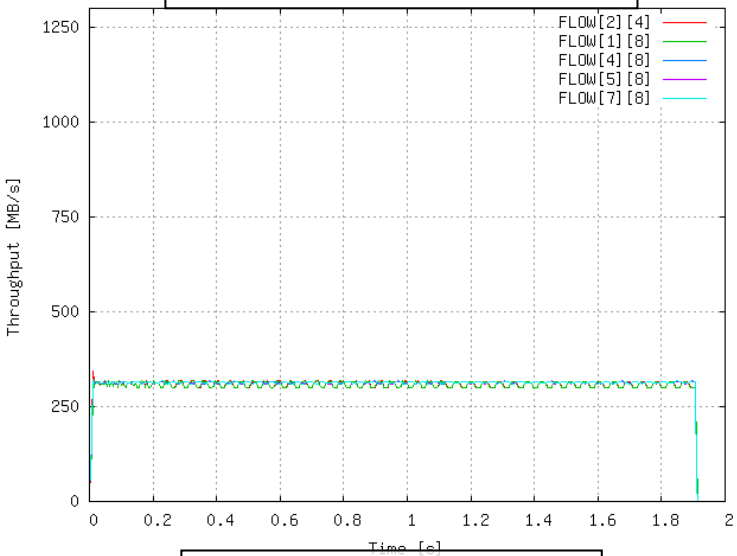
Case 5: Multi-Hop Dual Congestion Points (Light & Heavy)



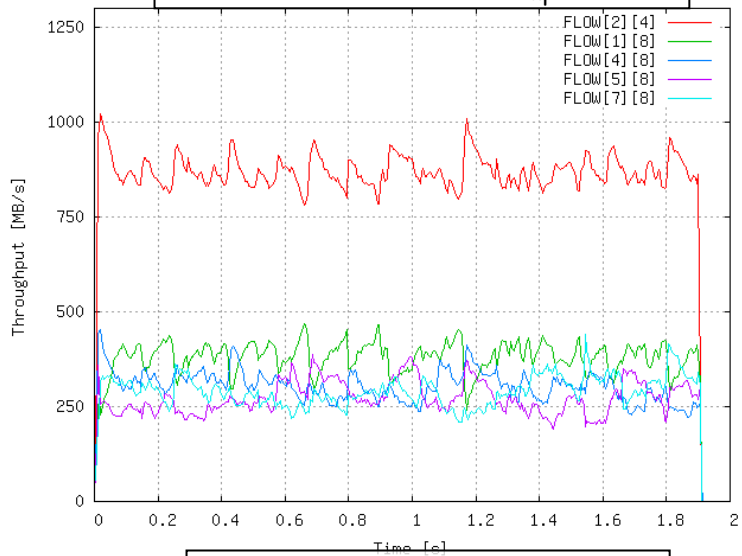
- Two switches, all links 10 Gb/s, no background traffic
- Four flows of 9 Gb/s each from nodes 1, 4, 5, 7 to node 8
- One flow of 9 Gb/s from node 2 to node 4
- Two congestion points
  - Port from switch 1 to switch 2
  - Port from switch 2 to node 8
- Fair allocation should provide 2.5 Gb/s for all flows to node 8 and 7.5 Gb/s for flow to node 4

Light/Heavy - Partitioned memory

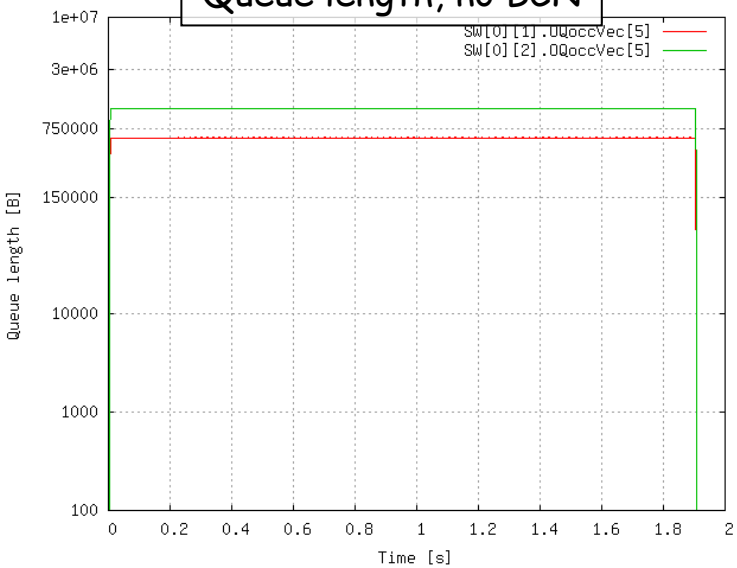
Flow throughput, no BCN



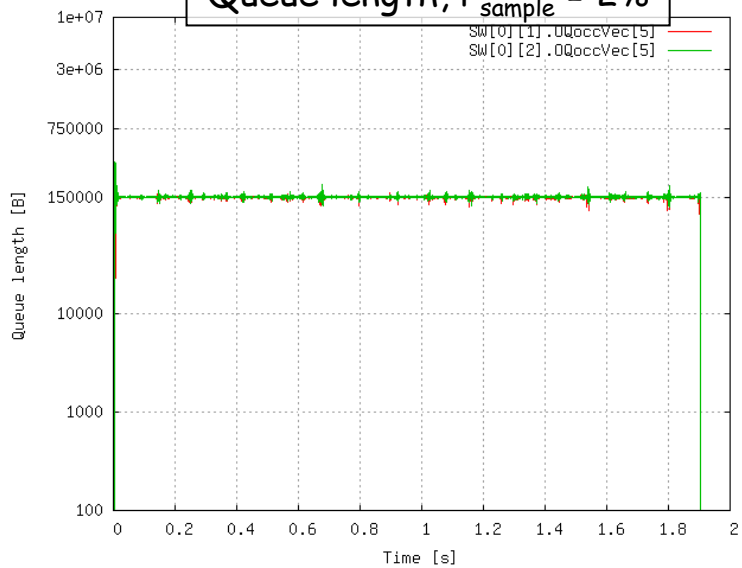
Flow throughput, P_{sample} = 2%



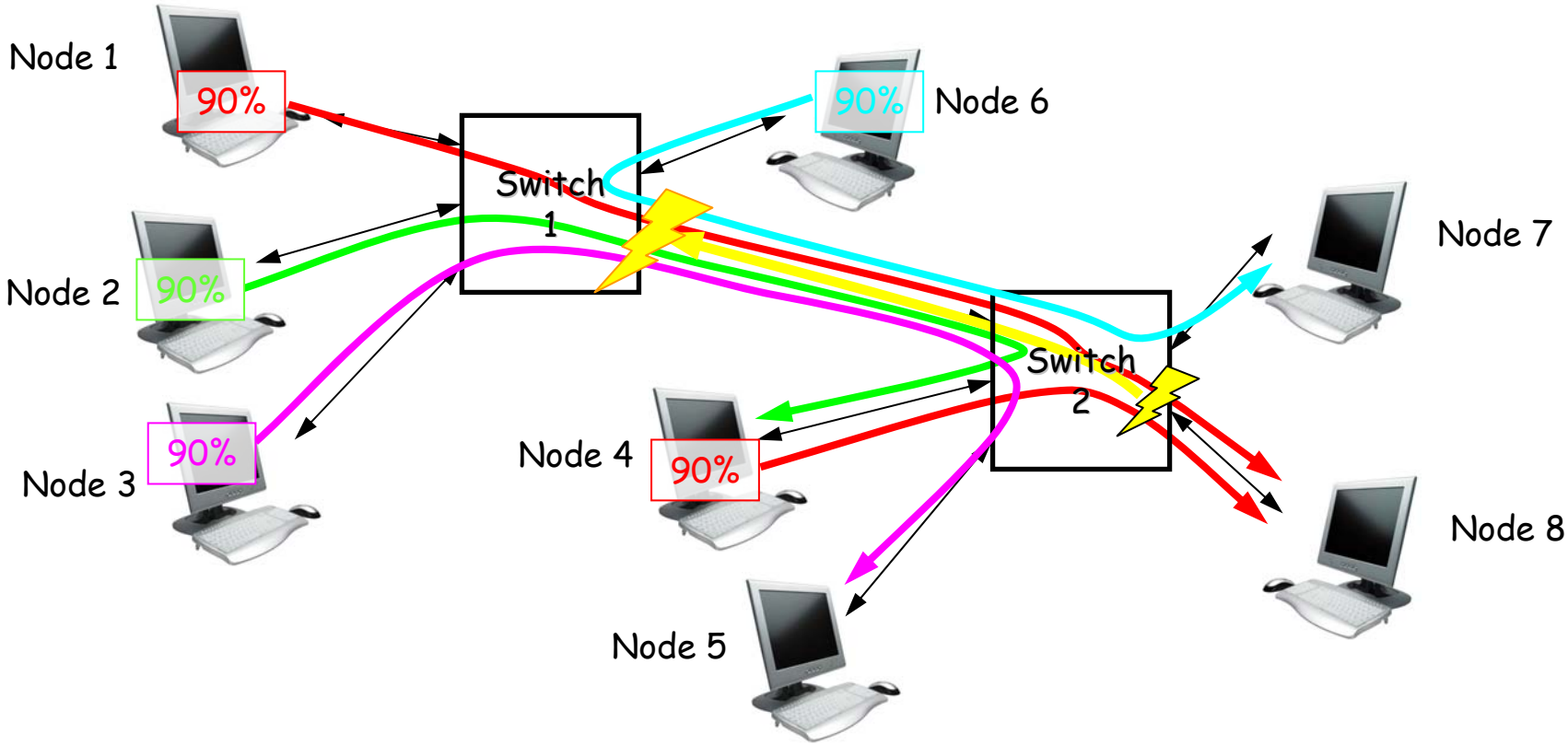
Queue length, no BCN



Queue length, P_{sample} = 2%



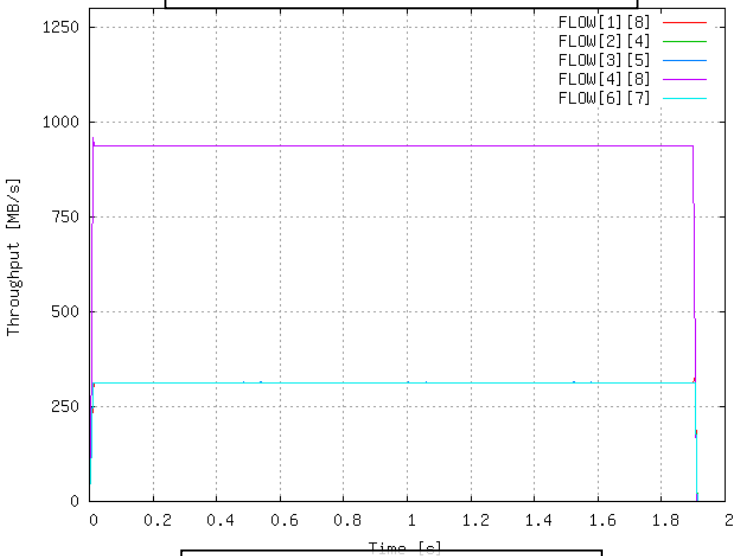
Case 6: Multi-Hop Dual Congestion Points (Heavy & Light)



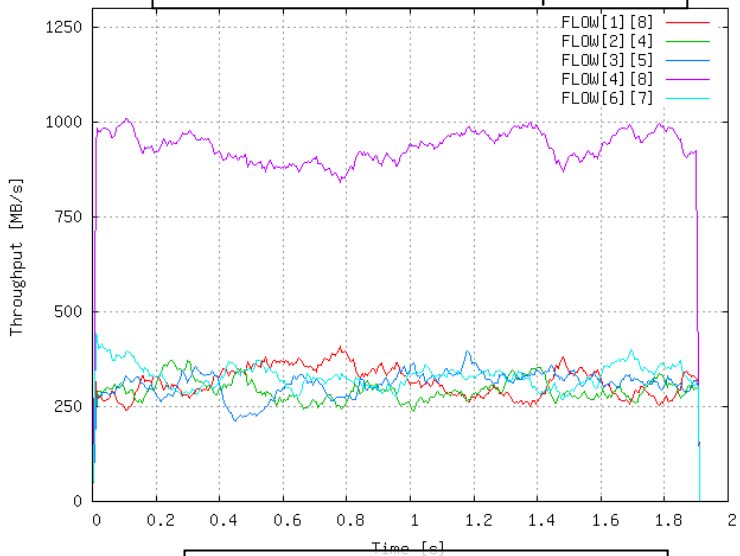
- Two switches, all links 10 Gb/s, no background traffic
- Two flows of 9 Gb/s each from nodes 1 and 4 to node 8
- Three flows of 9 Gb/s each from node 2 to node 4, 3 to 5, and 6 to 7
- Two congestion points
  - Port from switch 1 to switch 2
  - Port from switch 2 to node 8
- Fair allocation should provide 2.5 Gb/s for all flows to switch 2 and 7.5 Gb/s for flow from node 4 to node 8

Heavy/Light - Partitioned memory

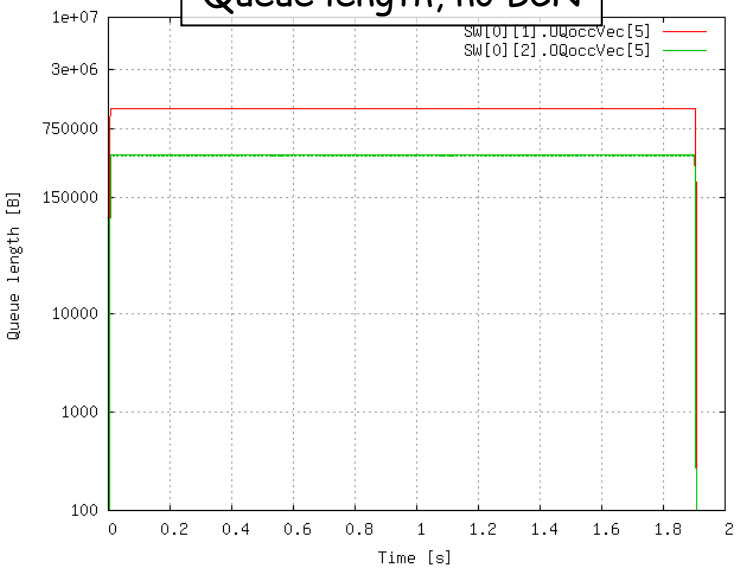
Flow throughput, no BCN



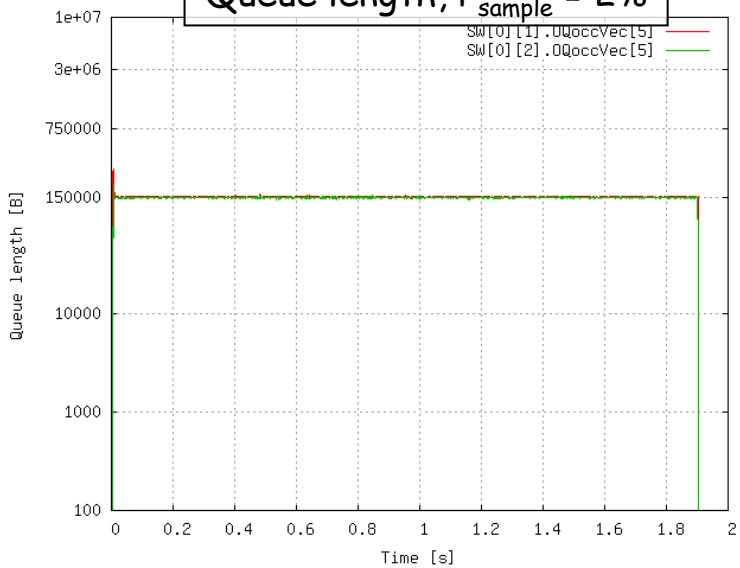
Flow throughput, P_{sample} = 2%



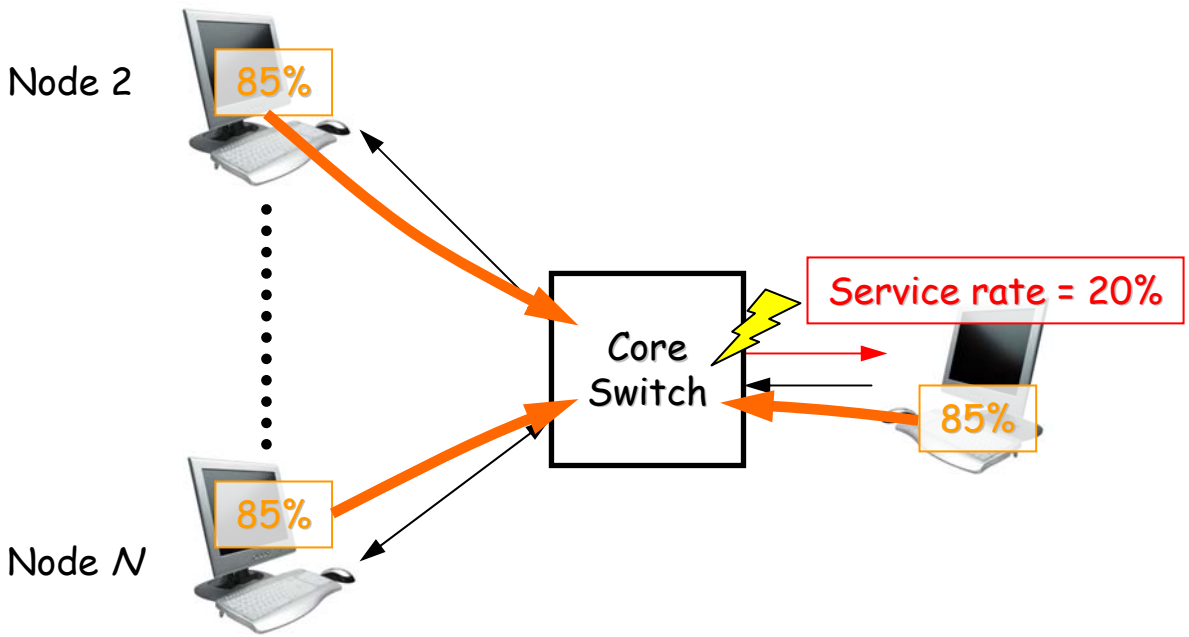
Queue length, no BCN



Queue length, P_{sample} = 2%



Case 7: Output-Generated Single-Hop Sweeping Hotspot

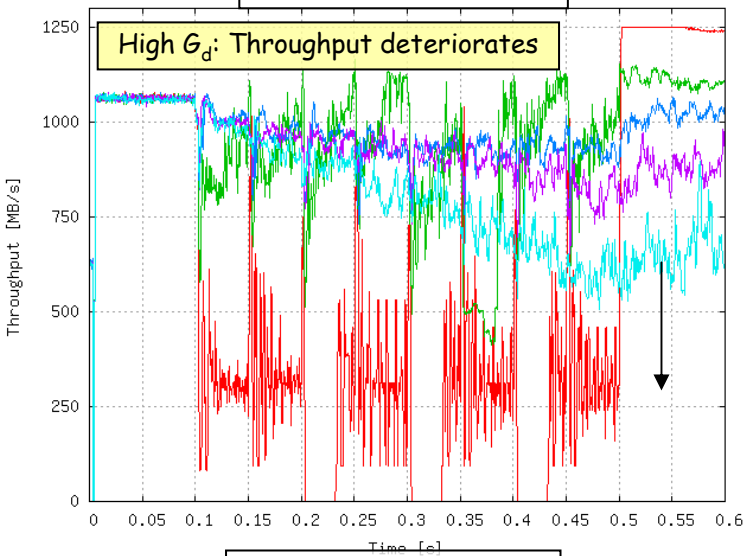


Time	Hotspot
[0.10-0.15]	Node 1
[0.15-0.20]	Node 2
[0.20-0.25]	Node 3
[0.25-0.30]	Node 4
[0.30-0.35]	Node 5
[0.35-0.40]	Node 6
[0.40-0.45]	Node 7
[0.45-0.50]	Node 8

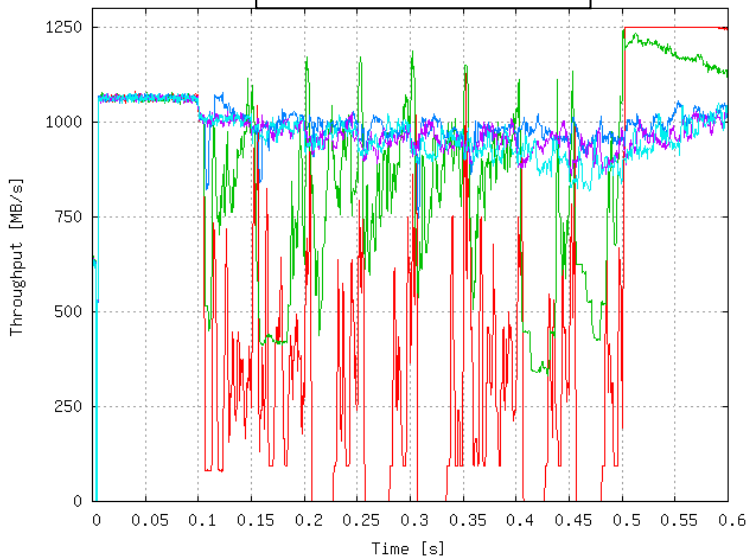
- All nodes: Uniform destination distribution, load = 85% (8.5 Gb/s)
- Hotspot moves every 50 ms from node 1 -> node 2 -> … -> node 8
  - Stress congestion control reaction speed
- Hot node service rate = 20%
- One congestion point
  - Hotspot degree = N-1
  - All flows affected

Output-Generated Single-Hop Sweeping Hotspot: N = 16

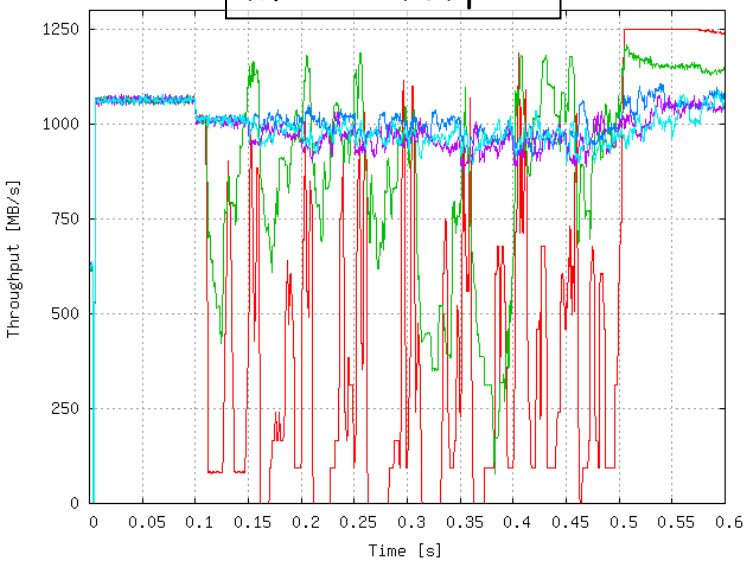
M = 150 KB/port



M = 300 KB/port



M = 600 KB/port

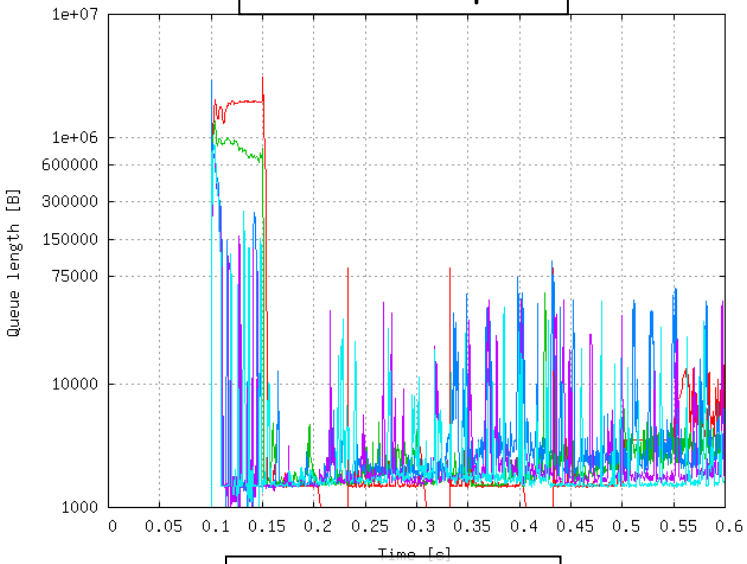


Aggregate throughput

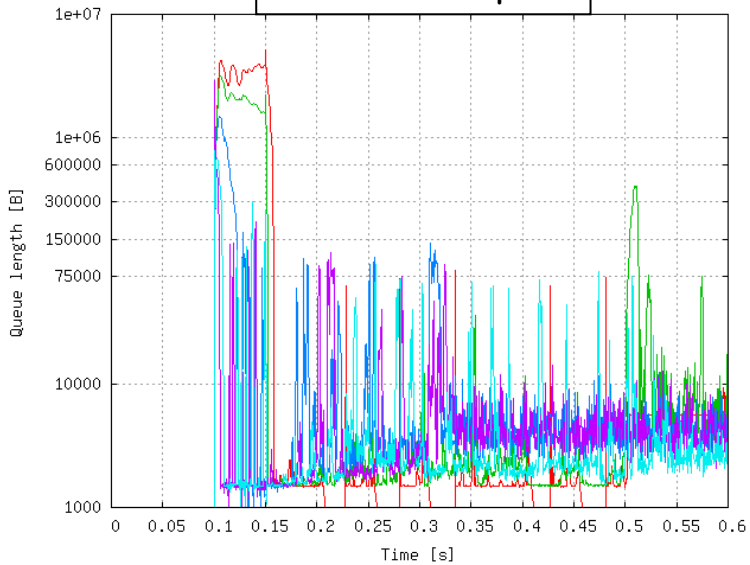
P_s = 0%
P_s = 2%, G_d = 1.0*G_{d0}
P_s = 2%, G_d = 2.5*G_{d0}
P_s = 2%, G_d = 5.0*G_{d0}
P_s = 2%, G_d = 10.0*G_{d0}

Output-Generated Single-Hop Sweeping Hotspot: N = 16

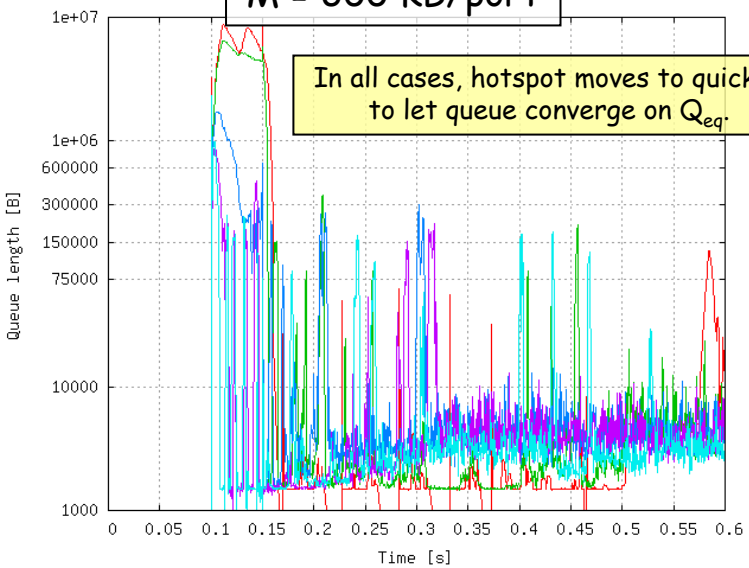
M = 150 KB/port



M = 300 KB/port



M = 600 KB/port



In all cases, hotspot moves to quickly to let queue converge on Q_{eq}.

Queue length

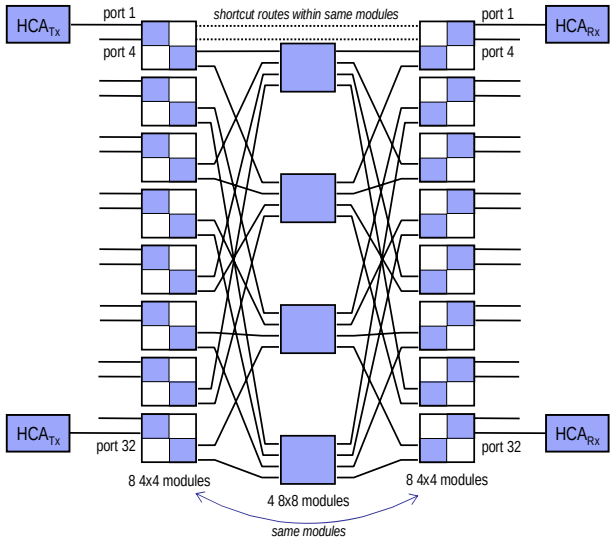
P_s = 0%
P_s = 2%, G_d = 1.0*G_{d0}
P_s = 2%, G_d = 2.5*G_{d0}
P_s = 2%, G_d = 5.0*G_{d0}
P_s = 2%, G_d = 10.0*G_{d0}

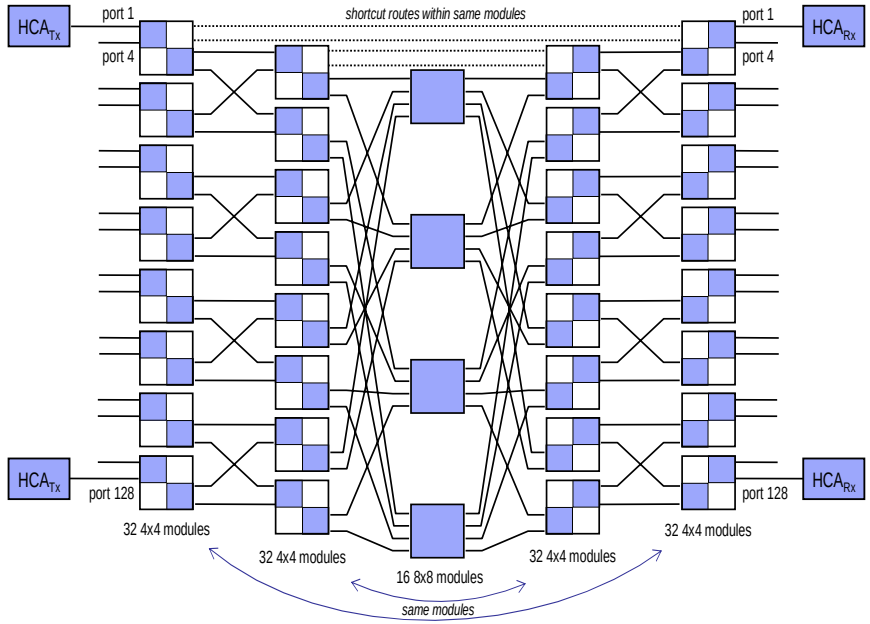
Next Simulation Steps

1. Dynamic flows: Markov-modulated and bursty
   a) Study transient response characterstics
   b) Aggregate Throughput
   c) Fairness
   d) Flow completion time

2. Move from small topologies to fat tree
   a) Initially a 3-hop, later 5-hop
   b) Agree on one "baseline" routing algorithm
   c) Solve the "congestion point" association issue

Baseline MIN Proposal: Bidir Fat Trees (FT)





- 2-level / 3-stage bidir MIN
- Simulate: 8 – 32 nodes
- Time per run: < 1hr

- 3-level / 5-stage bidir MIN
- Simulate: 128 – 2K nodes
- Time per run: TBD

Fat-trees: Scalable, w/ excellent routing and performance properties. Optimum performance/cost with current trends in technology. Can emulate any k-ary n-fly *and* n-cube topology. Large body of knowledge.

Conclusions

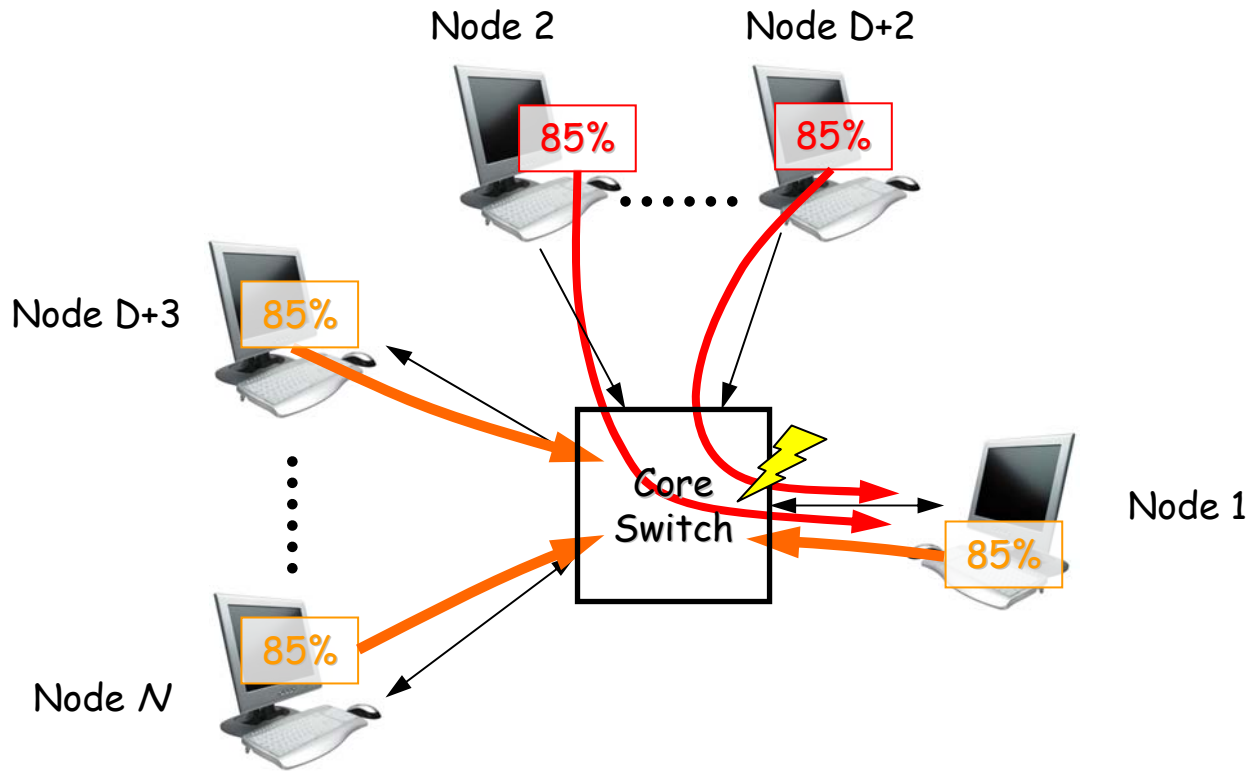
- BCN works for all the ZHB scenarios simulated
  - No surprises were found
  - Most findings are control systems-explainable

- Potential for improvement
  - Correct parameter setting remains open
    - particularly G_d and G_i require attention
  - Sampling remains also promising
    - Large improvements seem achievable

Contributors: Ronald Luijten

Backup

Input-Generated Single-Hop Hotspot

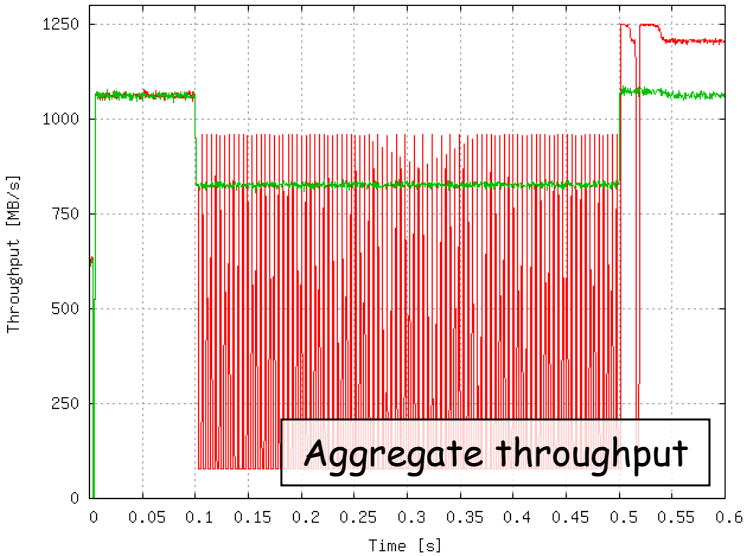


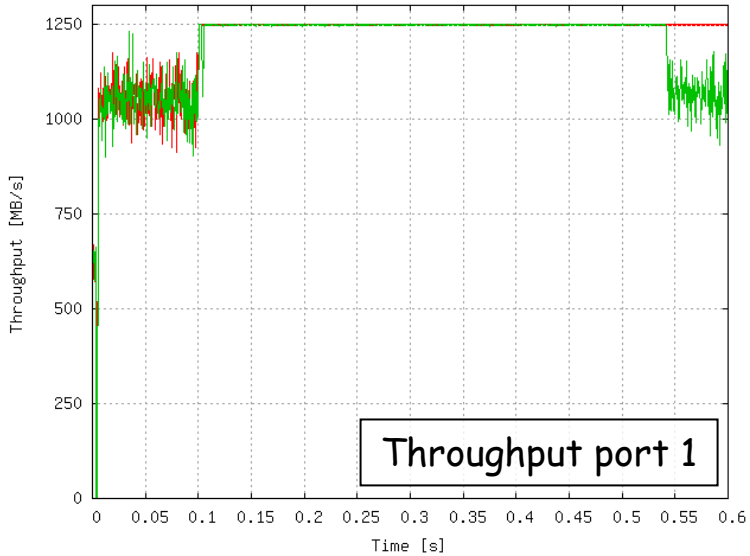
- Nodes 2 … D+2: All traffic to node 1
- Nodes 1, D+3 … N: Uniform destination distribution, load = 85% (8.5 Gb/s)
  - Uniform = same rate to all nodes except self; rate = load/(N-1)
- Number of hotspotting nodes = D
  - 1 < D < N
  - Hotspot degree = N-1: D heavy flows + N-D-1 light flows
- Results in one congestion point

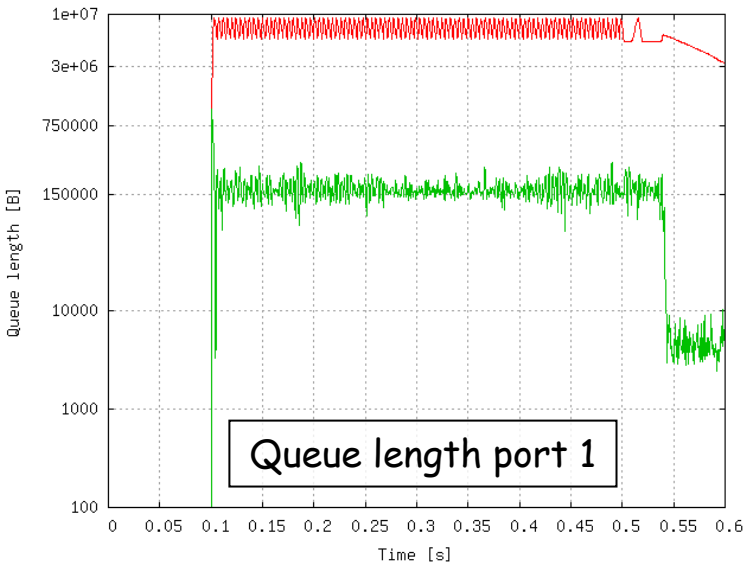
Simulation parameters

- Scenario
  - Single-hop input-generated hotspot
  - All nodes send at 85% loading
  - Four nodes target only hotspot
  - Remaining nodes generate uniform loading
    - Uniform = sending at same rate to all nodes except self
- Network
  - N = 16
  - M = 600 KB/port
  - Shared memory
    - PAUSE applied to all ports simultaneously based on global high/low watermarks
    - $watermark_{high}$ = N*(M – rtt*bw)
    - $watermark_{low}$ = $watermark_{high}$ / 2
  - Partitioned memory per input
    - Deadlock prevention
    - PAUSE applied on a per input basis based on local high/low watermarks
    - $watermark_{high}$ = M – rtt*bw
    - $watermark_{low}$ = $watermark_{high}$ / 2
- BCN
  - W = 2.0
  - G_i = 6.6667*10-4
  - G_d = 1.6667*10-6
  - Q_{eq} = 150 KB (= M/4)
  - P_{sample} = 2%
  - R_u = R_{min} = 10 Mb/s
  - No BCN(0,0) or BCN_MAX

Throughput & queue length – Shared memory



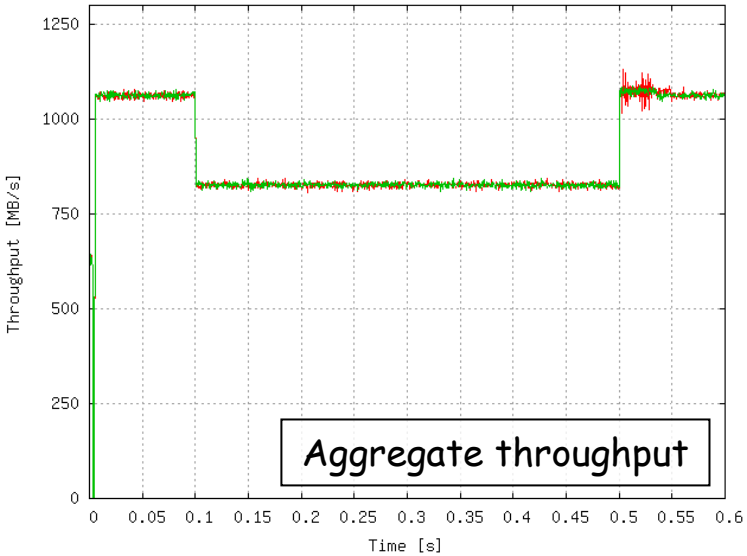


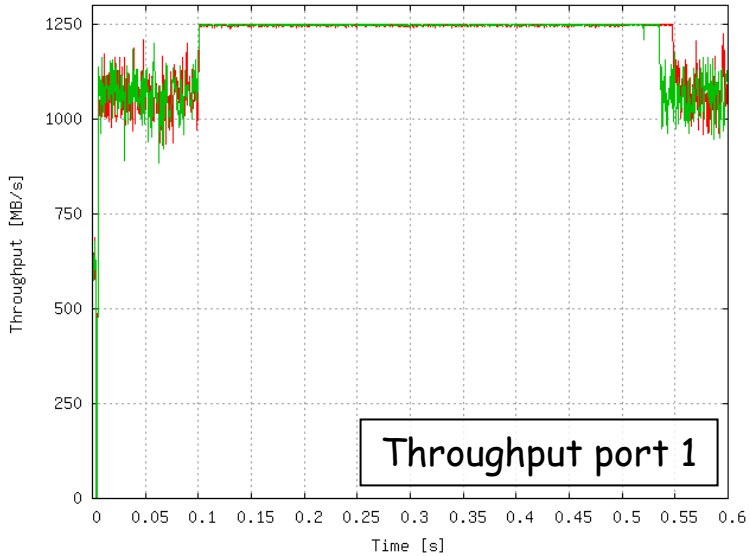


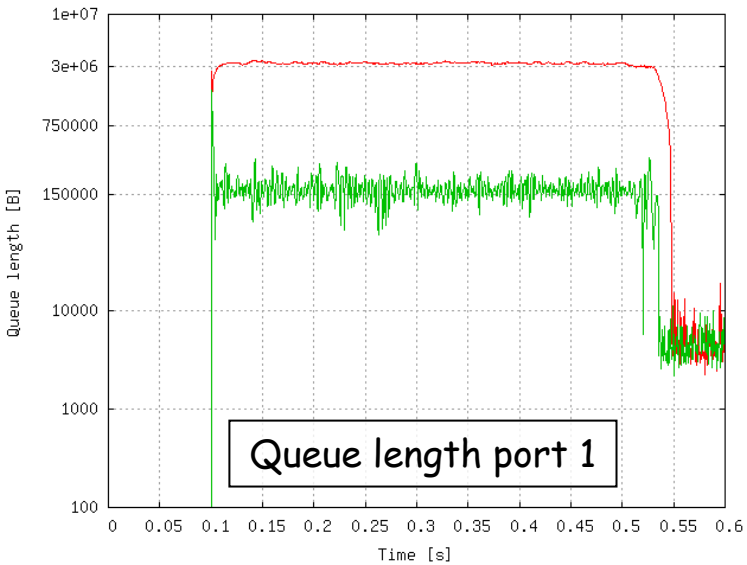
P_{sample} = 2%

No BCN

Throughput & queue length – Partitioned memory



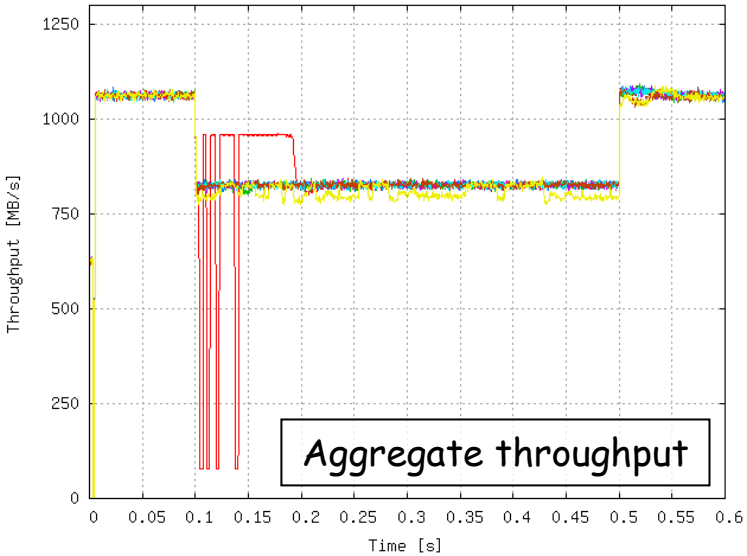


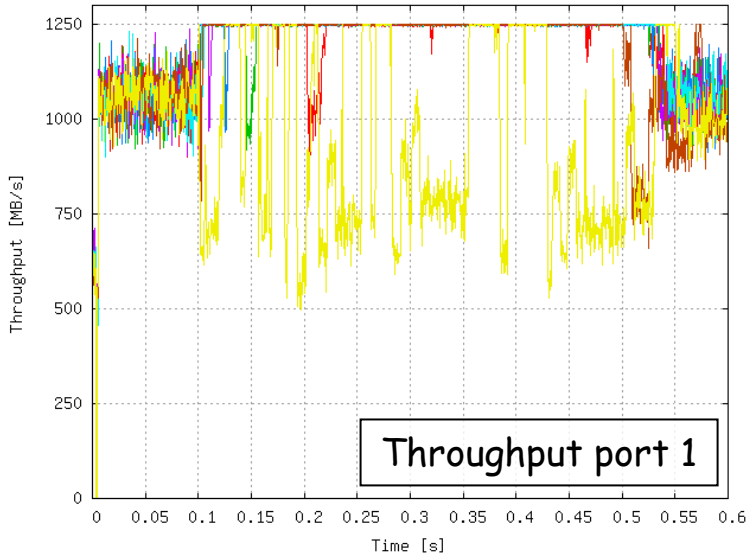


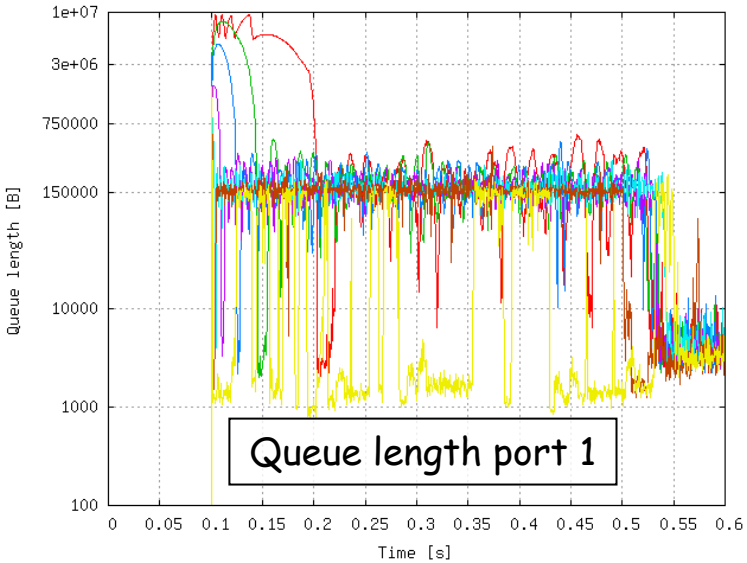
P_{sample} = 2%

No BCN

G_d sensitivity – Shared memory

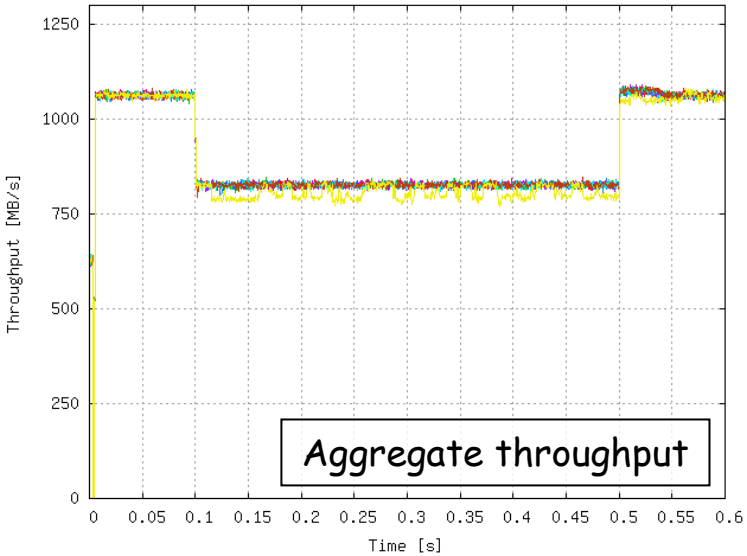


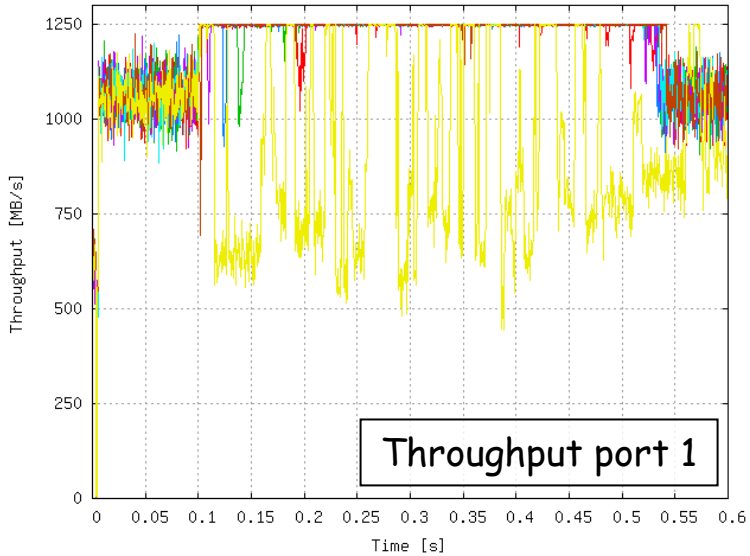


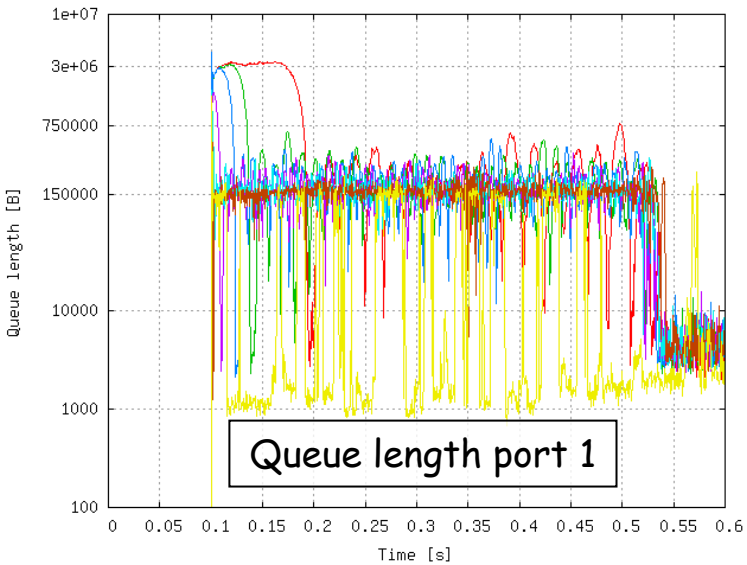


$G_{d0} = 6.6667*10^{-7}$
$G_d = 0.10*G_{d0}$
$G_d = 0.25*G_{d0}$
$G_d = 0.50*G_{d0}$
$G_d = 1.0*G_{d0}$
$G_d = 2.5*G_{d0}$
$G_d = 5.0*G_{d0}$
$G_d = 10.0*G_{d0}$

G_d sensitivity – Partitioned memory







$G_{d0} = 6.6667*10^{-7}$
$G_d = 0.10*G_{d0}$
$G_d = 0.25*G_{d0}$
$G_d = 0.50*G_{d0}$
$G_d = 1.0*G_{d0}$
$G_d = 2.5*G_{d0}$
$G_d = 5.0*G_{d0}$
$G_d = 10.0*G_{d0}$