Congestion Management Protocols
Simulation Results and Protocol Variations

Guenter Roeck, Teak Technologies

July 2007

- Independent protocol validation
- Determine performance of CP<->RP probing protocols

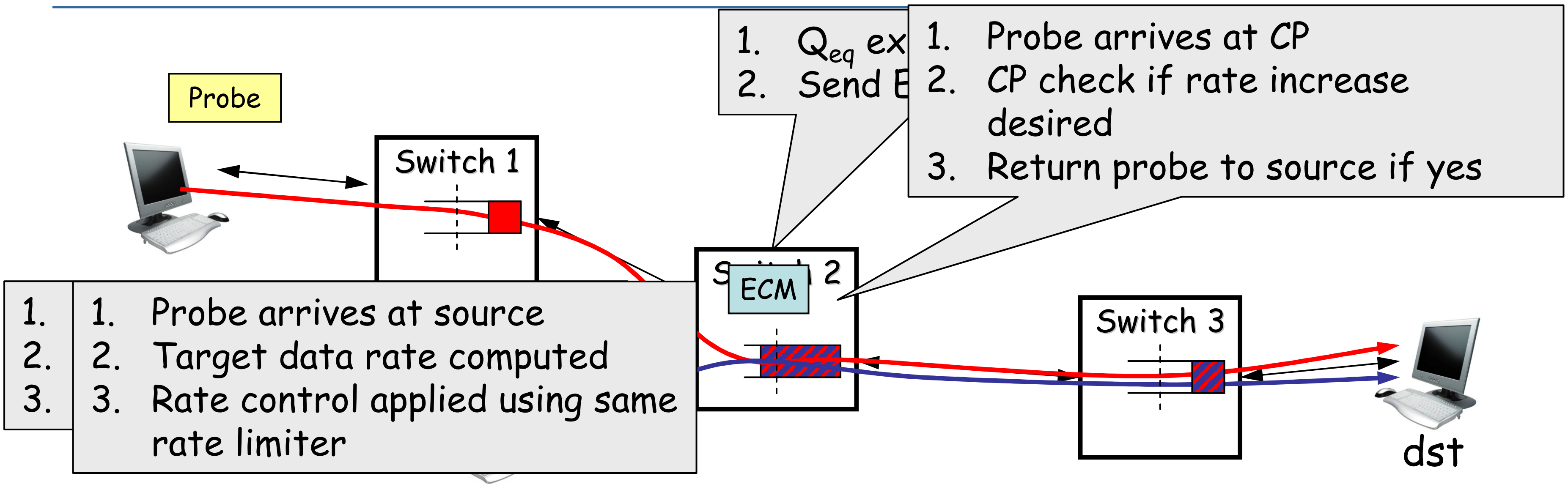
Outline

- Simulated Protocols
- Simulation Environment and Parameters
- Simulation Results
- Summary and Observations
- Conclusion

Simulated Protocols

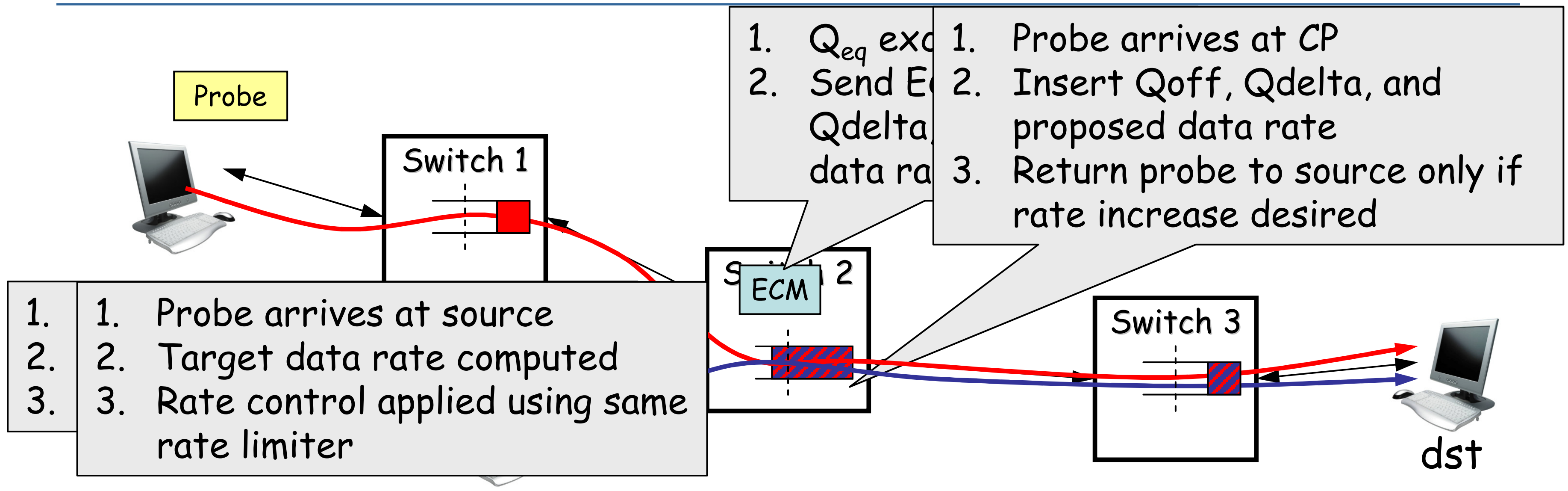
- ECM
  - As specified in au-bergamasco-ecm-v0.1.pdf
- E2CM
  - As specified in au-sim-IBM-ZRL-E2CM-proposal-r1.09b.ppt
- QCN
  - 2-point architecture
- FECN
  - As per March 2007 document, using probes (non-tagging)
- FECN-B
  - Modifications as proposed in Geneva (BCN-00, fast start)
- QCN-P
  - QCN 2-point architecture with added probes from RP to CP
- E2CM-P
  - Similar to ECM/E2CM, with probes from RP to CP replacing tags/path probing
- E2CM-PR, QCN-PR
  - Similar to E2CM-P/QCN-P, with added data rate guidance from CP to RP

QCN-P, E2CM-P Operation



- Probes sent for rate limited flows in regular intervals
- Probe destination address is most recent CP requesting a rate decrease
- Only rate limited flows are probed
- Probes sent as high priority frames

QCN-PR, E2CM-PR Operation



- Probes sent for rate limited flows in regular intervals
- Probe destination address is most recent CP requesting a rate decrease
- Only rate limited frames are probed
- Probes sent as high priority frames
- ECM packets and probe responses include suggested and maximum data rate in addition to Qoff, Qdelta
- Reaction Point takes suggested and maximum data rate into account when adjusting its transmit rate

Protocol Classification

Protocols requiring end-to-end support

	Tagging	Non-Tagging
Forward Notification	QCN-3 FECN	E2CM
Backward Notification	ECM, QCN-2	E2CM-P, E2CM-PR, QCN-P, QCN-PR

Notice

- While I have tried to implement all protocols as specified, there is no guarantee that I got everything right
- Simulations results reflect my implementation, not necessarily the intend of the protocol authors
- My sincere apologies if I got something wrong ...

- OMNET++
- INET framework
  - Added support for different CM protocols
    - Some 6,500 LOC total
    - Three weeks development time including simulation runs
  - Switch between protocols by changing configuration parameters

Simulation setup

- Two simulation runs per protocol, with different algorithm parameters
- Presenting only first set of results
  - 100 slides is bad enough ...
  - Second set of results typically does not change the trend
  - Results for second set of tests are typically better for most of the protocols
- Test topologies
  - Baseline test as proposed in au-sim-bergamasco-baseline-sim-scenario-092806v06.pdf
  - Tests 1-3, 5-8 as proposed in au-sim-wadekar-reqd-extended-sim-list-020807.pdf
    - No time to implement framework changes required to run test 4
  - Multi-hop test with several (7, 12) congestion points

- System parameters
  - Switch latency (processing time) = 1us
  - Link latency = 500ns
  - Switch frame capacity = 200 packets (300 kB)
  - Packet length = 1500 bytes
  - No PAUSE generated by switch
    - Did not have time to implement necessary framework changes
  - Using PAUSE to create output generated hotspots

Simulation Parameters

ECM Run 1

- Qeq = 375
- Qsc = 1600
- Qmc = 2400
- Qsat disabled
- Gi = 0.53333
- Gd = 0.00026667
- Ru = 1000000
- Rd = 1000000
- Td = 1ms
- Rmin = 1000000
- W = 2.0
- **samplingInterval = 150000**

ECM Run 2

- Qeq = 375
- Qsc = 1600
- Qmc = 2400
- Qsat disabled
- Gi = 0.53333
- Gd = 0.00026667
- Ru = 1000000
- Rd = 1000000
- Td = 1ms
- Rmin = 1000000
- W = 2.0
- **samplingInterval = 75000**

Simulation Parameters

E2CM, E2CM-P, E2CM-PR Run 1

- Qeq = 375
- Qsc = 1600
- Qmc = 2400
- Qsat disabled
- Gi = 0.53333
- Gd = 0.00026667
- Ru = 1000000
- Rd = 1000000
- Td = 1ms
- Rmin = 1000000
- W = 2.0
- flowQeq = 15000
- rateTimer = 1ms [PR]
- switchRateWeight = 0.02 [PR]
- **samplingInterval = 150000**
- **probeInterval = 100000 [P, PR]**

E2CM, E2CM-P, E2CM-PR Run 2

- Qeq = 375
- Qsc = 1600
- Qmc = 2400
- Qsat disabled
- Gi = 0.53333
- Gd = 0.00026667
- Ru = 1000000
- Rd = 1000000
- Td = 1ms
- Rmin = 1000000
- W = 2.0
- flowQeq = 15000
- rateTimer = 1ms [PR]
- switchRateWeight = 0.02 [PR]
- **samplingInterval = 75000**
- **probeInterval = 50000 [P, PR]**

Simulation Parameters

FECN, FECN-B

- N0 = 10
- A = 1.1
- B = 1.002
- C = 0.1
- Alpha = 0.5
- minRate = 10000000
- Qeq = 192000 (bits)
- Qsc = 960000 (bits)
- T = 1ms

Simulation Parameters

QCN Run 1

extraFastRecovery = true

fastRecoveryThreshold = 5

hyperactiveIncrease = true

driftFactor = 1.0005

Gd = 0.0078125 (1/128)

timerPeriod = 200uS

minRate = 10000000

minDecFactor = 0.5

EfrMax = 1000000

A = 12000000

Qeq = 24000

W = 2.0

baseProbability = 1%

toThreshold = 150000

QCN Run 2

extraFastRecovery = true

fastRecoveryThreshold = 5

hyperactiveIncrease = true

driftFactor = 1.0005

Gd = 0.0078125 (1/128)

timerPeriod = 200uS

minRate = 10000000

minDecFactor = 0.5

EfrMax = 1000000

A = 12000000

Qeq = 24000

W = 2.0

baseProbability = 2%

toThreshold = 75000

Simulation Parameters

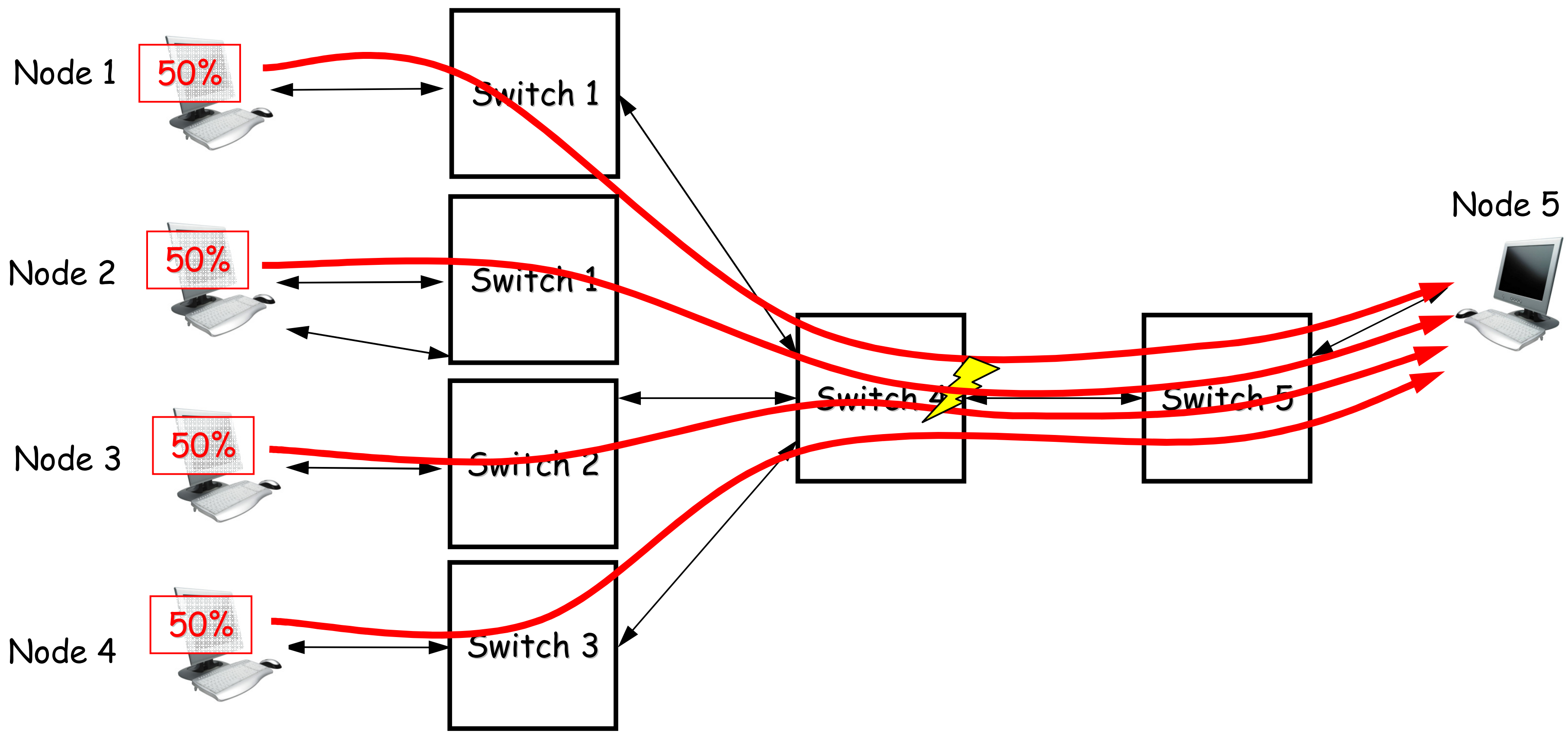
QCN-P, QCN-PR Run 1

extraFastRecovery = true
fastRecoveryThreshold = 5
hyperactiveIncrease = true
Gd = 0.0078125 (1/128)
timerPeriod = 200uS
minRate = 10000000
minDecFactor = 0.5
EfrMax = 1000000
A = 12000000
Qeq = 24000
W = 2.0
selfIncrease = 1000000 [P]
selfIncreaseFactor = 0.1% [PR]
rateT = 1ms [PR]
switchRateWeight = 0.002 [PR]
baseProbability = 1%
toThreshold = 100000

QCN-P, QCN-PR Run 2

extraFastRecovery = true
fastRecoveryThreshold = 5
hyperactiveIncrease = true
Gd = 0.0078125 (1/128)
timerPeriod = 200uS
minRate = 10000000
minDecFactor = 0.5
EfrMax = 1000000
A = 12000000
Qeq = 24000
W = 2.0
selfIncrease = 1000000 [P]
selfIncreaseFactor = 0.1% [PR]
rateT = 1ms [PR]
switchRateWeight = 0.002 [PR]
baseProbability = 2%
toThreshold = 50000

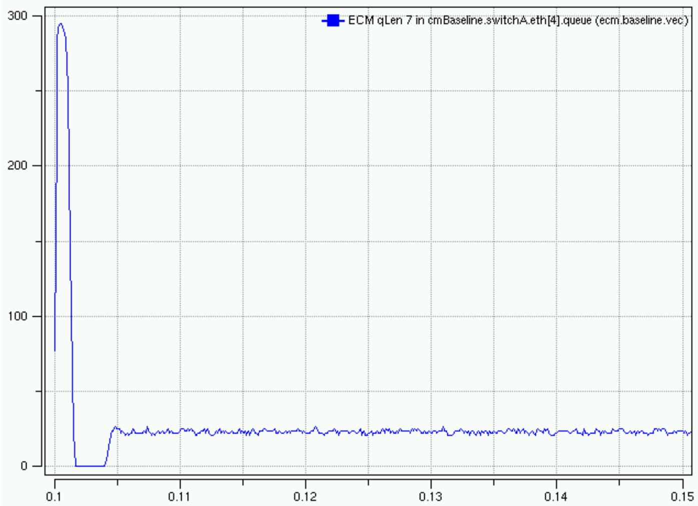
Baseline: Symmetric Topology, Single Hotspot



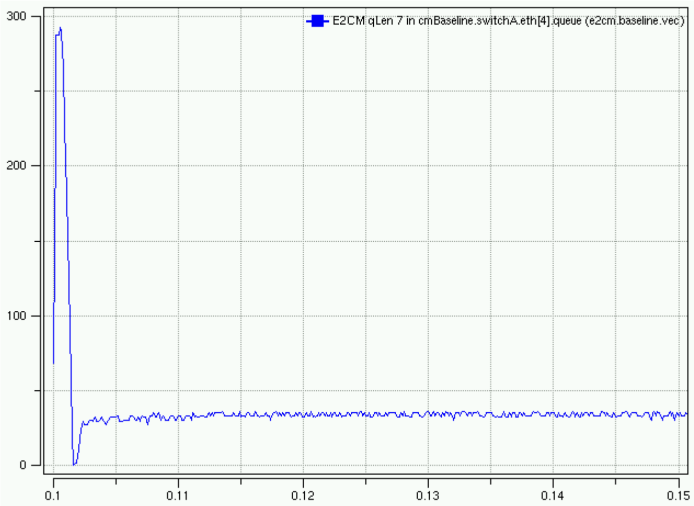
- Node 1 to 4 sending at 50% load to node 5

Baseline: Queue Length

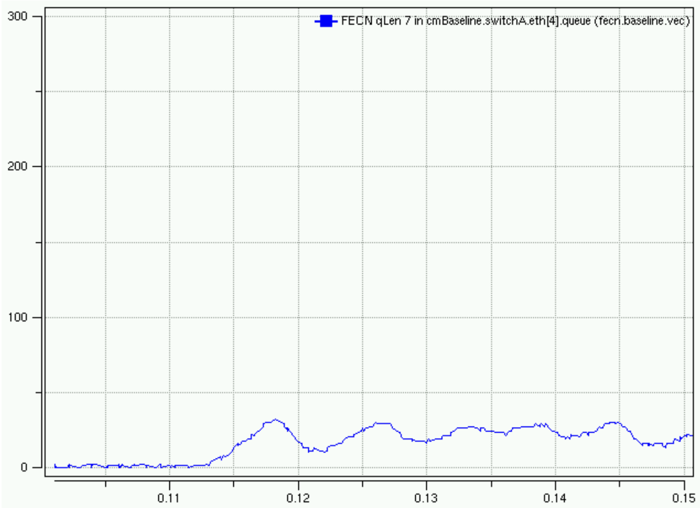
ECM



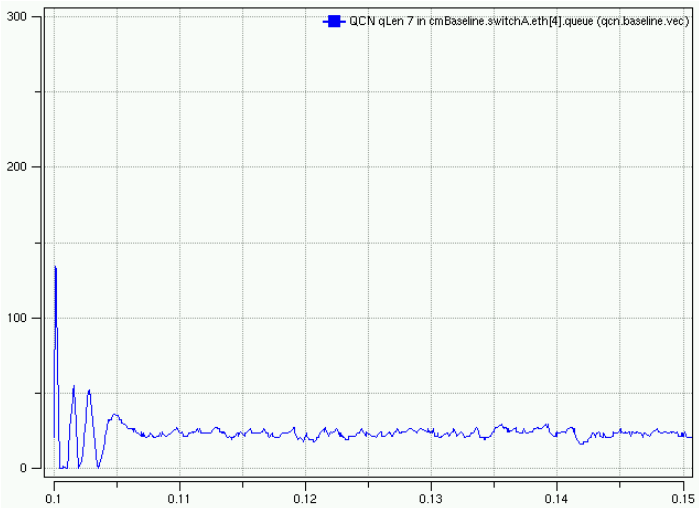
E2CM



FECN

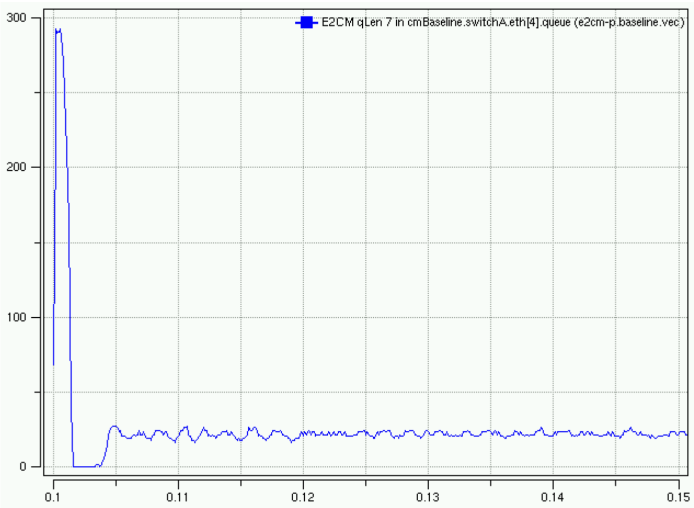


QCN

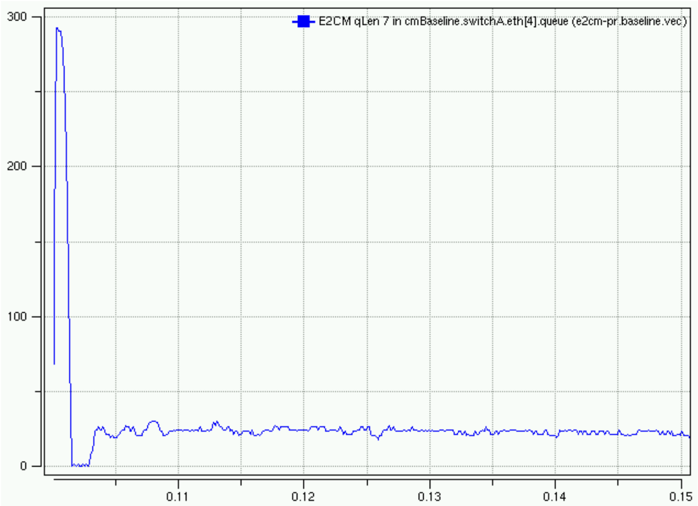


Baseline: Queue Length

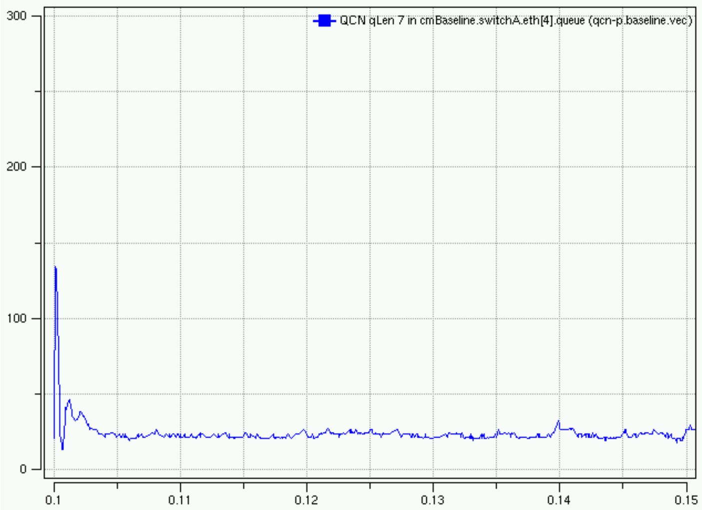
E2CM-P



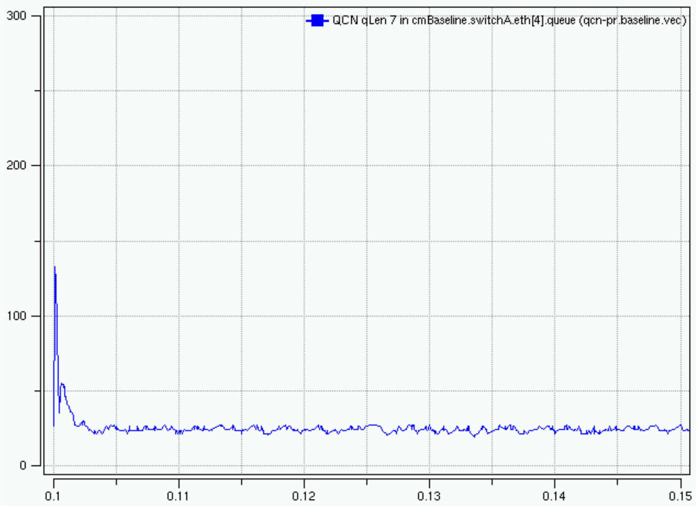
E2CM-PR



QCN-P

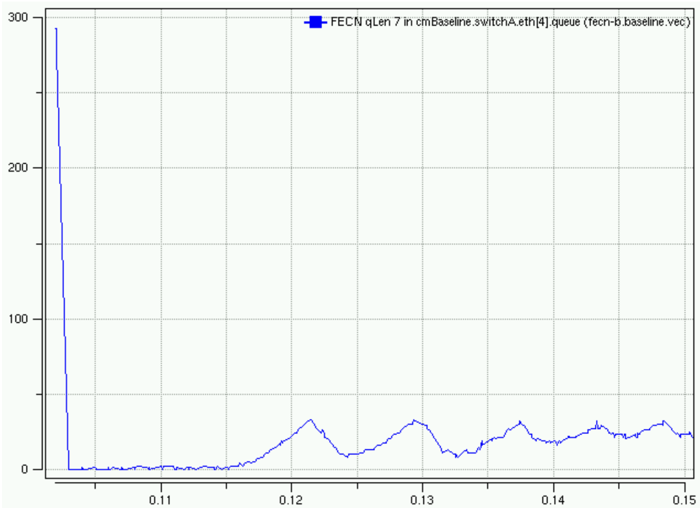


QCN-PR



Baseline: Queue Length

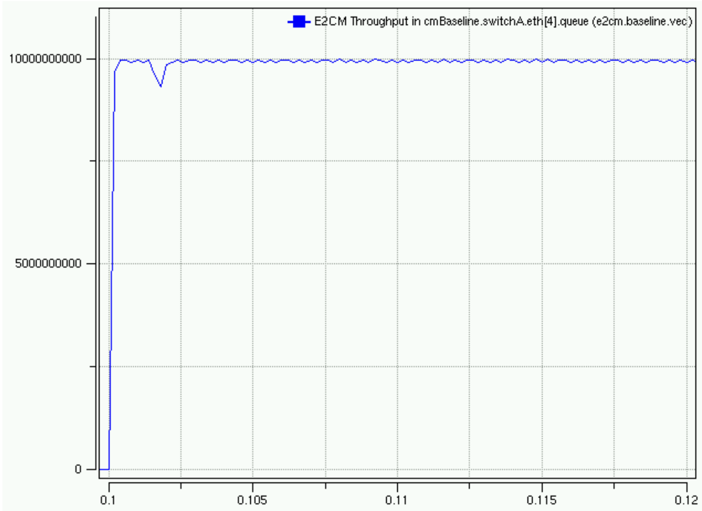
FECN-B



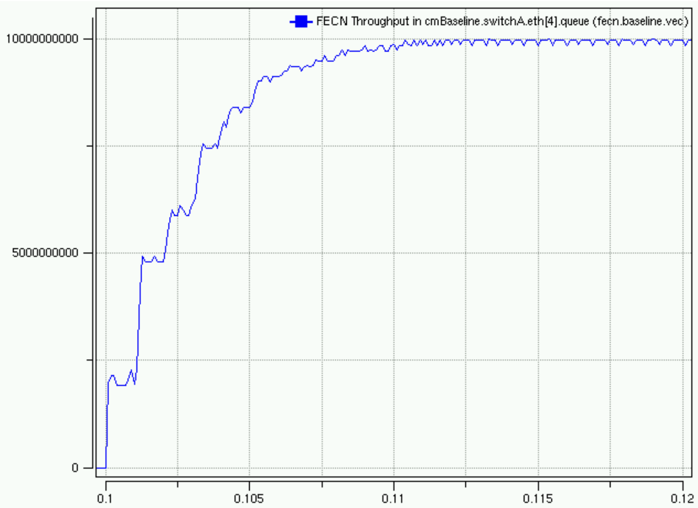
Baseline: Throughput

ECM

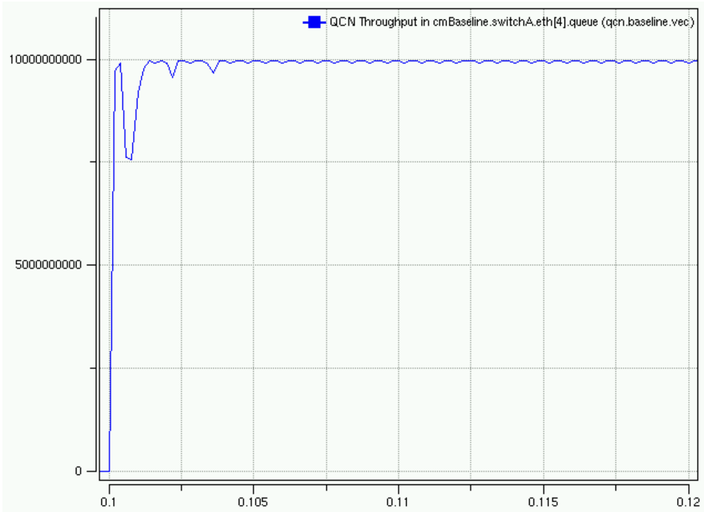
E2CM



FECN

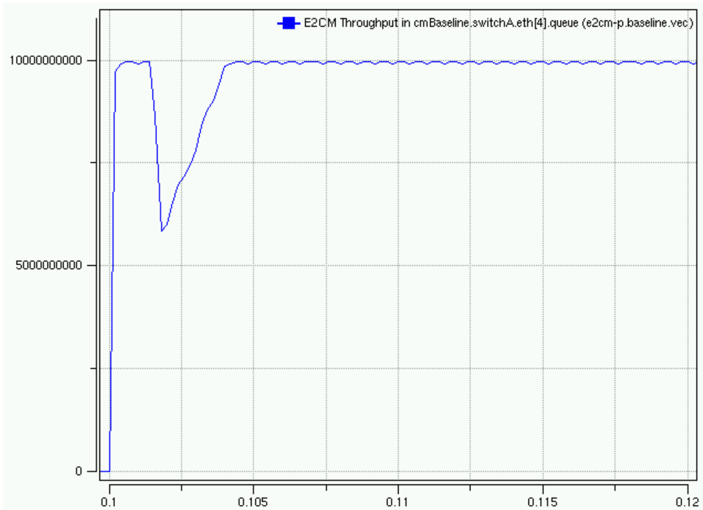


QCN

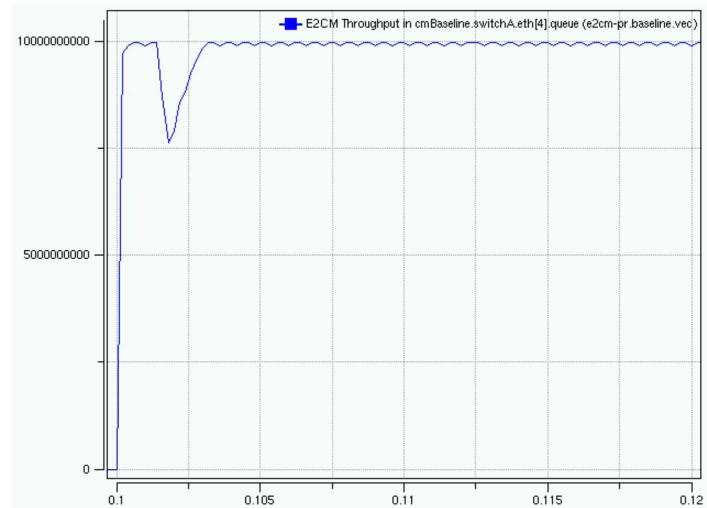


Baseline: Throughput

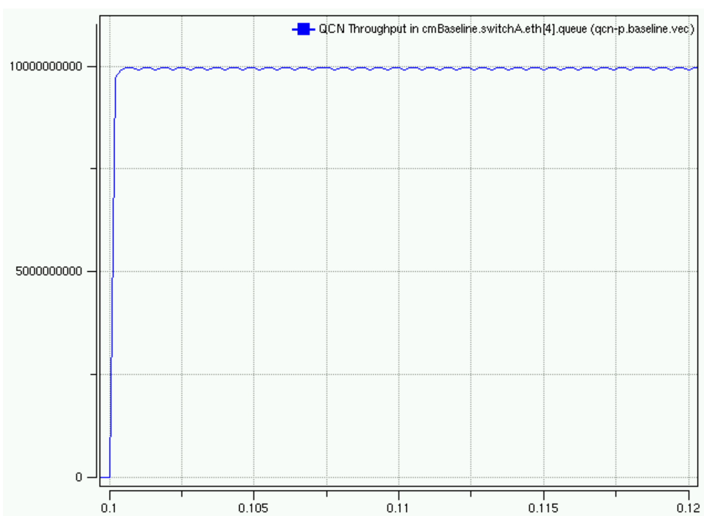
E2CM-P



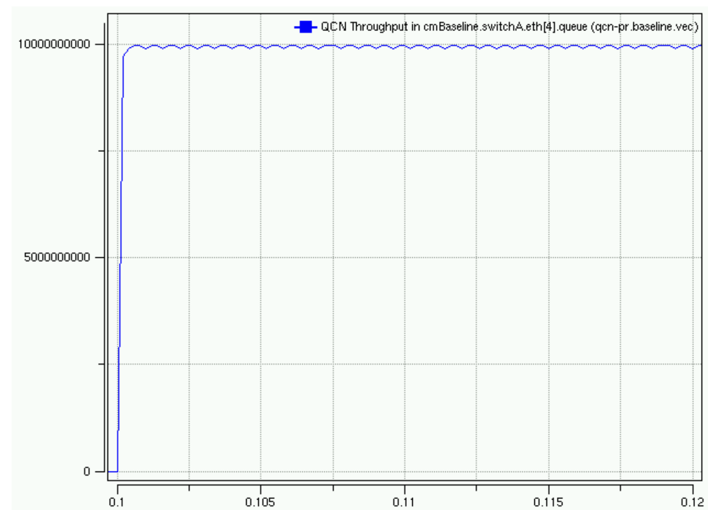
E2CM-PR



QCN-P

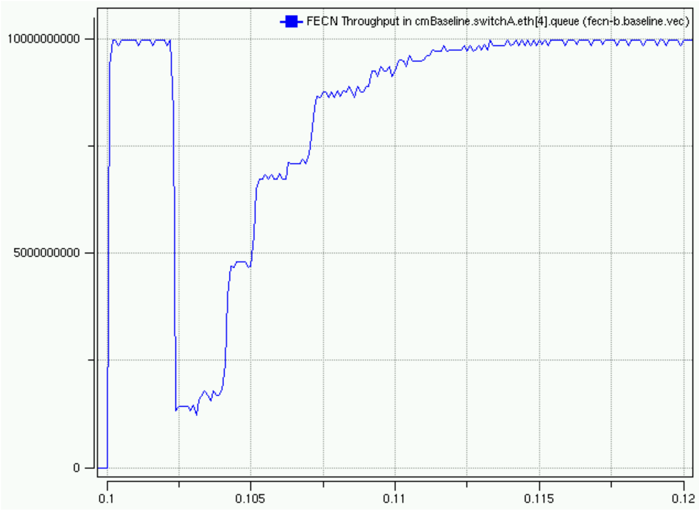


QCN-PR



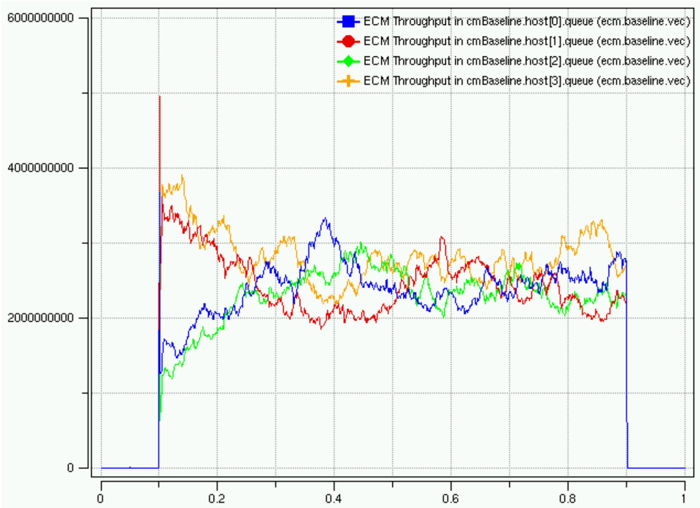
Baseline: Throughput

FECN-B

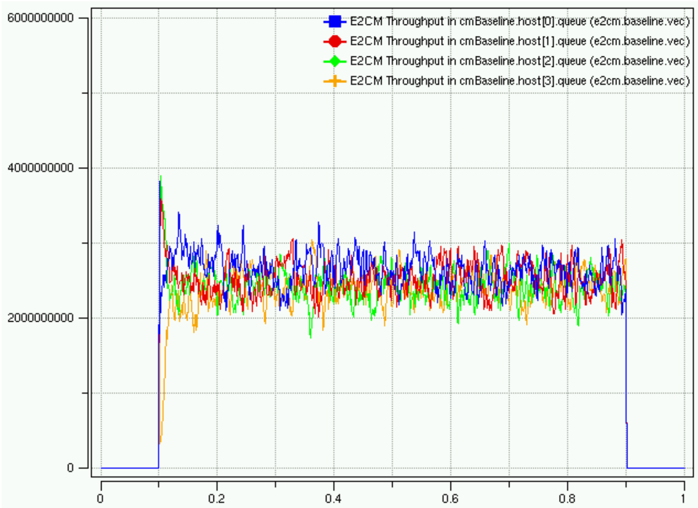


Baseline: Fairness

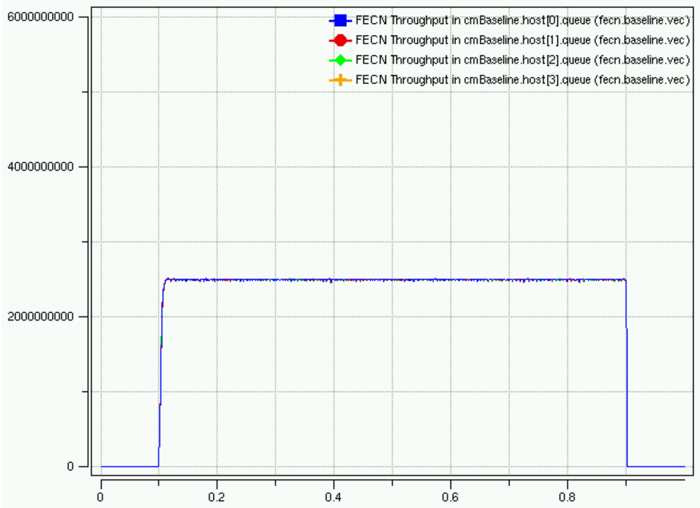
ECM



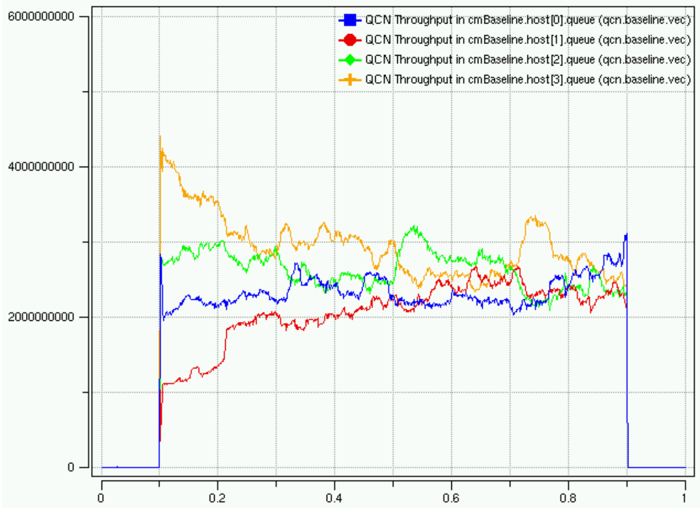
E2CM



FECN

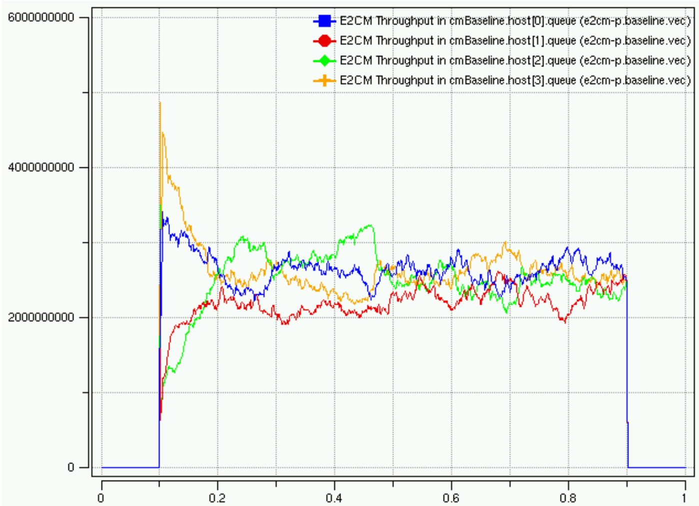


QCN

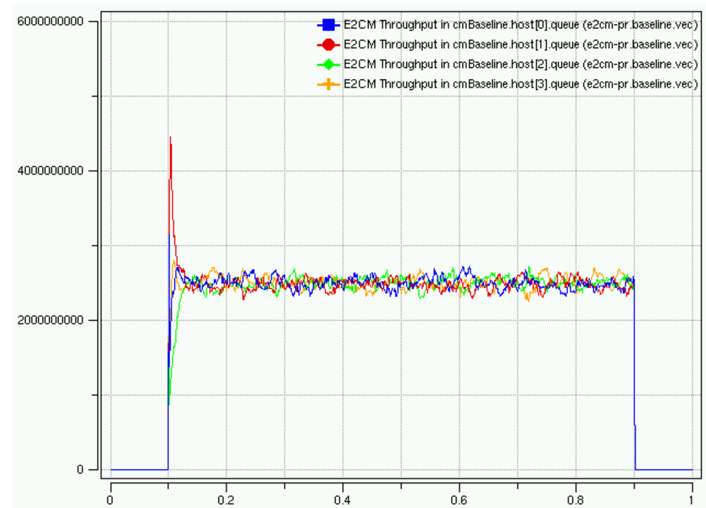


Baseline: Fairness

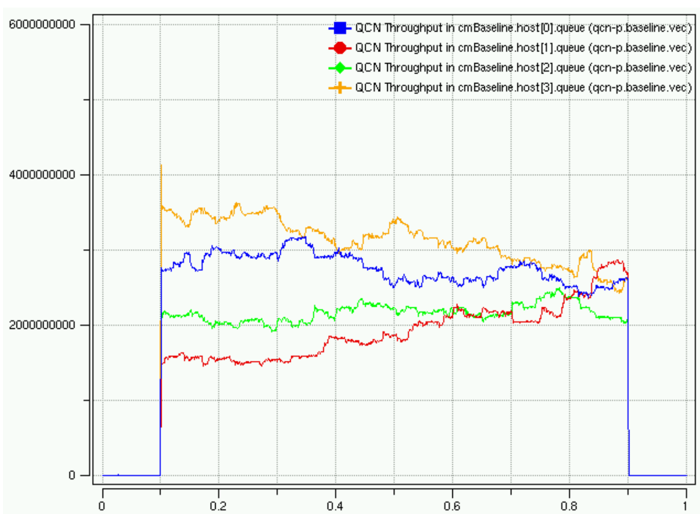
E2CM-P



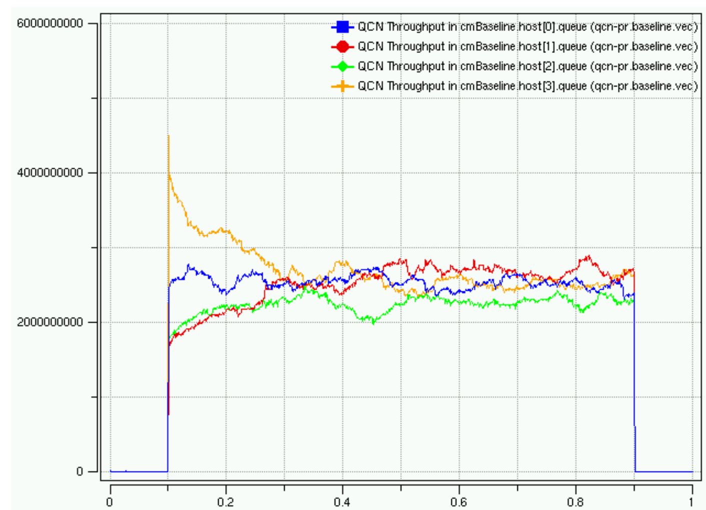
E2CM-PR



QCN-P

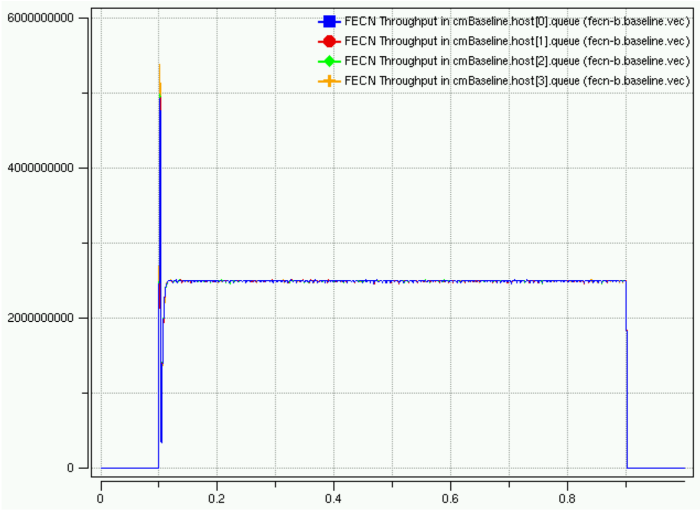


QCN-PR

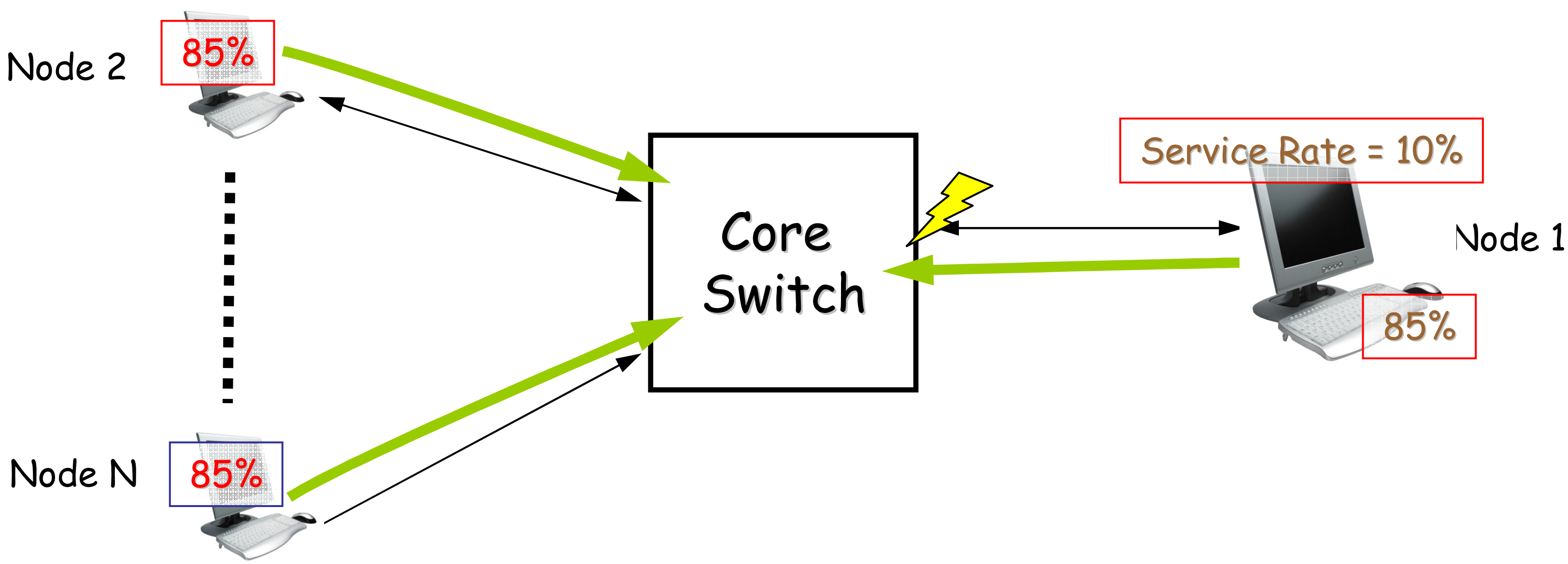


Baseline: Fairness

FECN-B



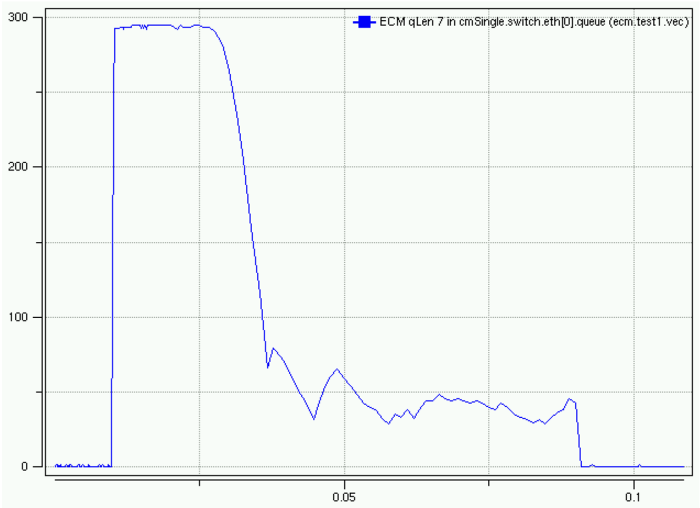
Test 1: Output Generated Single Hotspot



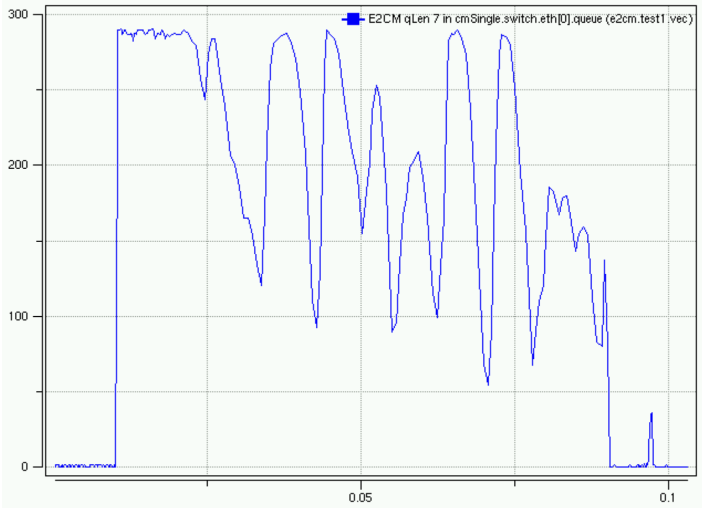
- All nodes (10): Uniform distribution, load: 8.5 Gb/s
- Node 1 (hotspot) service rate: 1Gb/s
  - Node1 limits service rates by sending PAUSE frames to switch
- One congestion point
  - Duration: 80mS from ti=10ms to 90 ms

Test 1: Queue Length

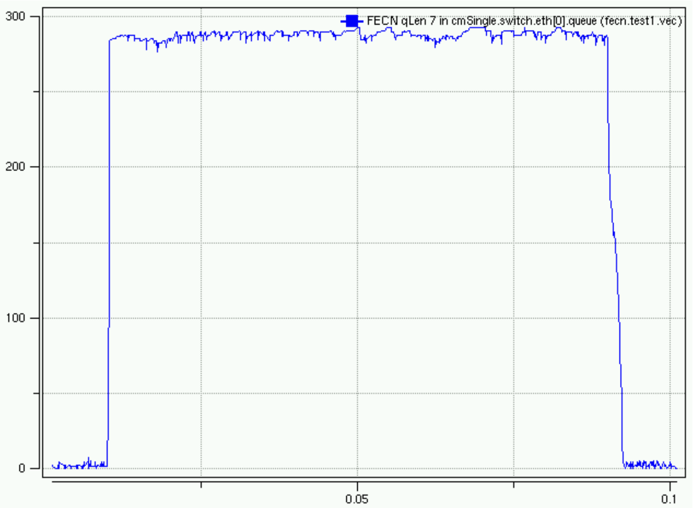
ECM



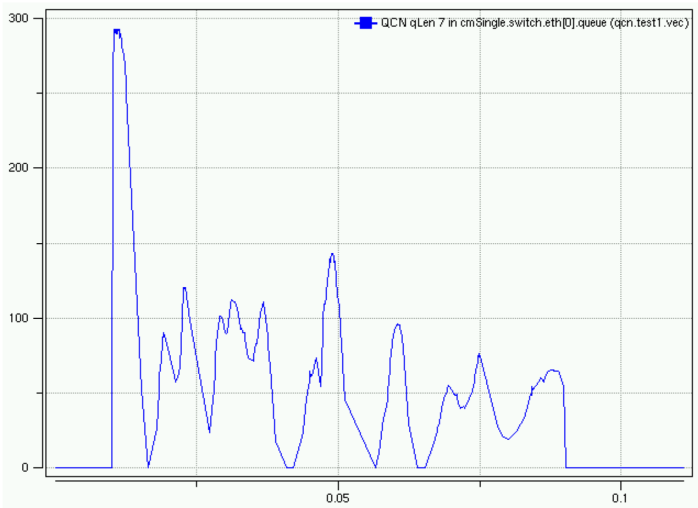
E2CM



FECN

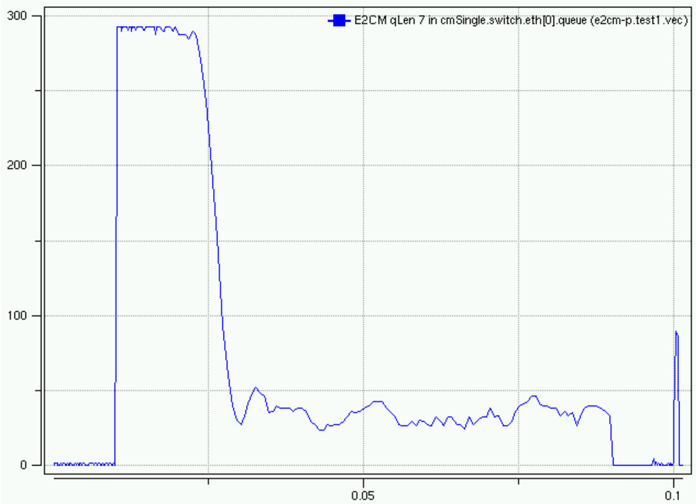


QCN

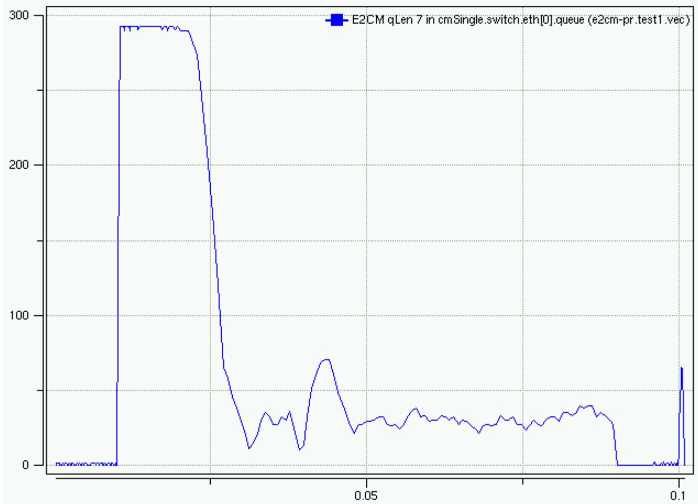


Test 1: Queue Length

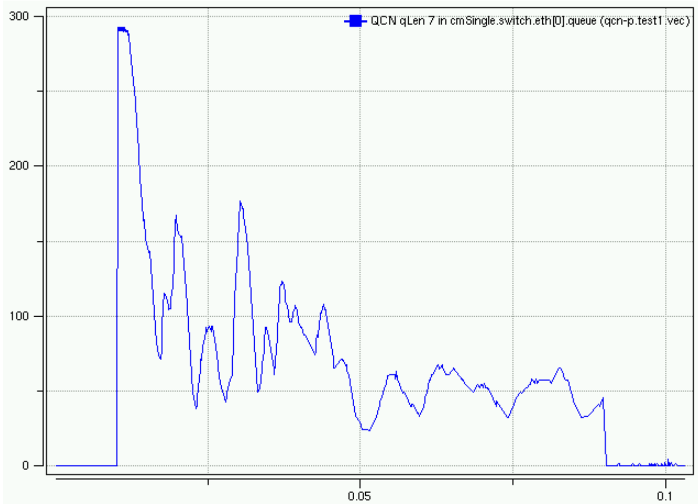
E2CM-P



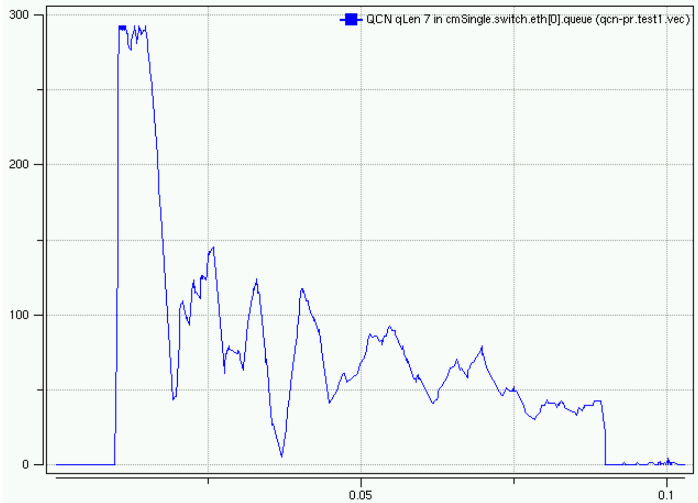
E2CM-PR



QCN-P

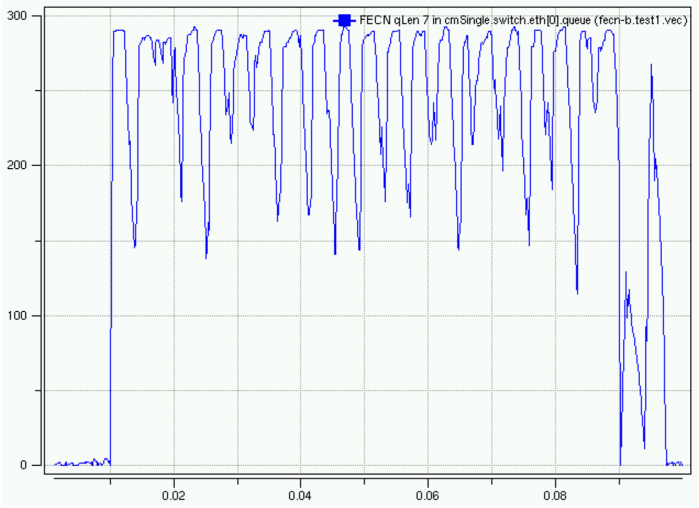


QCN-PR



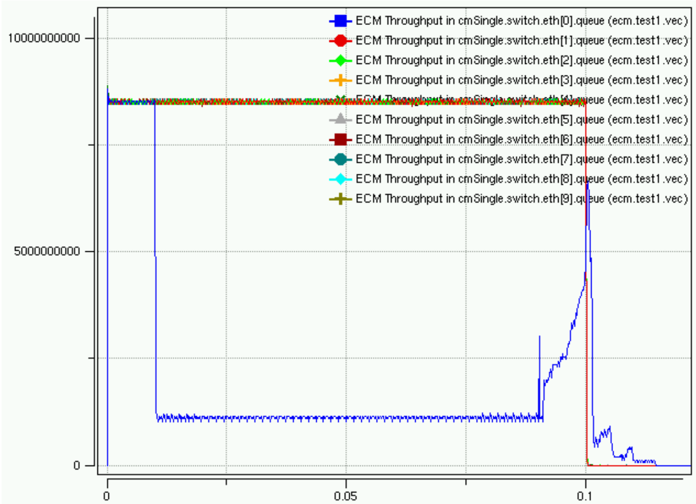
Test 1: Queue Length

FECN-B

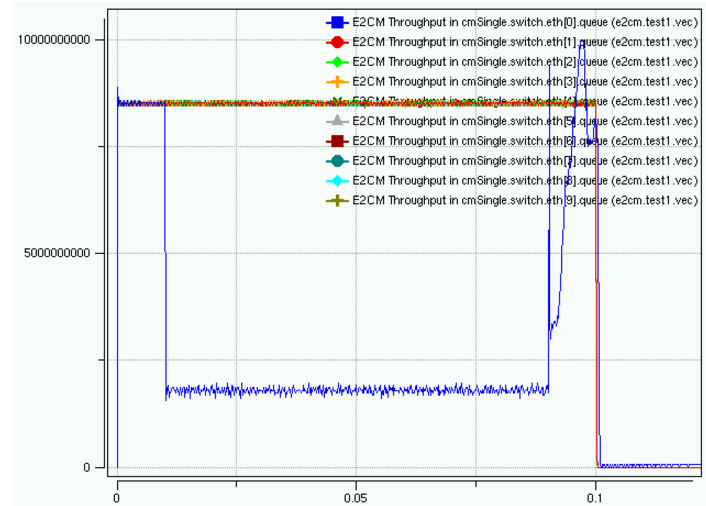


Test 1: Throughput

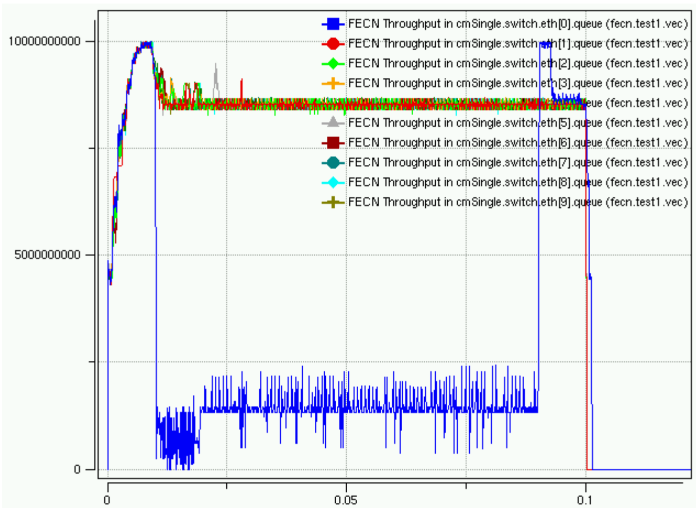
ECM



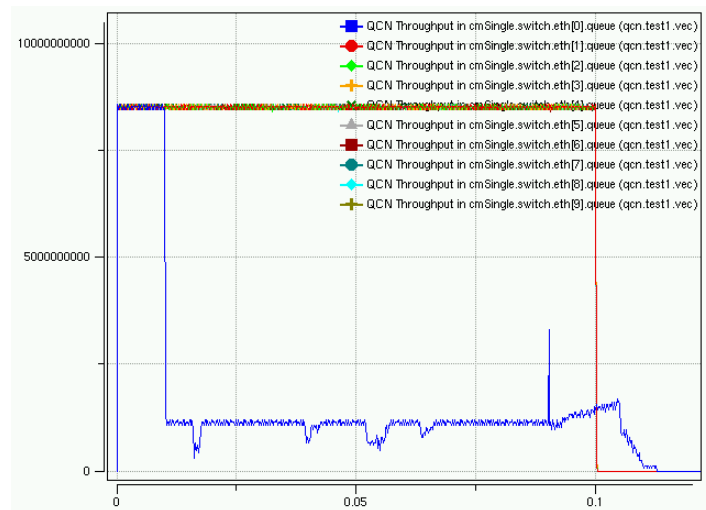
E2CM



FECN

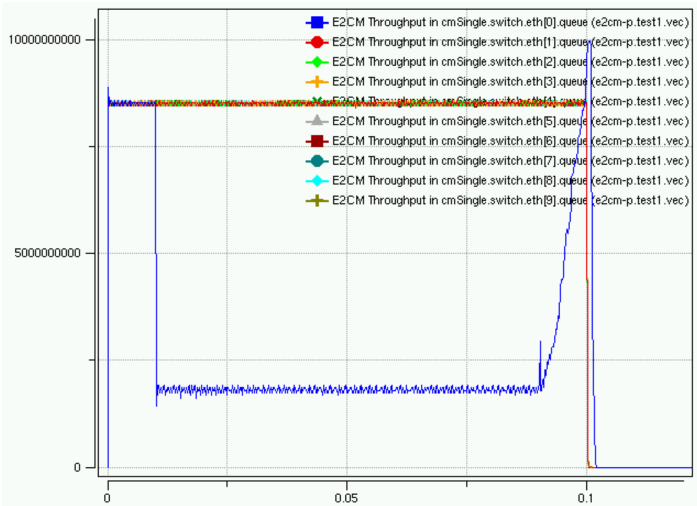


QCN

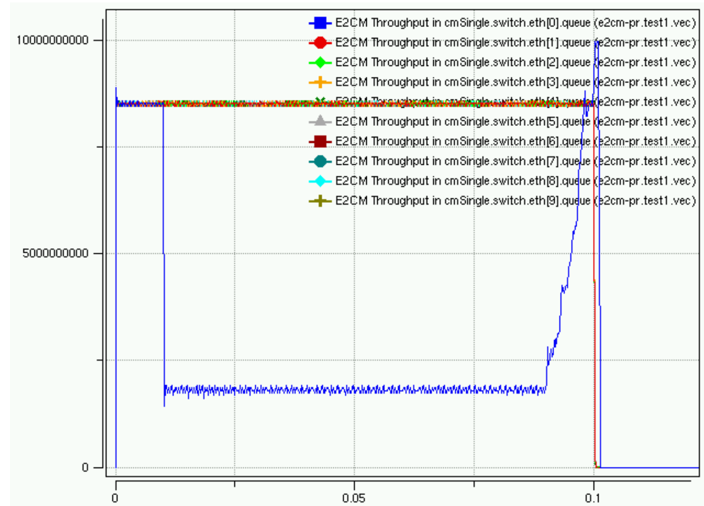


Test 1: Throughput

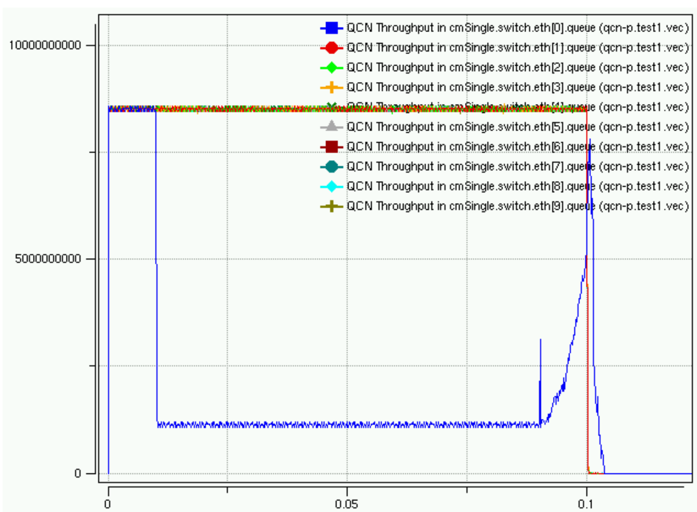
E2CM-P



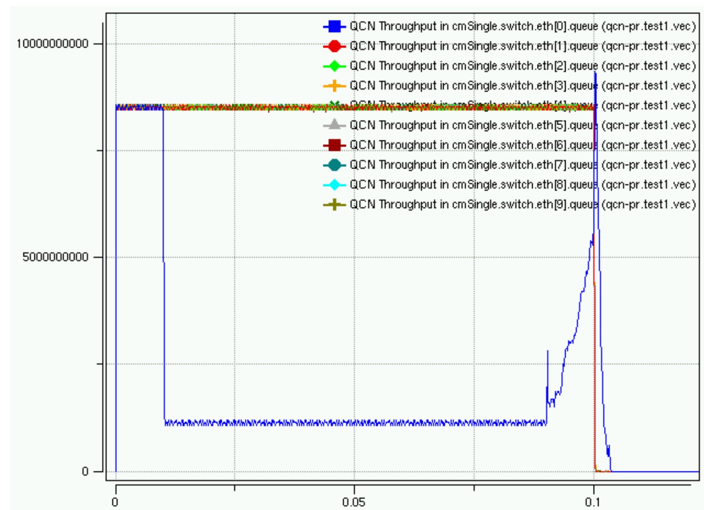
E2CM-PR



QCN-P



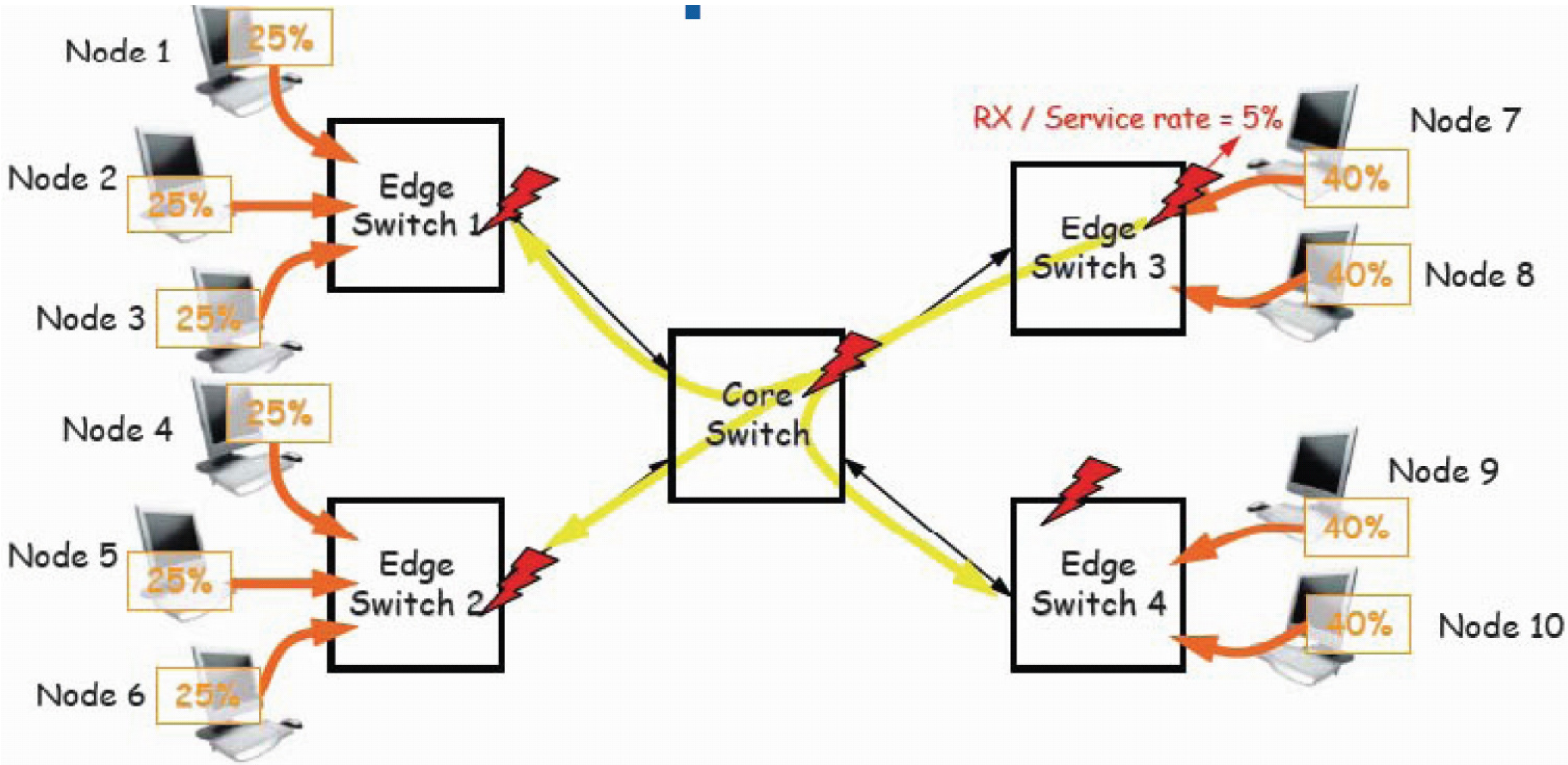
QCN-PR



Test 1: Throughput

FECN-B

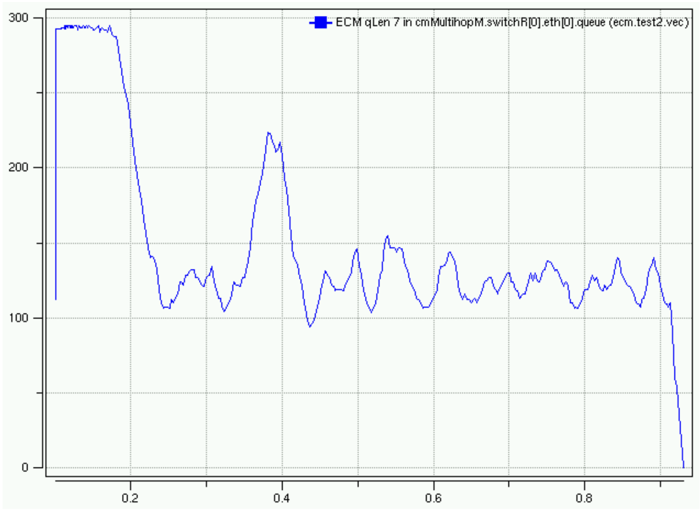
Test 2: Output-Generated Hotspot, Multi-Hop



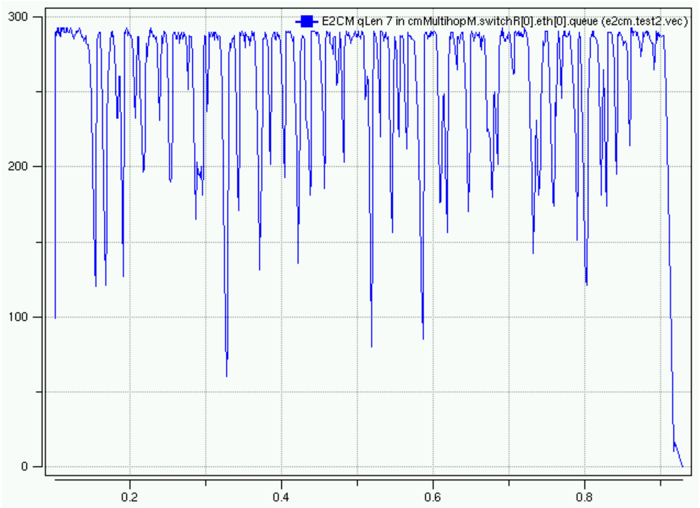
- All: Uniform distribution traffic (background traffic)
- Nodes 1-6: 25% (2.5 Gbps), Nodes 7-10: 40% (4 Gbps)
- Primary Hotspot: Node 7 service rate = 5% (Rx only)

Test 2: Queue Length

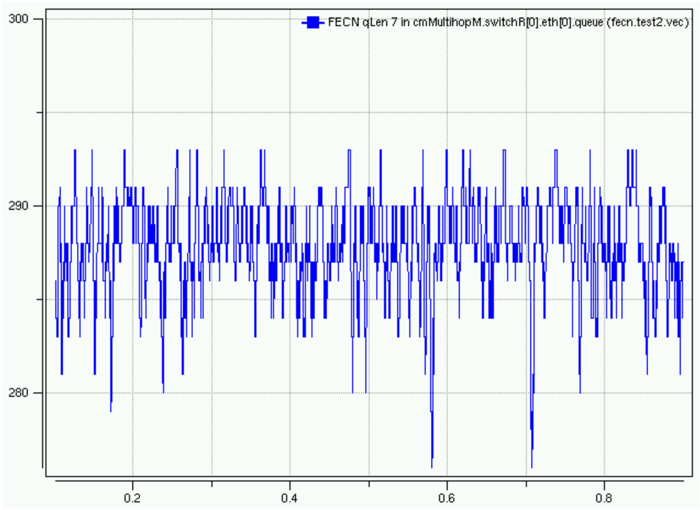
ECM



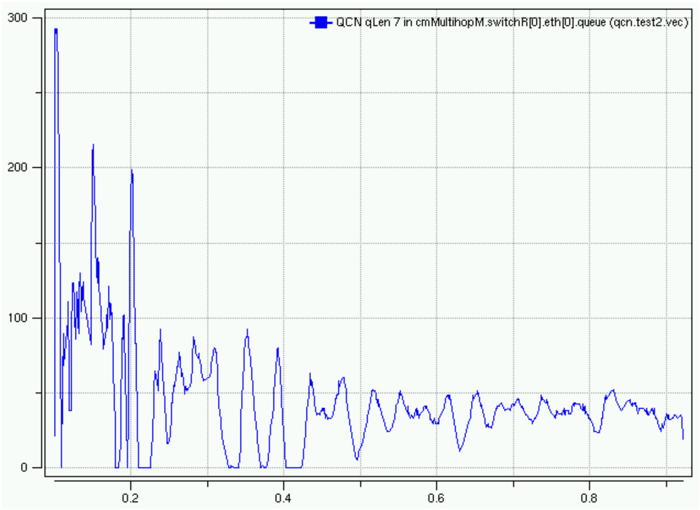
E2CM



FECN

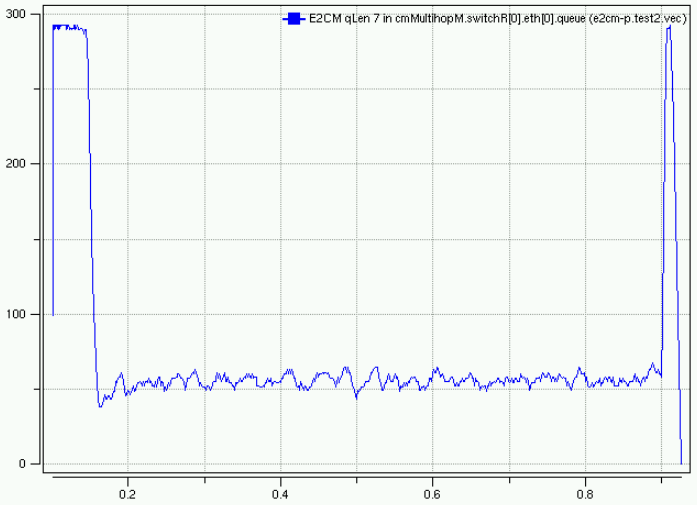


QCN

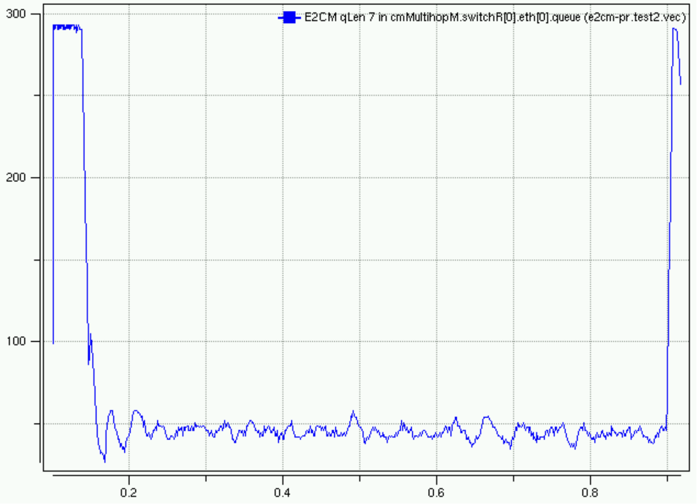


Test 2: Queue Length

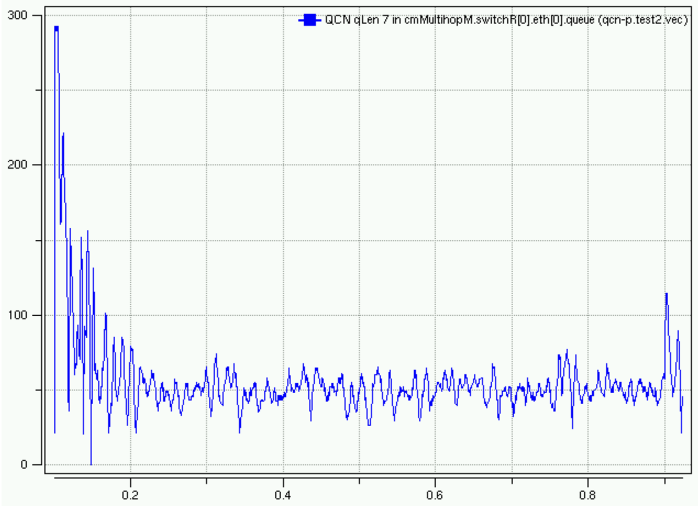
E2CM-P



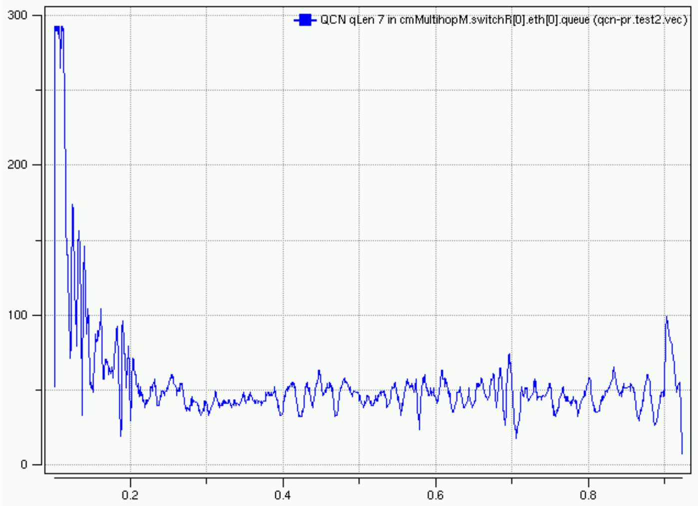
E2CM-PR



QCN-P

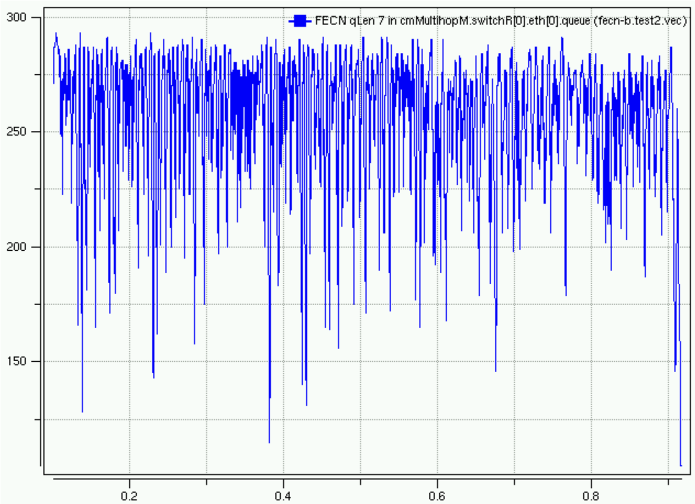


QCN-PR



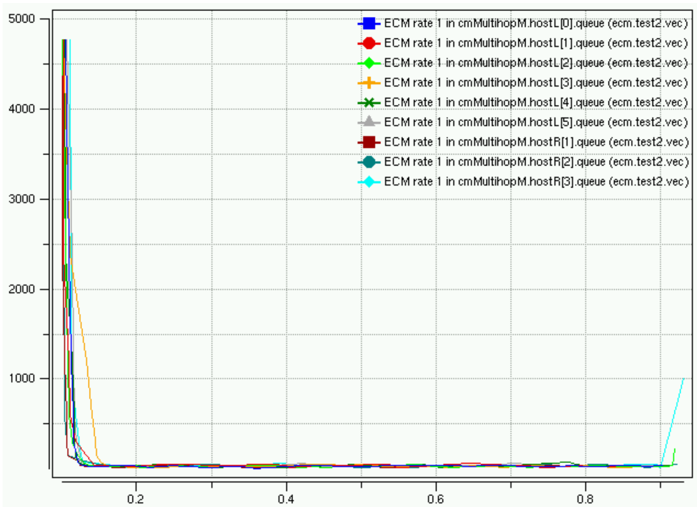
Test 2: Queue Length

FECN-B

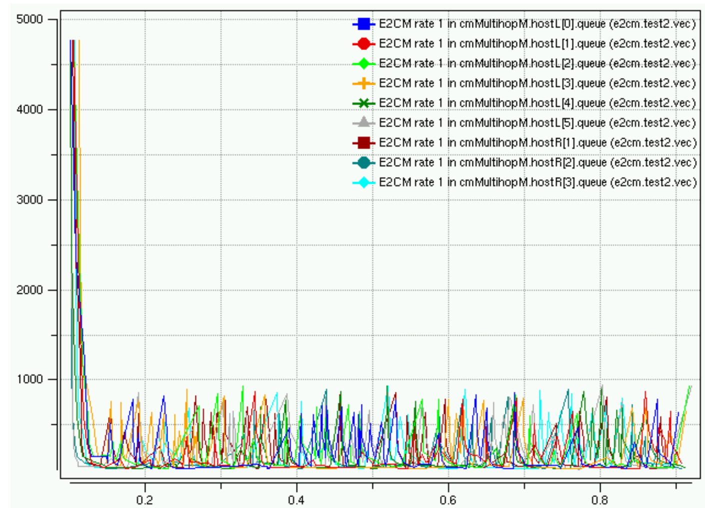


Test 2: Data Rate to Node 7

ECM

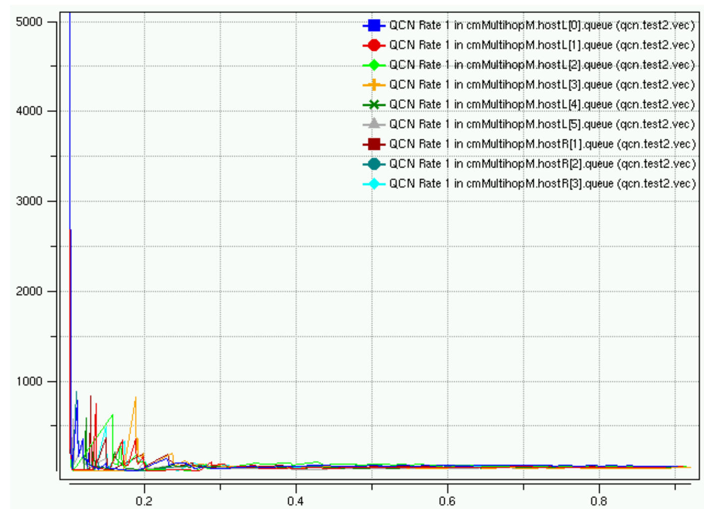


E2CM



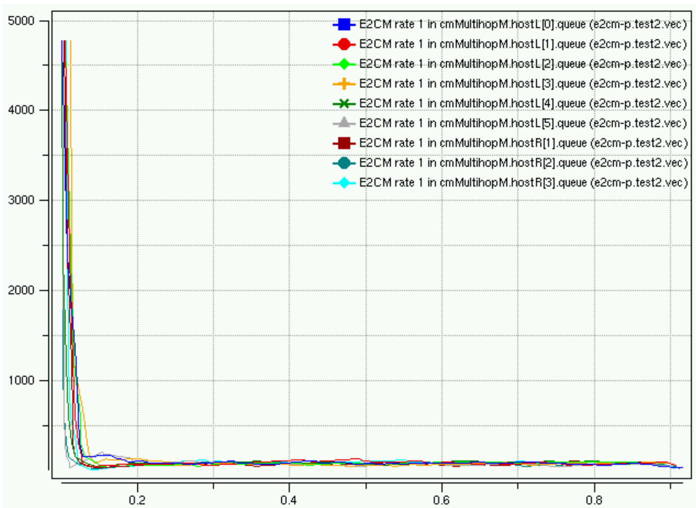
FECN

QCN

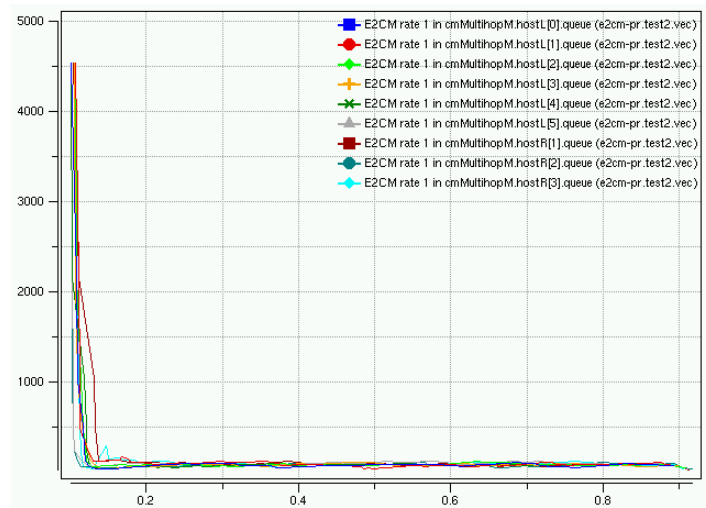


Test 2: Data Rate to Node 7

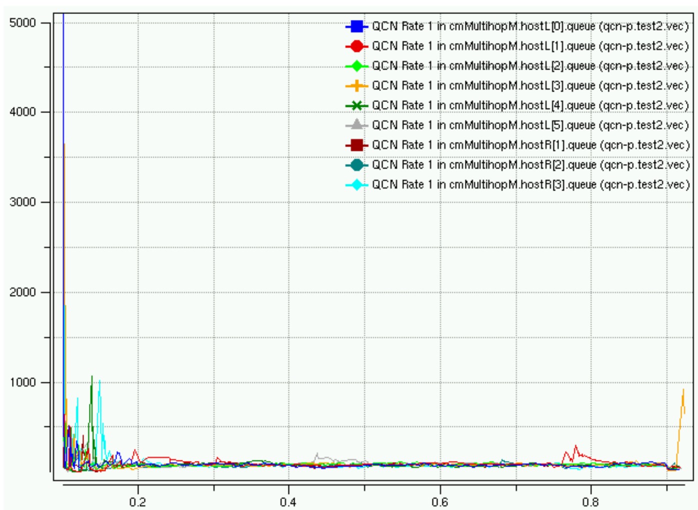
E2CM-P



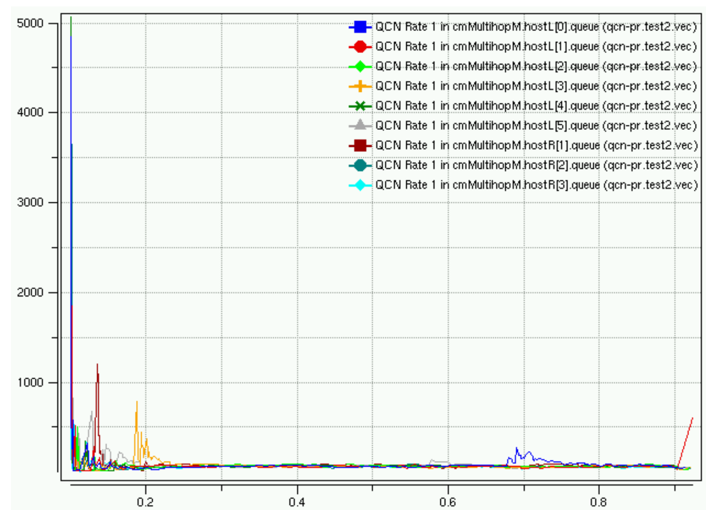
E2CM-PR



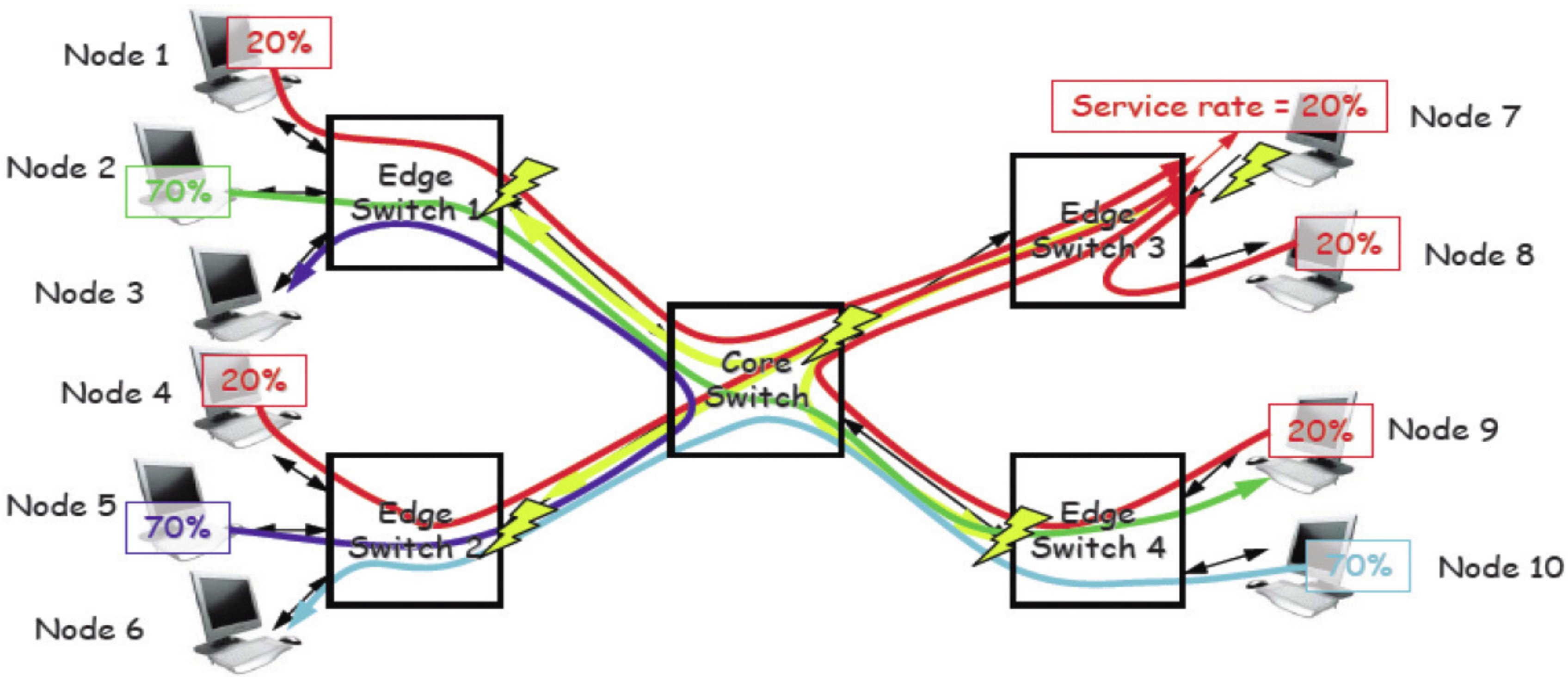
QCN-P



QCN-PR



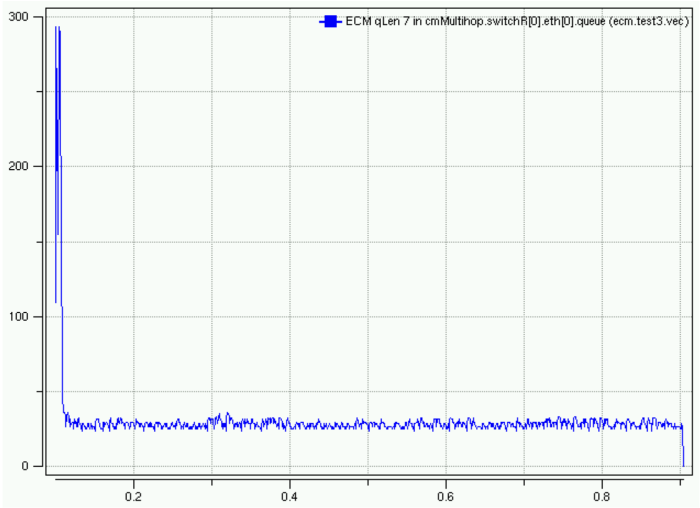
Test 3: Output-generated Hotspot; Multi-hop, Selected Victims



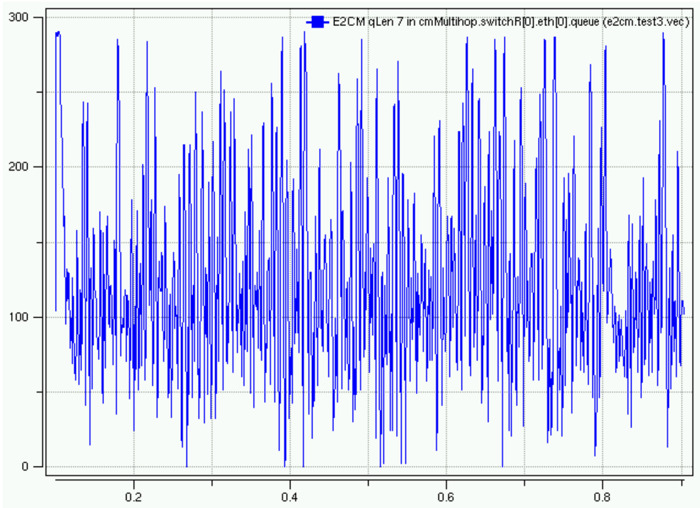
- Four culprit flows of 1Gb/s each from nodes 1,4,8,9 to node 7 (hotspot)
- Three victim flows of 7 Gb/s each: node 2 to 9, 5 to 3, and 10 to 6
- Node 7 service rate: 20%
- Five congestion points; all switches and all flows are affected
- Fair allocation provides 0.5 Gb/s to all culprits and 7 Gb/s to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

Test 3: Queue Length

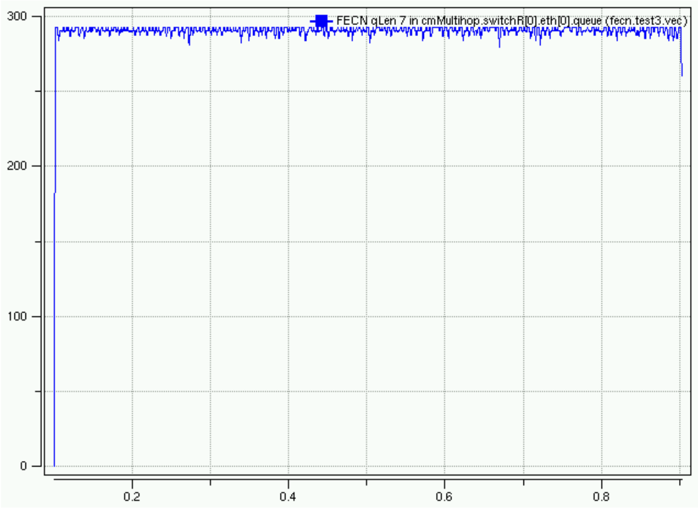
ECM



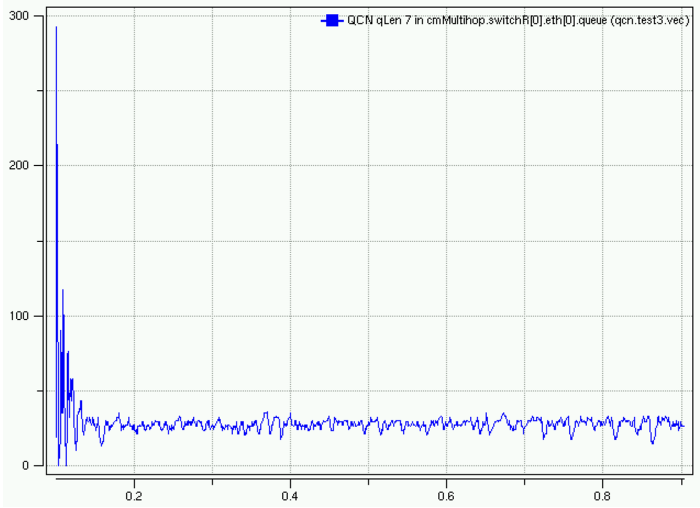
E2CM



FECN

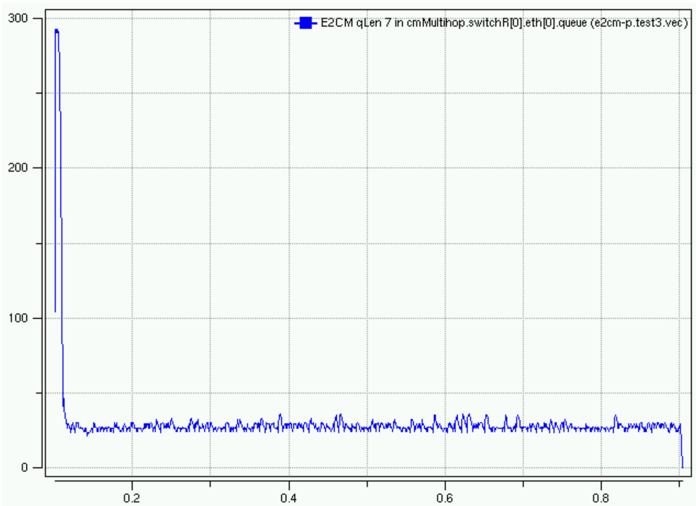


QCN

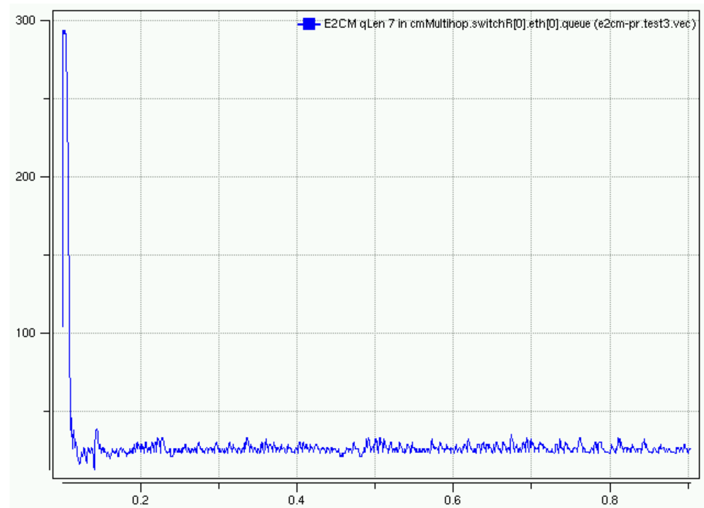


Test 3: Queue Length

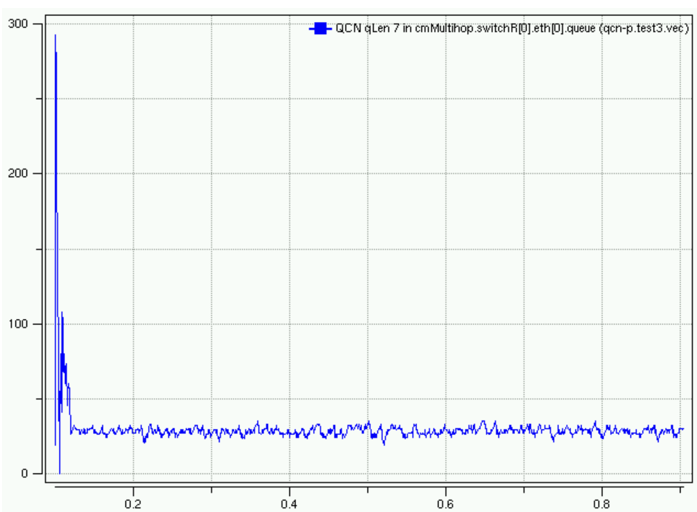
E2CM-P



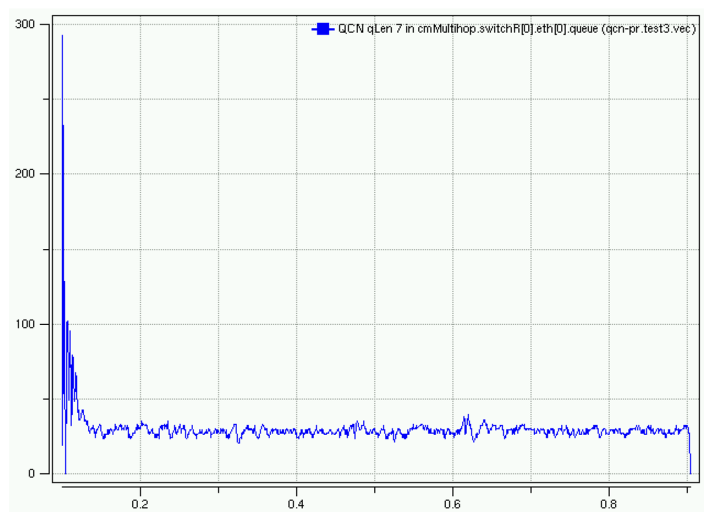
E2CM-PR



QCN-P

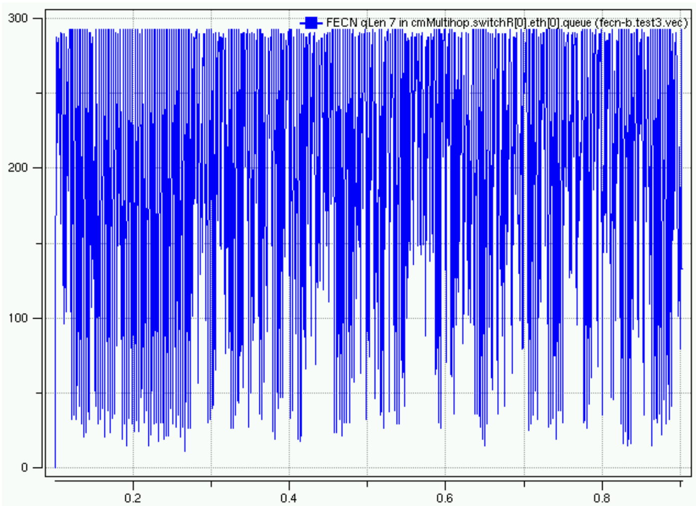


QCN-PR

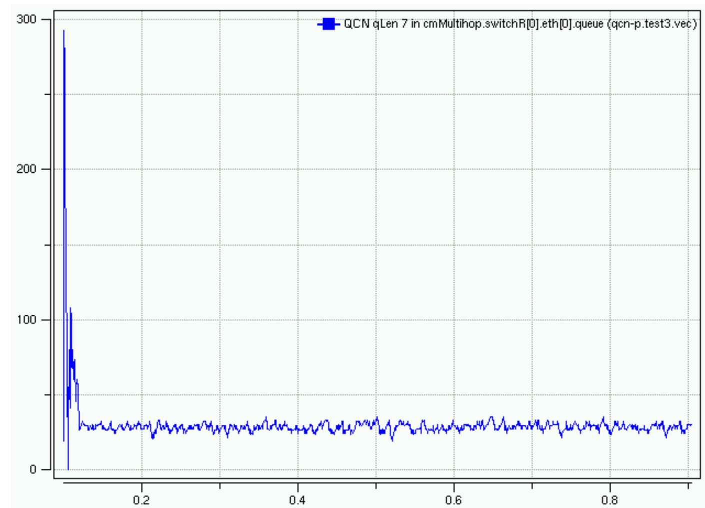


Test 3: Queue Length

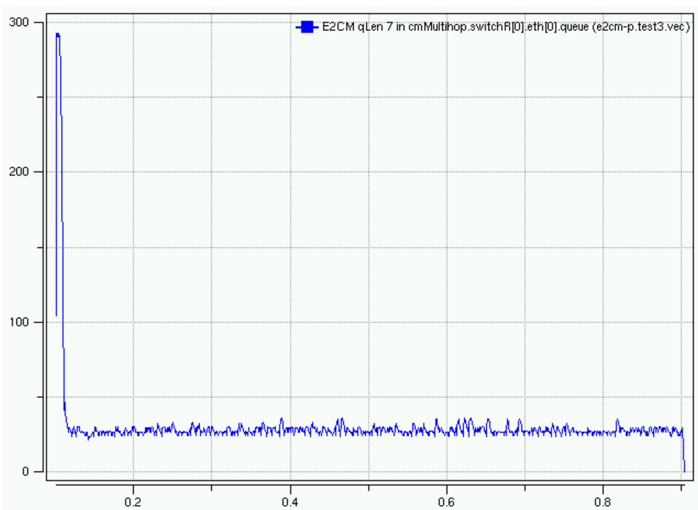
FECN-B



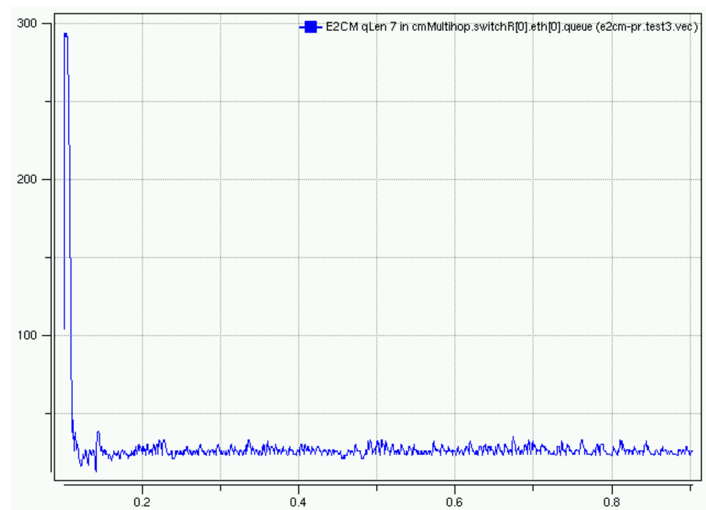
QCN-P



E2CM-P

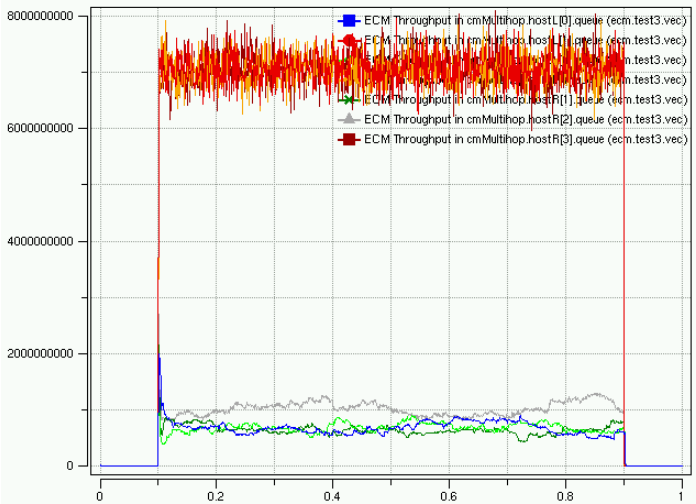


E2CM-PR

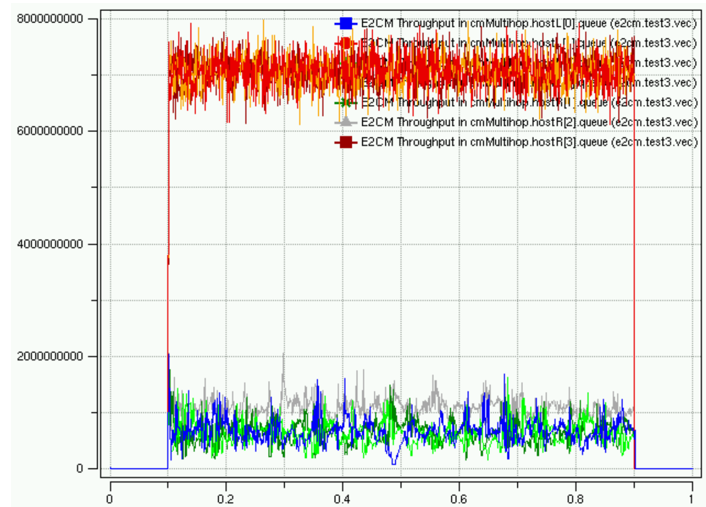


Test 3: Fairness

ECM

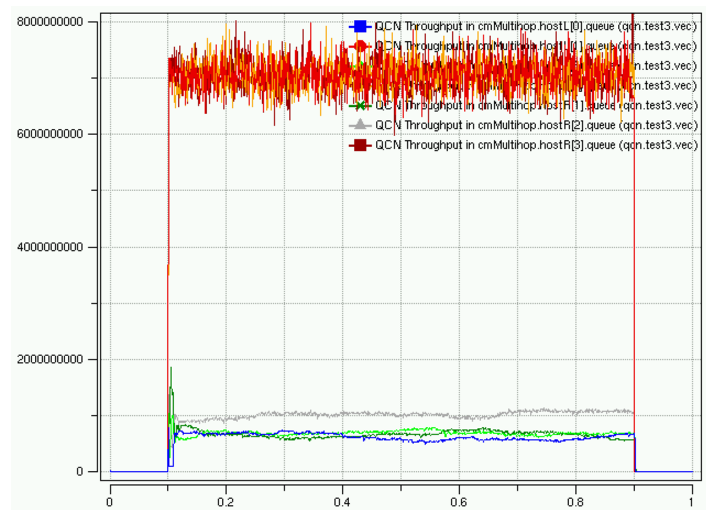


E2CM



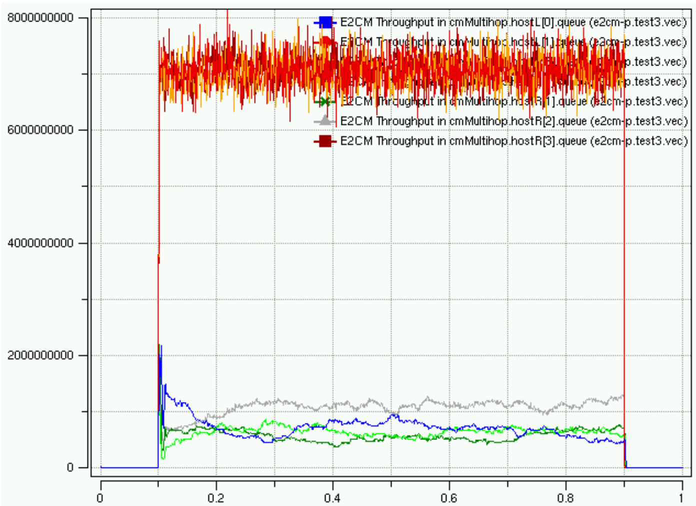
FECN

QCN

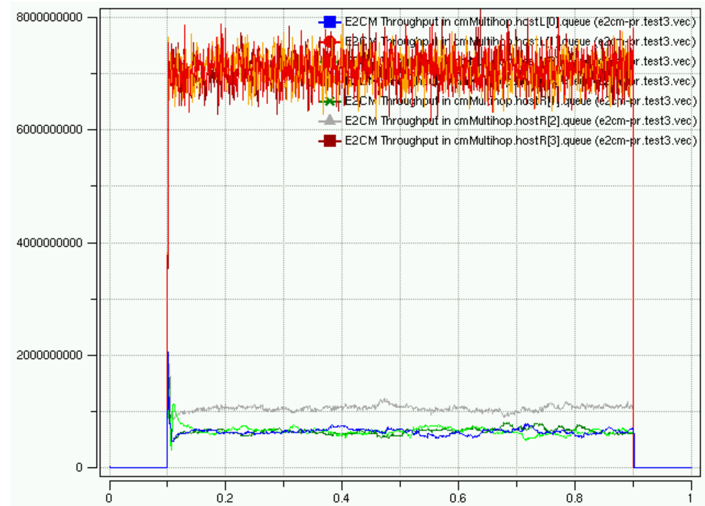


Test 3: Fairness

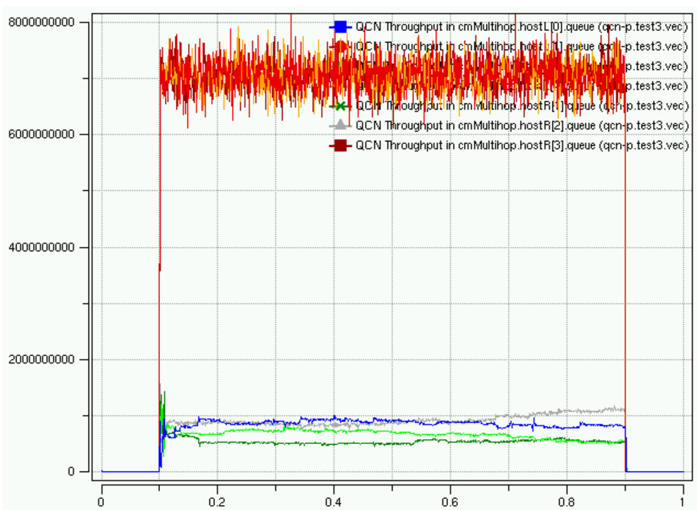
E2CM-P



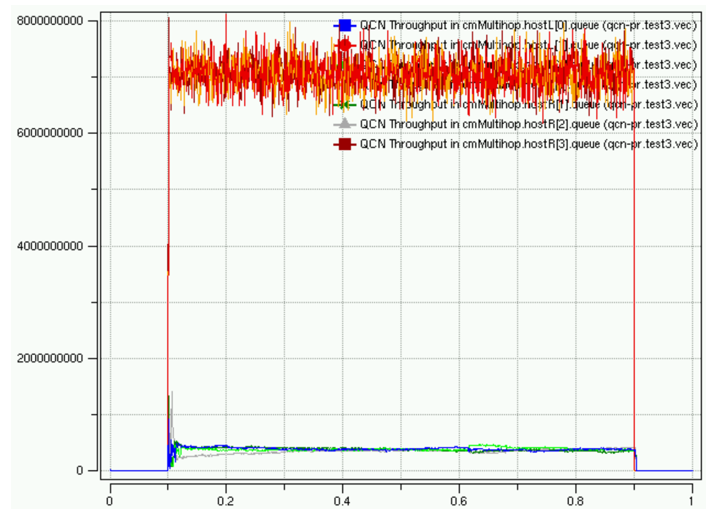
E2CM-PR



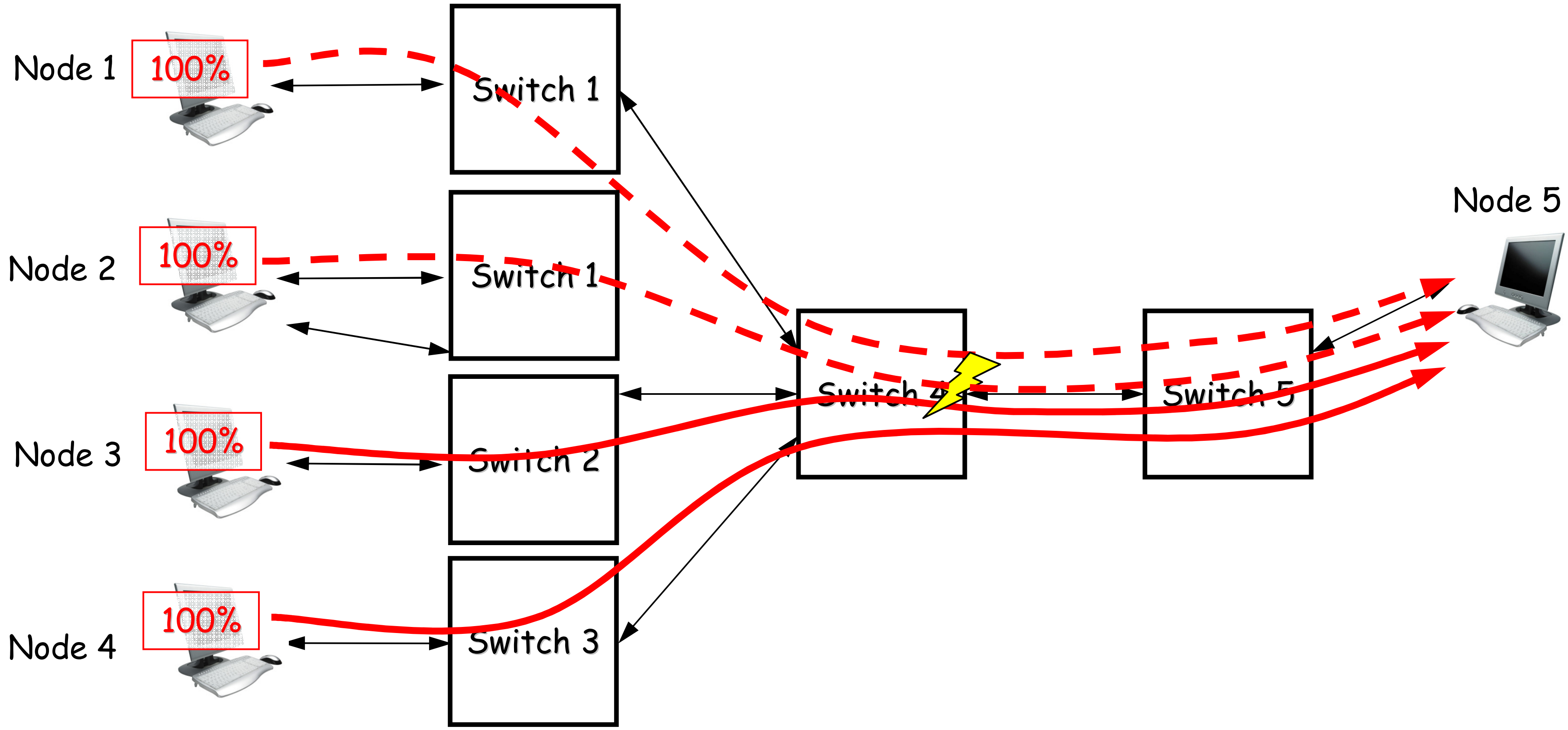
QCN-P



QCN-PR



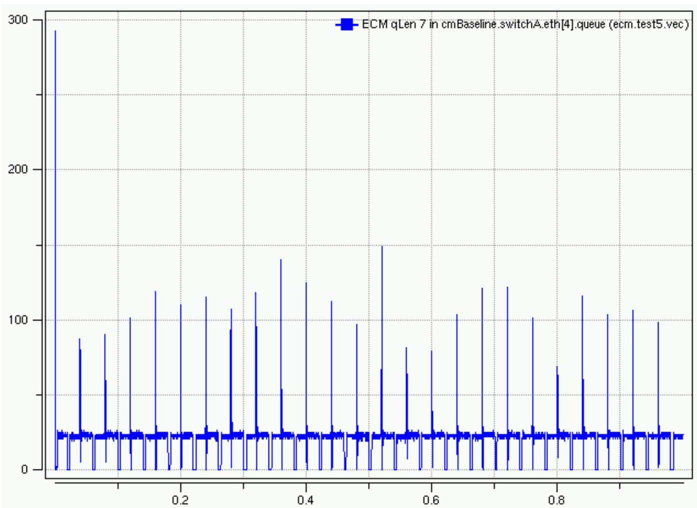
Test 5: Symmetric Topology, Single Hotspot, Bursty



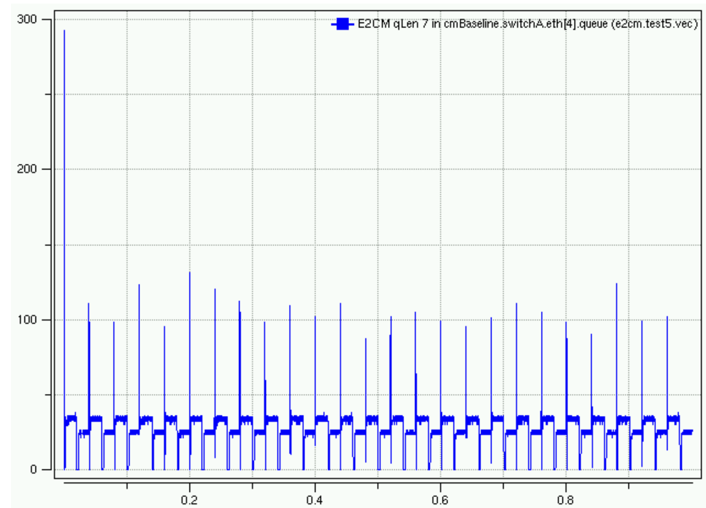
- Point-to-point from node 1..4 to node 5
- Load: 100%
- Node 1 and 2 On/Off Sources (Ton=Toff=20ms)

Test 5: Queue Length

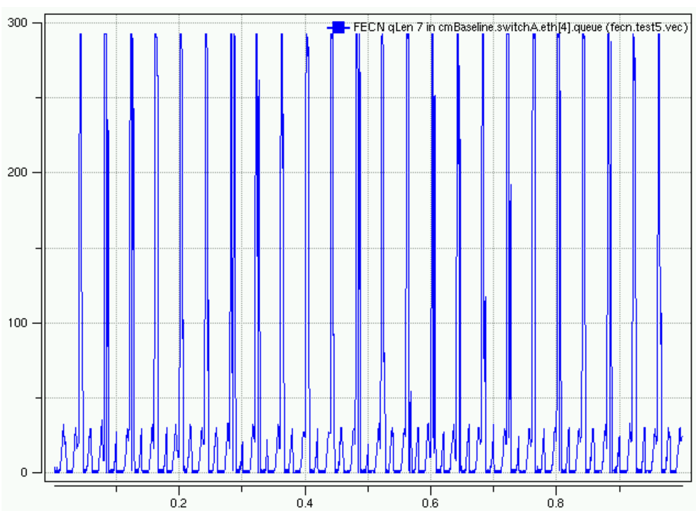
ECM



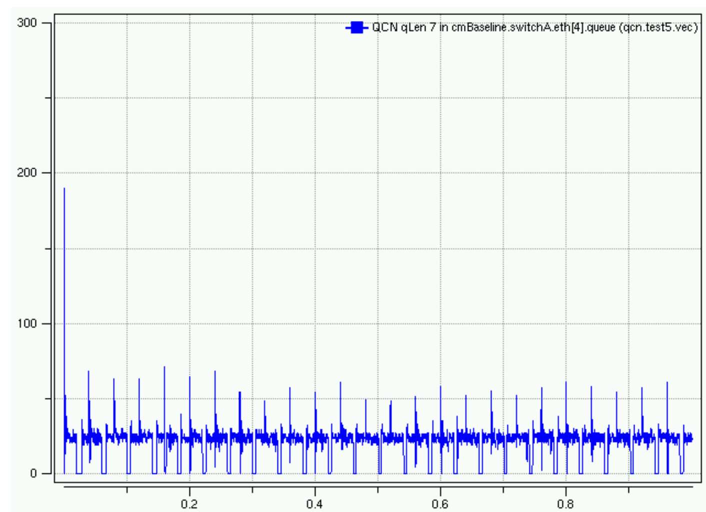
E2CM



FECN

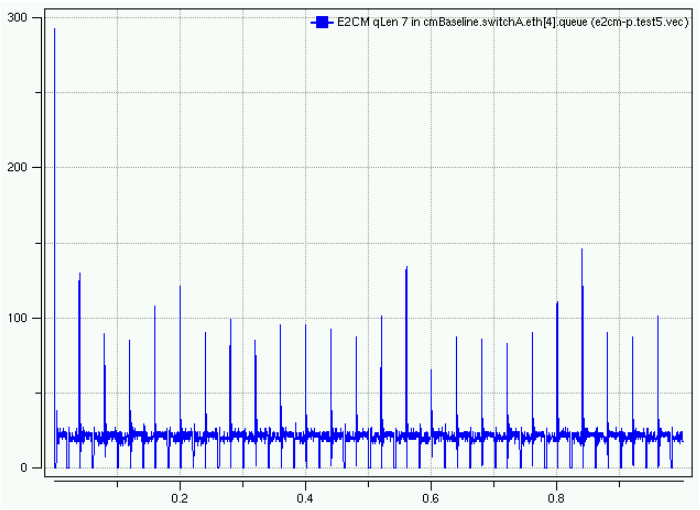


QCN

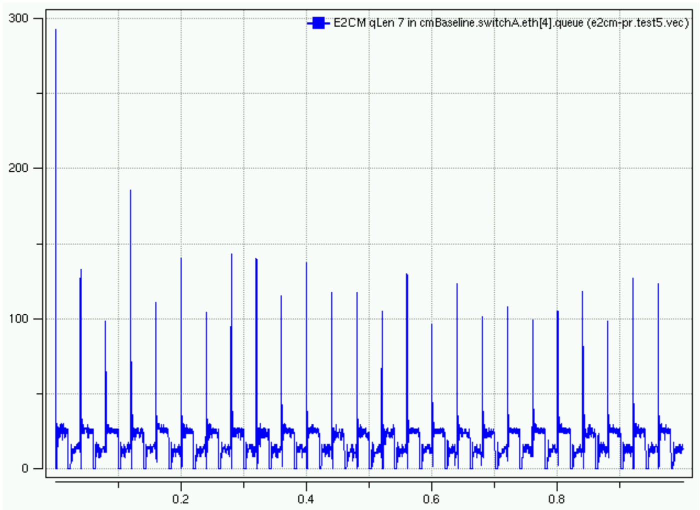


Test 5: Queue Length

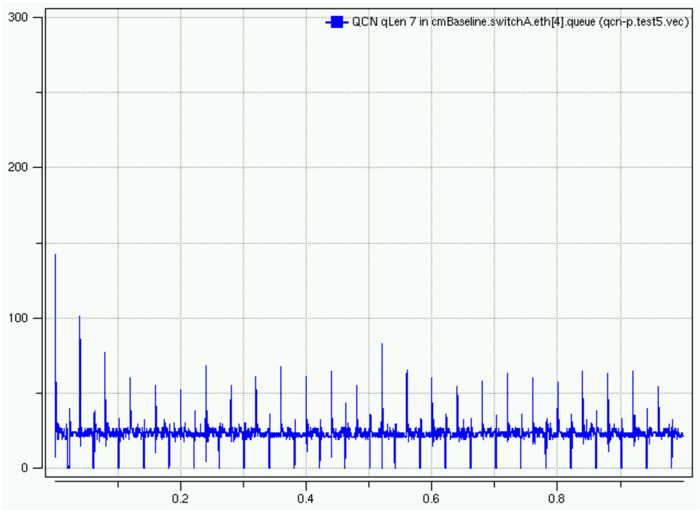
E2CM-P



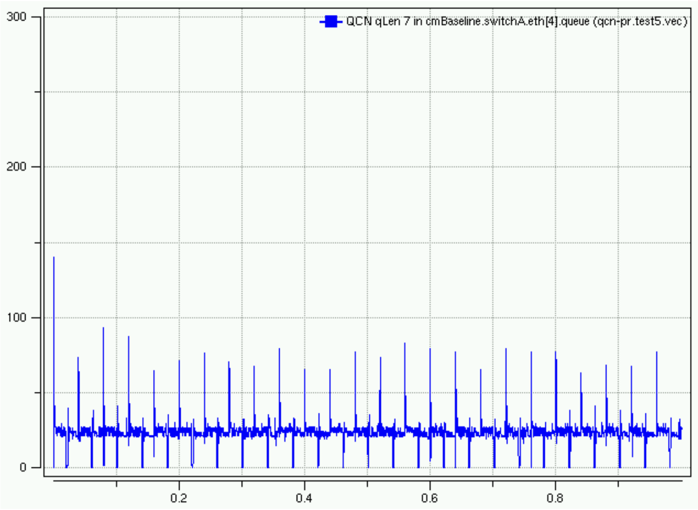
E2CM-PR



QCN-P

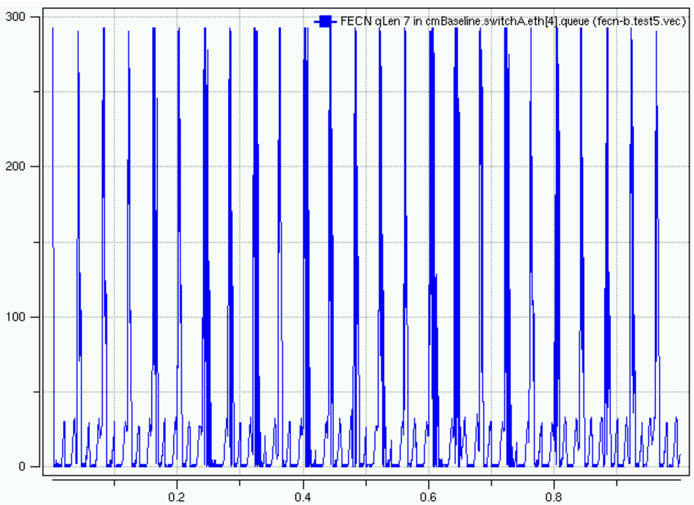


QCN-PR



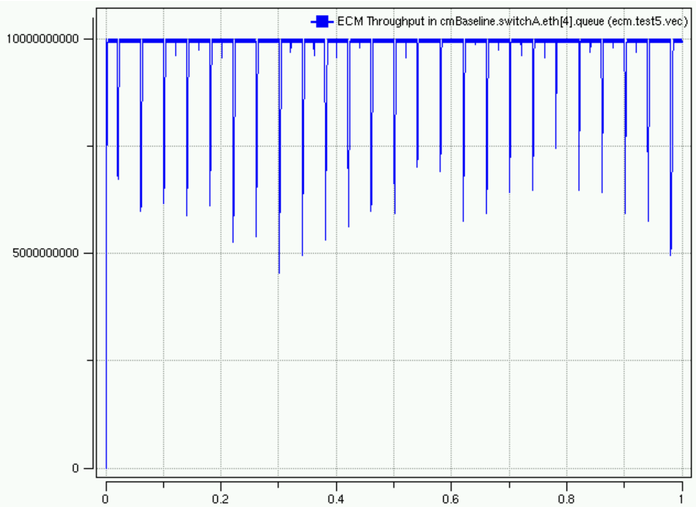
Test 5: Queue Length

FECN-B

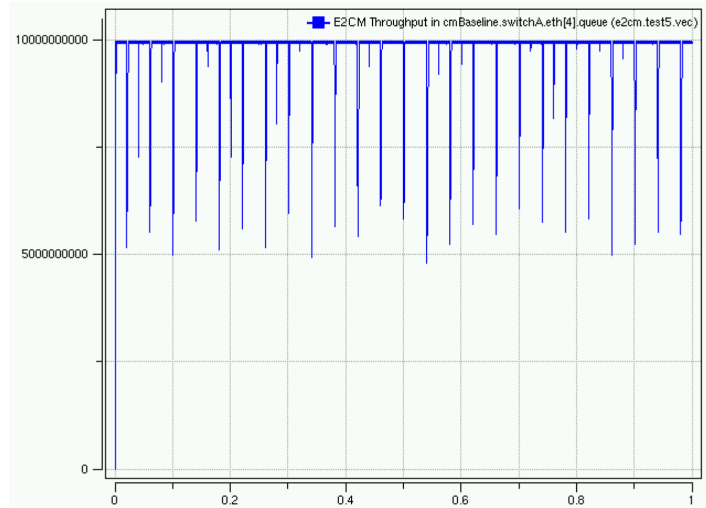


Test 5: Throughput

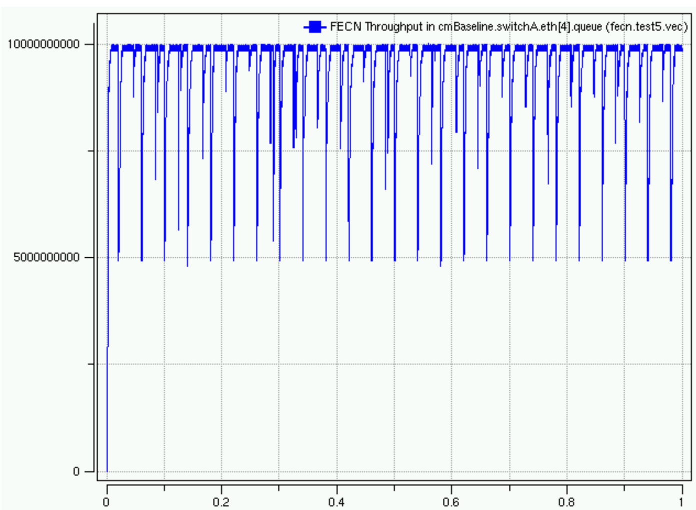
ECM



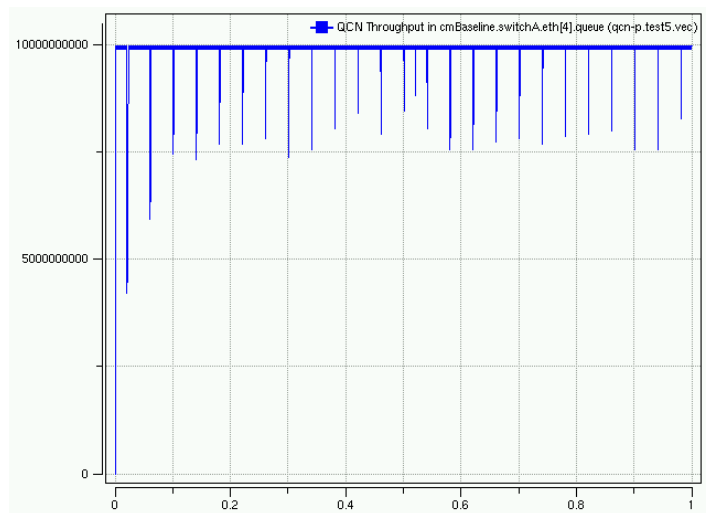
E2CM



FECN

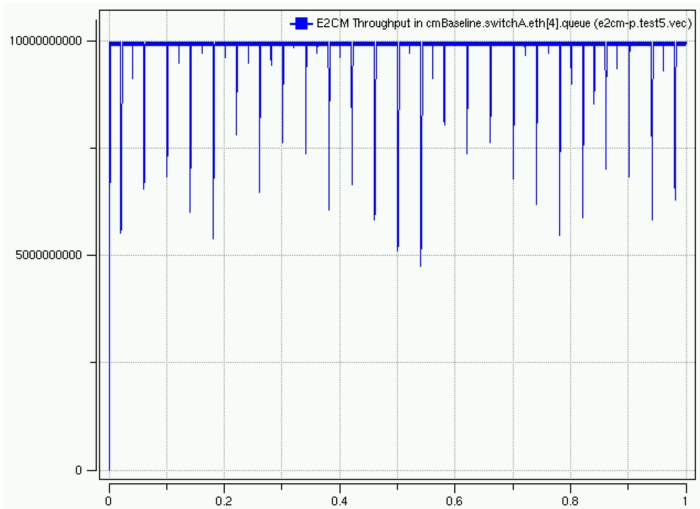


QCN

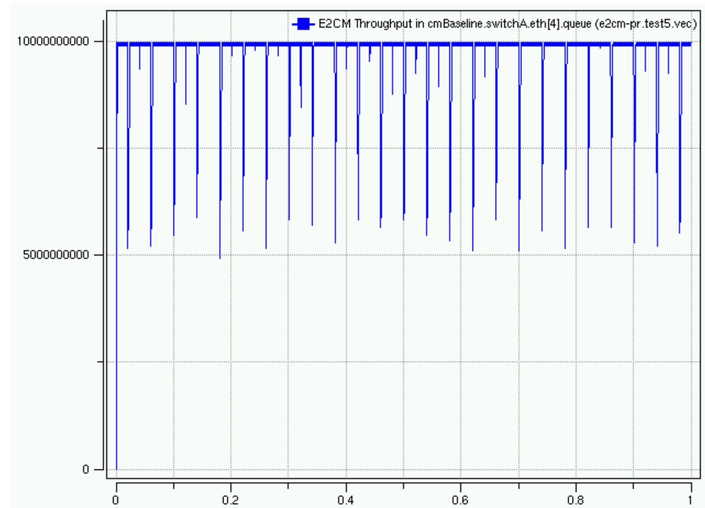


Test 5: Throughput

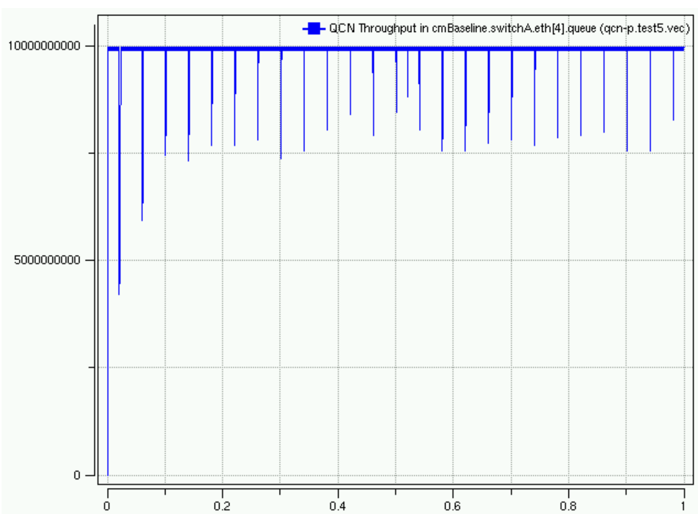
E2CM-P



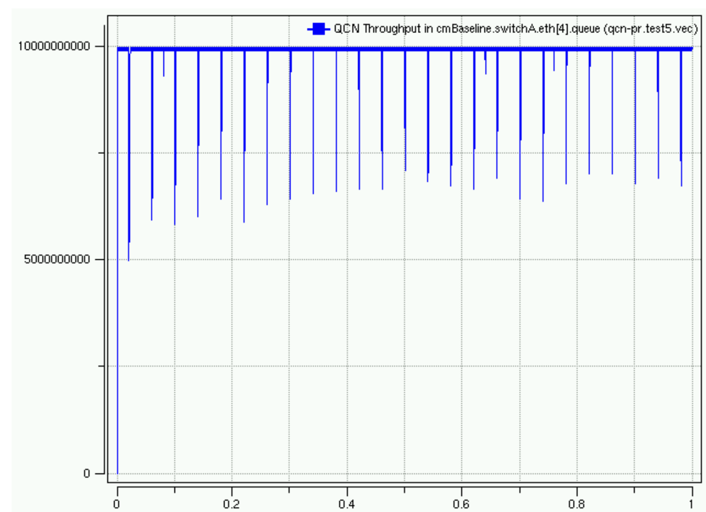
E2CM-PR-P



QCN-P

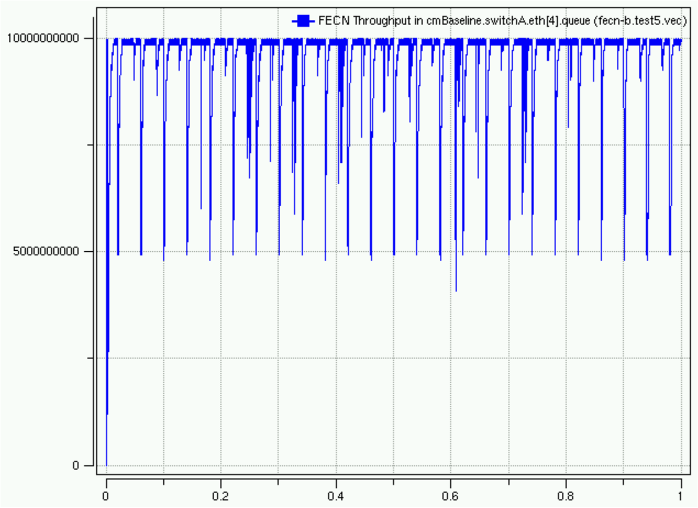


QCN-PR



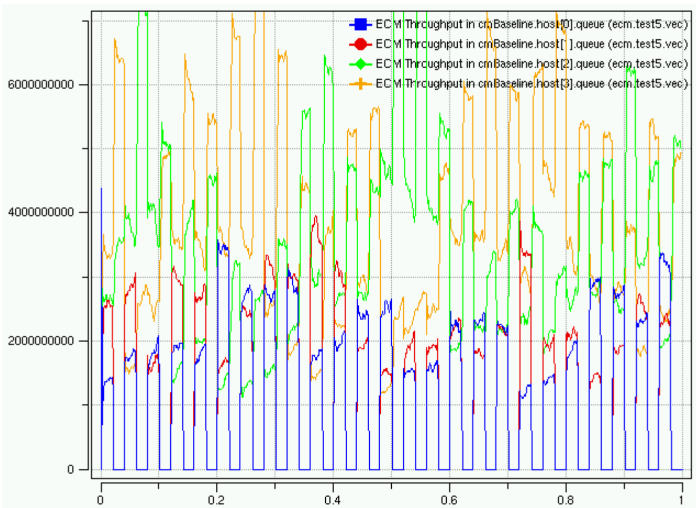
Test 5: Throughput

FECN-B

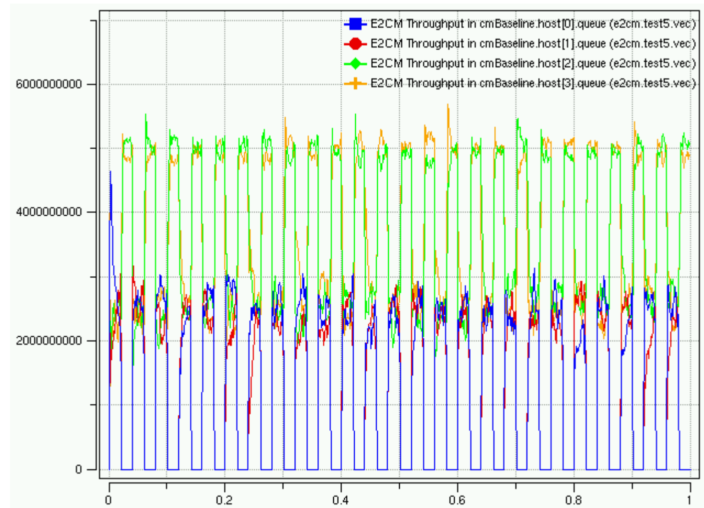


Test 5: Fairness

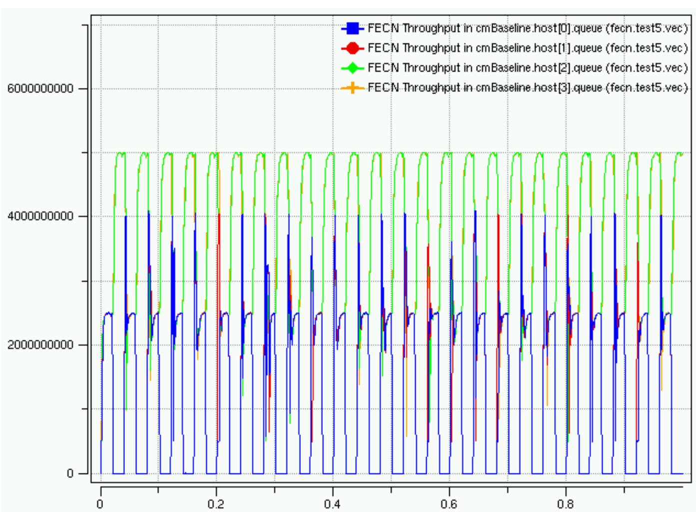
ECM



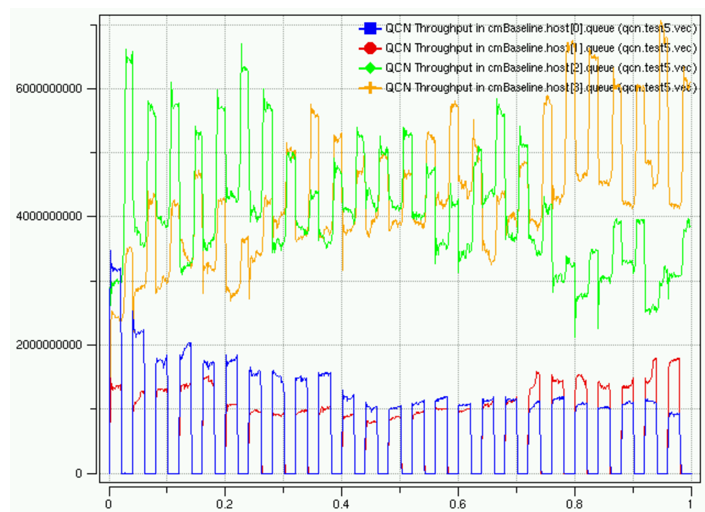
E2CM



FECN

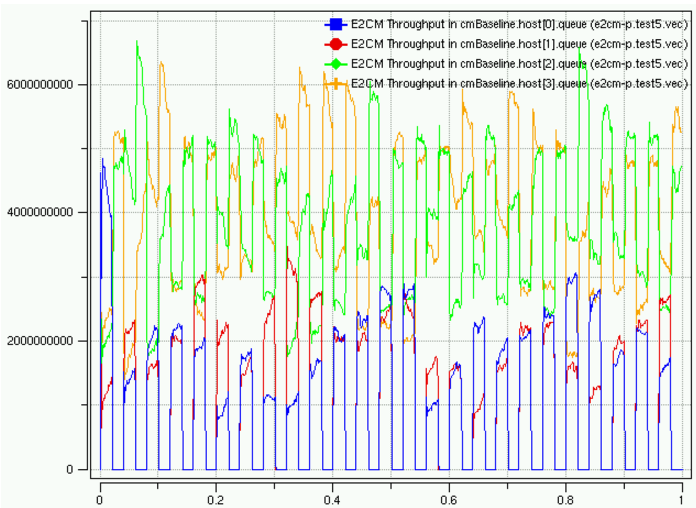


QCN

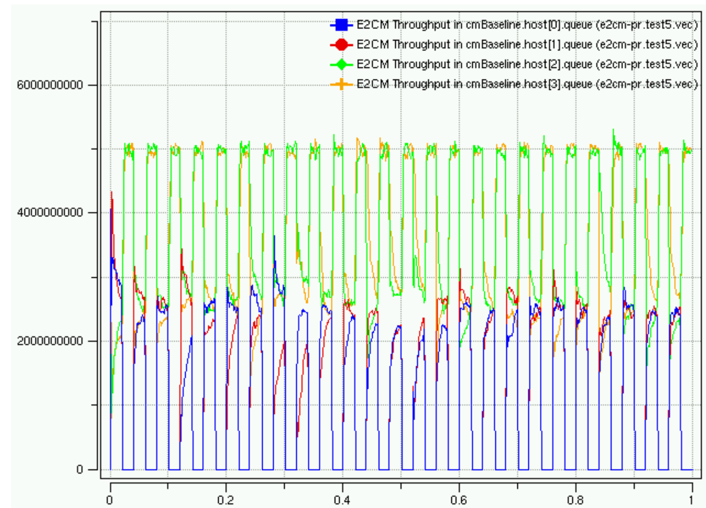


Test 5: Fairness

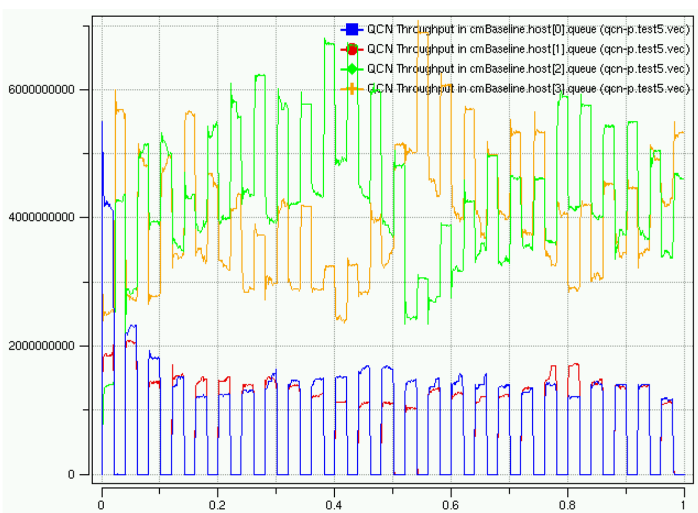
E2CM-P



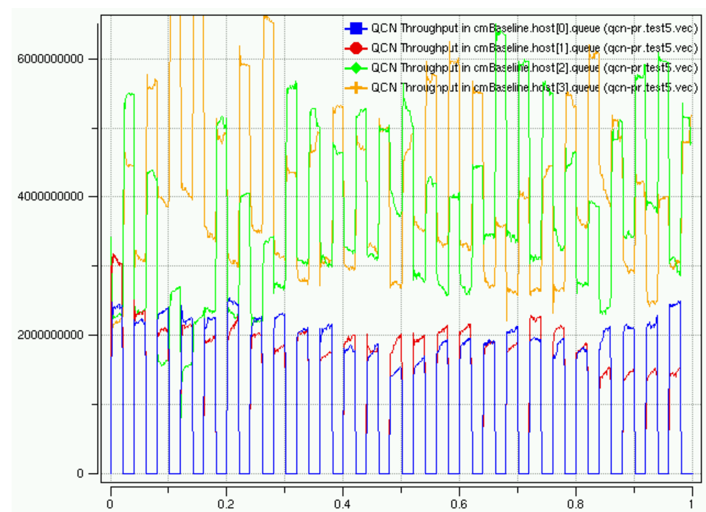
E2CM-PR



QCN-P

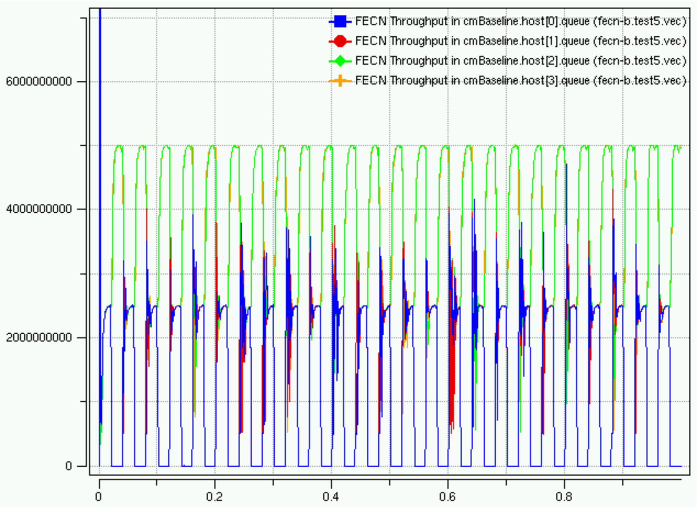


QCN-PR

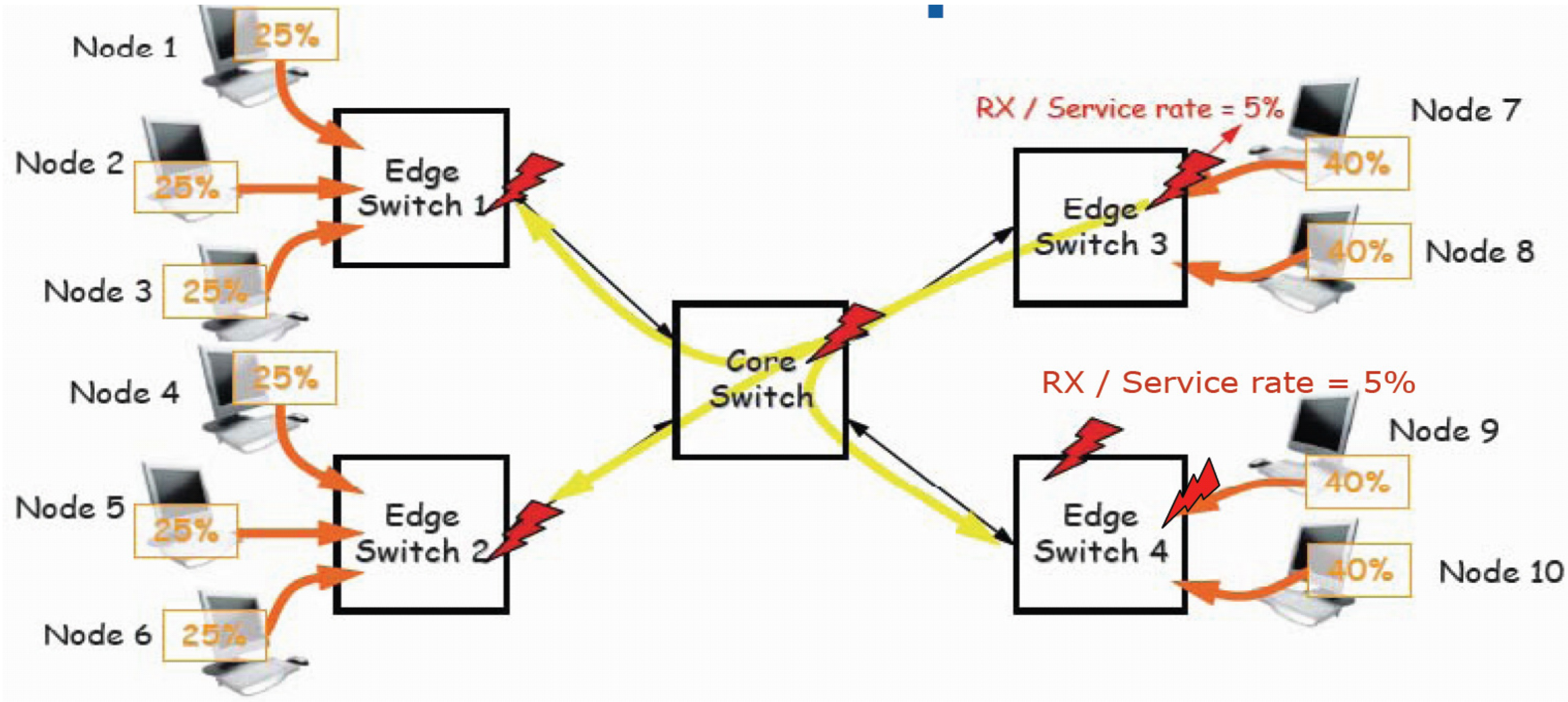


Test 5: Fairness

FECN-B



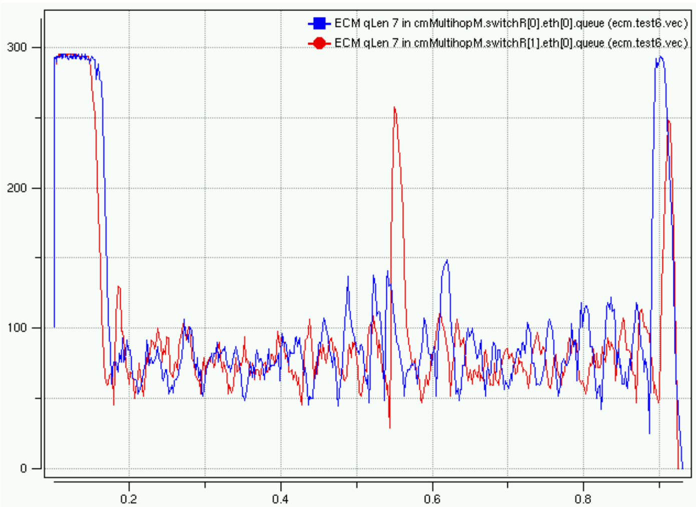
Test 6: Output-Generated Dual Hotspot, Multi-Hop



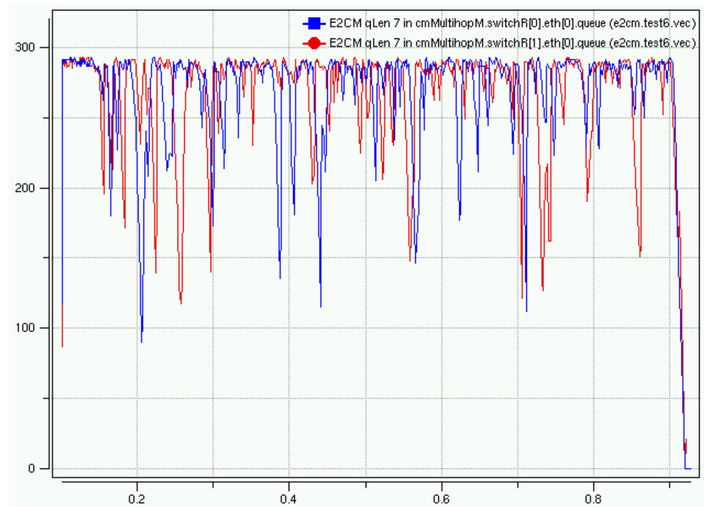
- All: Uniform distribution traffic (background traffic)
- Nodes 1-6: 25% (2.5 Gbps), Nodes 7-10: 40% (4 Gbps)
- Two Hotspots: Node 7 & 9 service rate = 5% (Rx only)

Test 6: Queue Length

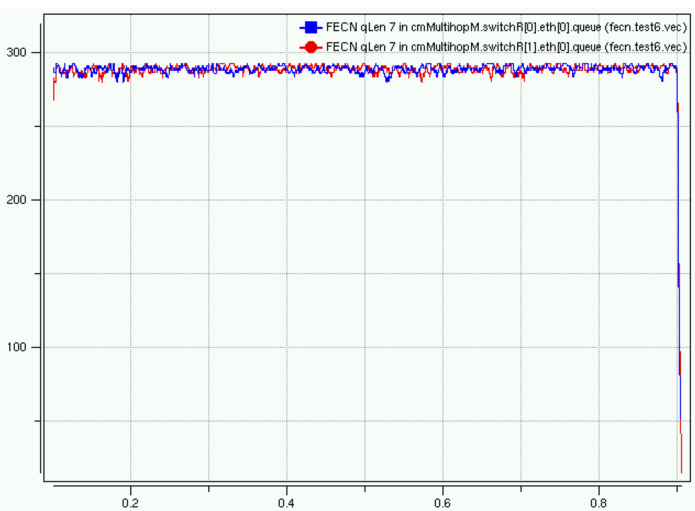
ECM



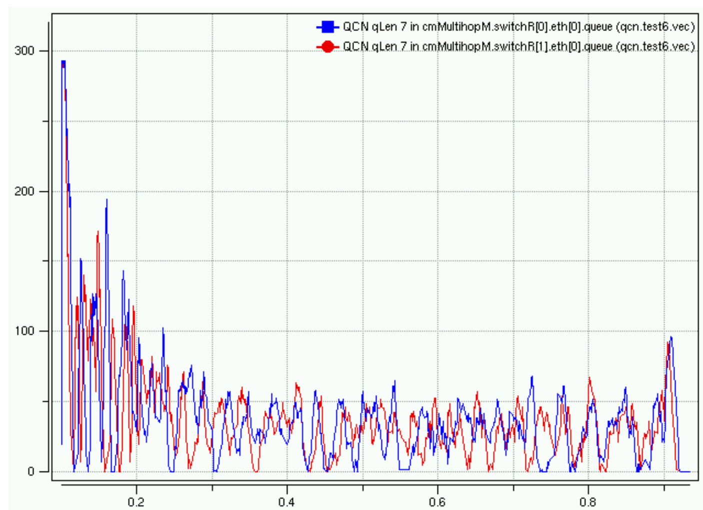
E2CM



FECN

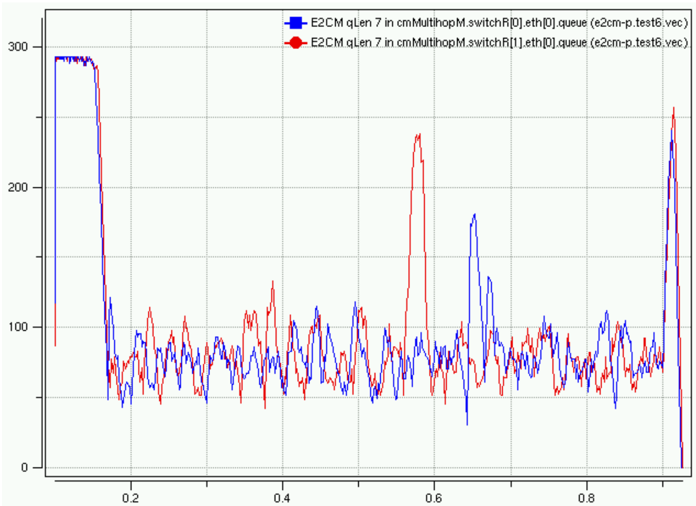


QCN

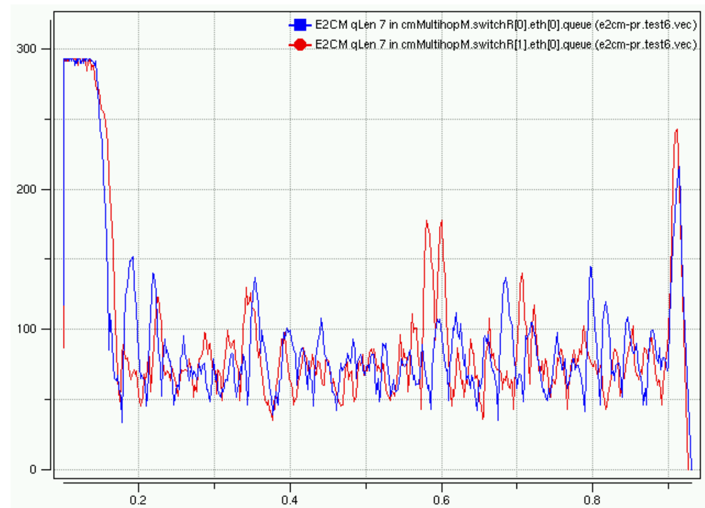


Test 6: Queue Length

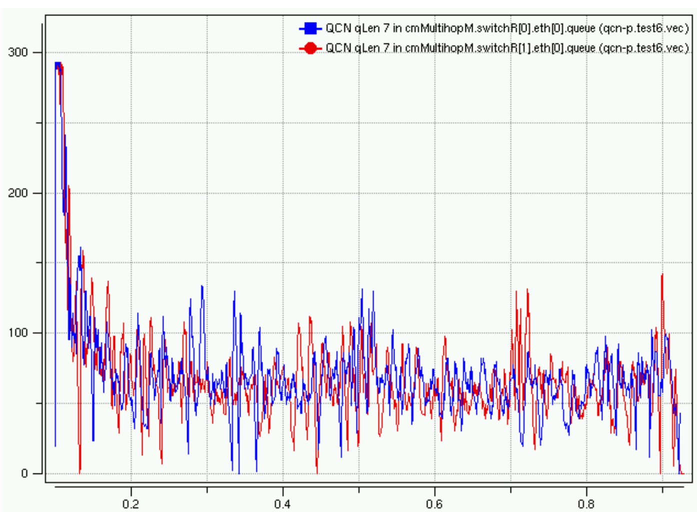
E2CM-P



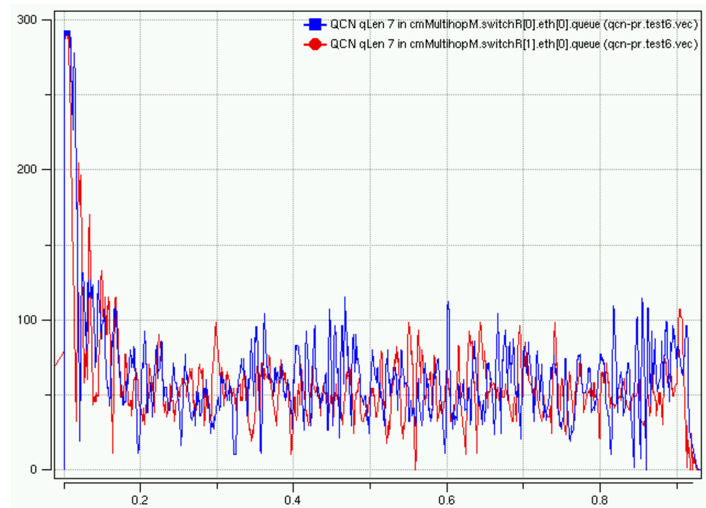
E2CM-PR



QCN-P

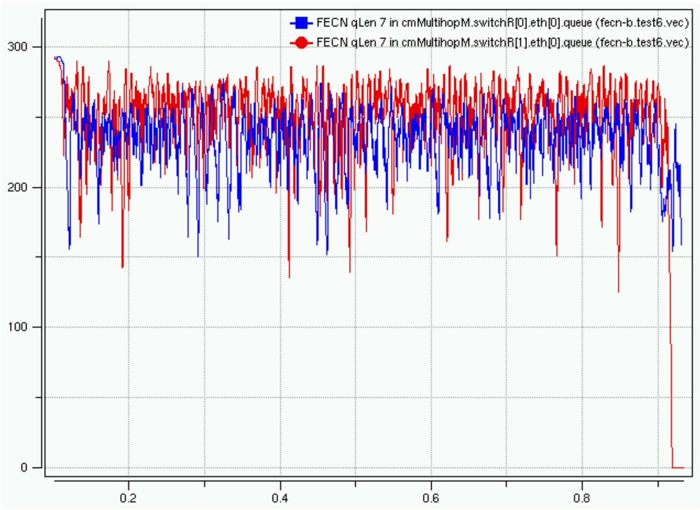


QCN-PR

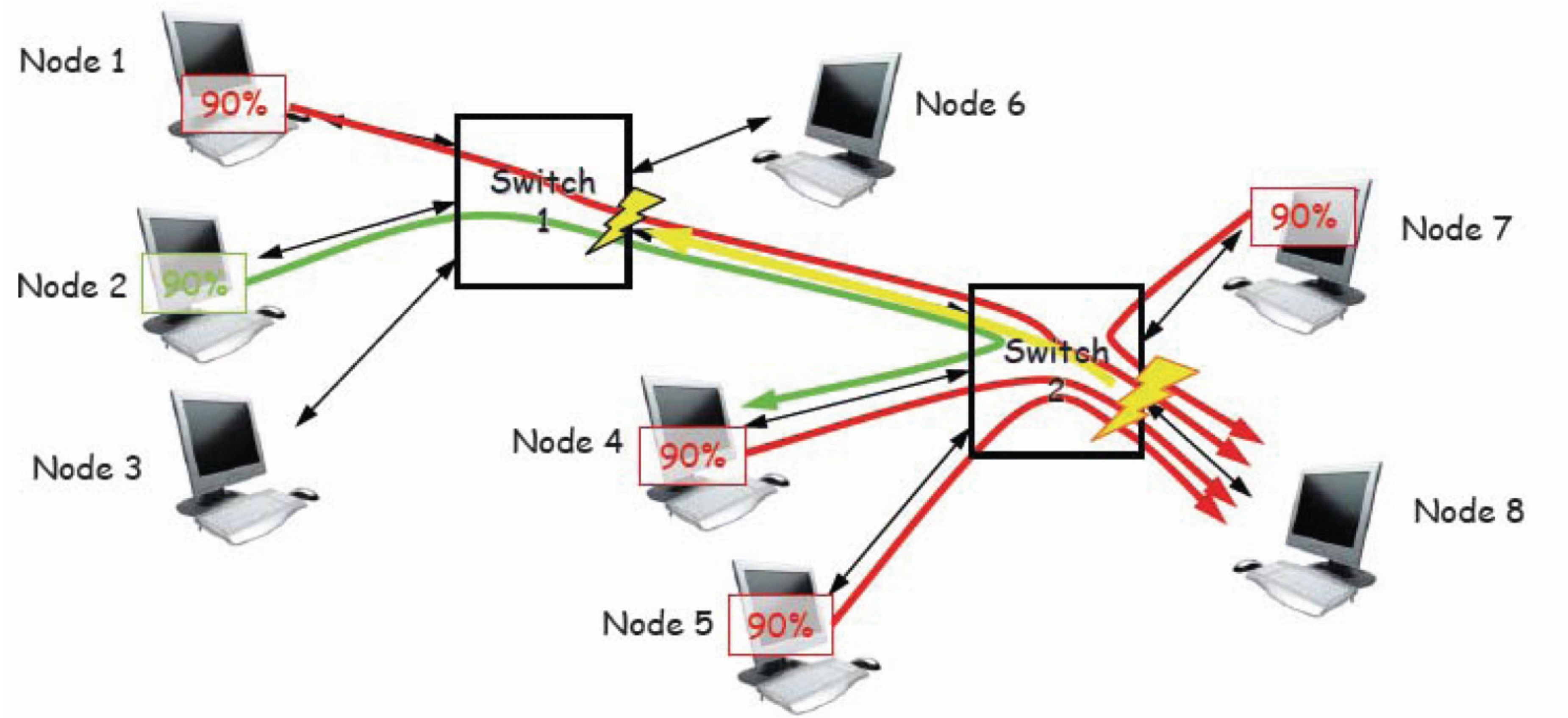


Test 6: Queue Length

FECN-B



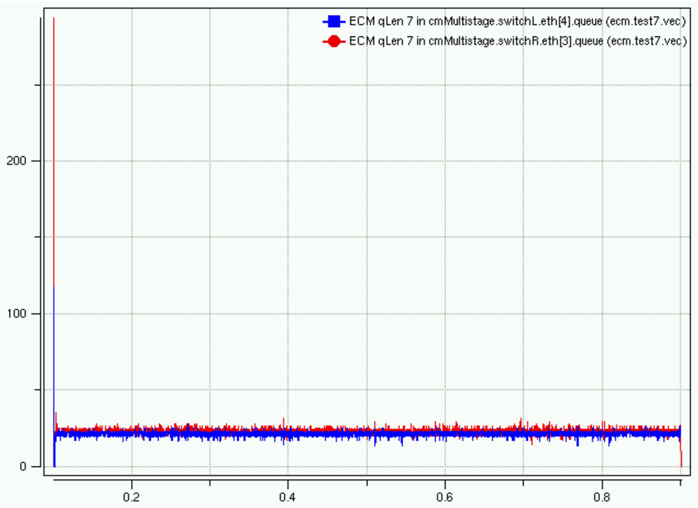
Test 7: Multi-stage Dual Hotspot (Light & Heavy)



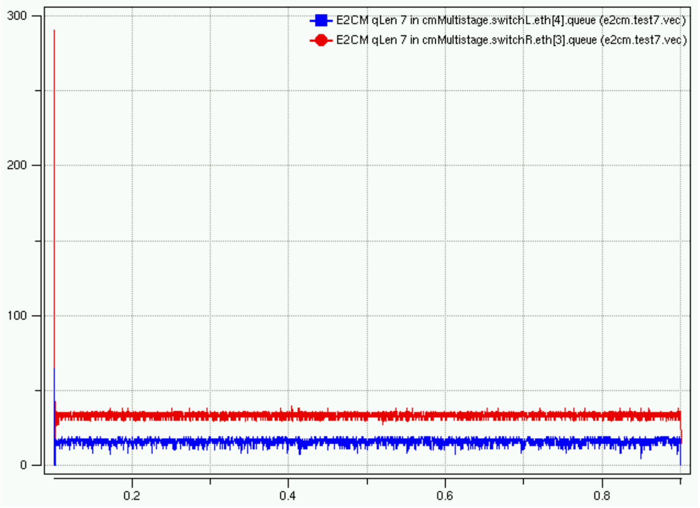
- Two switches, all links 10 Gbps, no background traffic
- Four flows of 9 Gbps each from nodes 1,4,5,7 to node 8
- One flow of 9 Gbps each from node 2 to node 4
- Two congestion points
  - Port from switch 1 to switch 2
  - Port from switch 2 to node 8
- Fair allocation should provide 2.5 Gbps for all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

Test 7: Queue Length

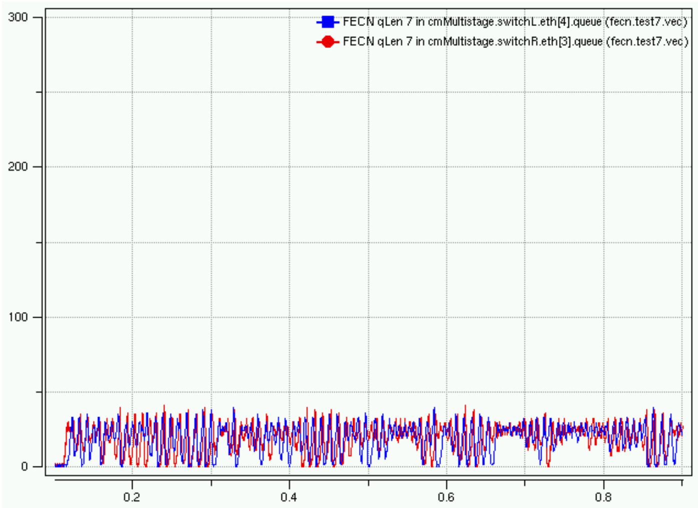
ECM



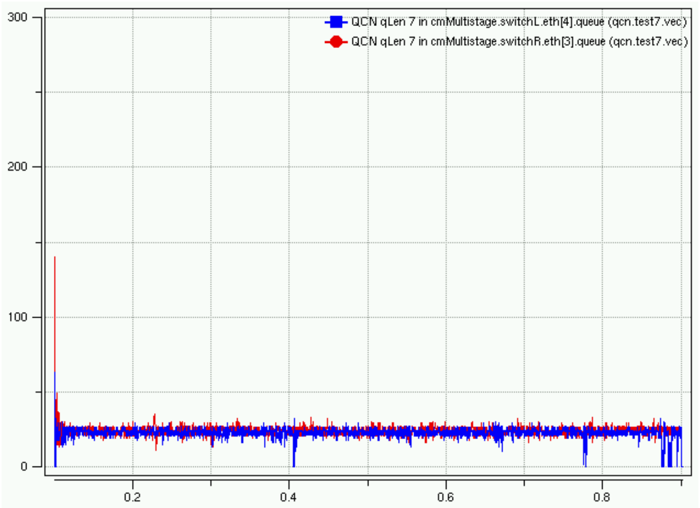
E2CM



FECN

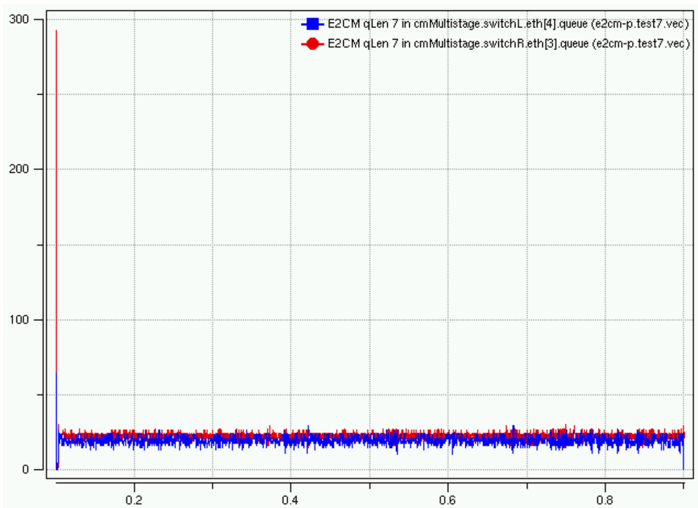


QCN

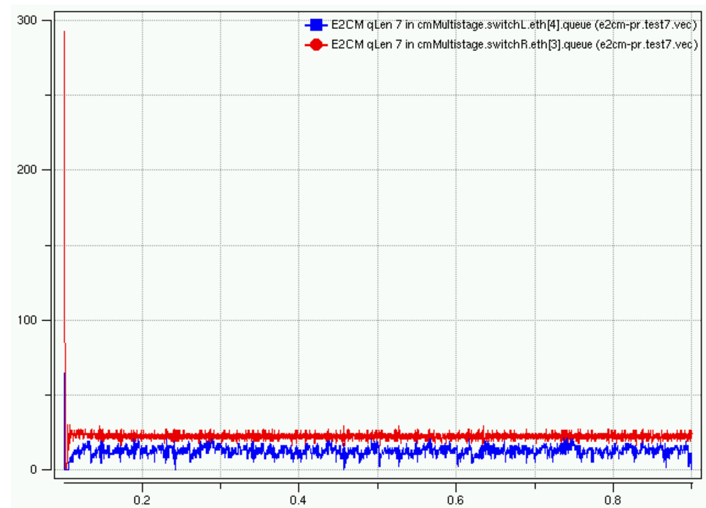


Test 7: Queue Length

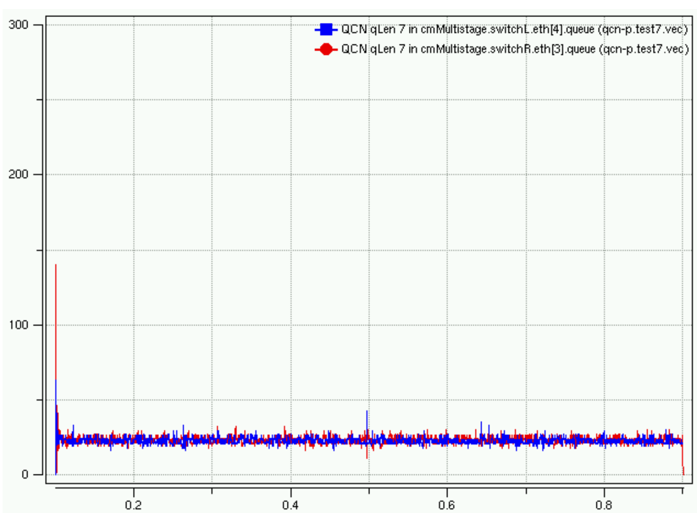
E2CM-P



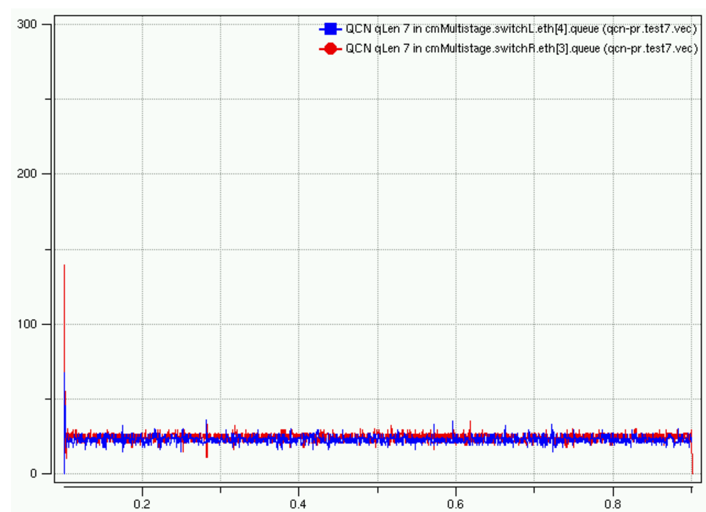
E2CM-PR



QCN-P

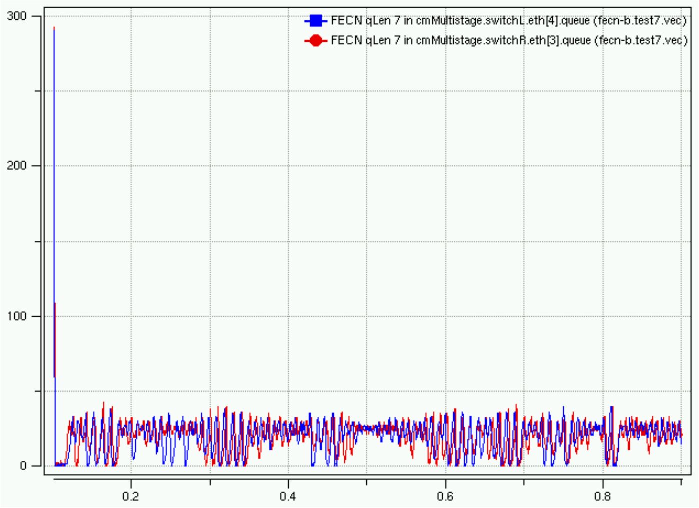


QCN-PR



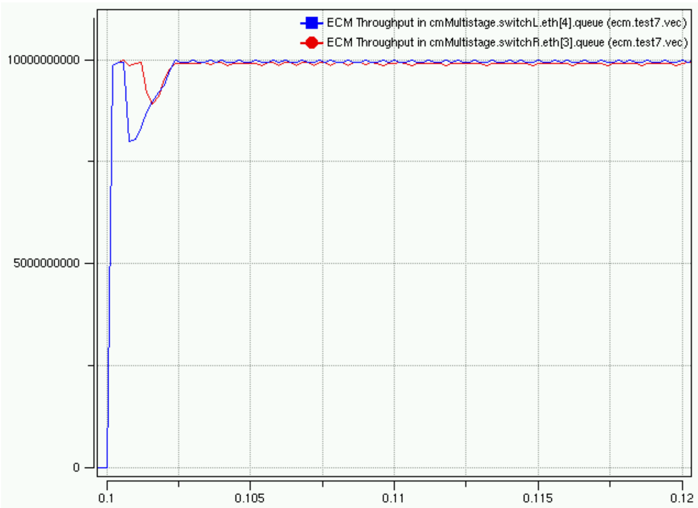
Test 7: Queue Length

FECN-B



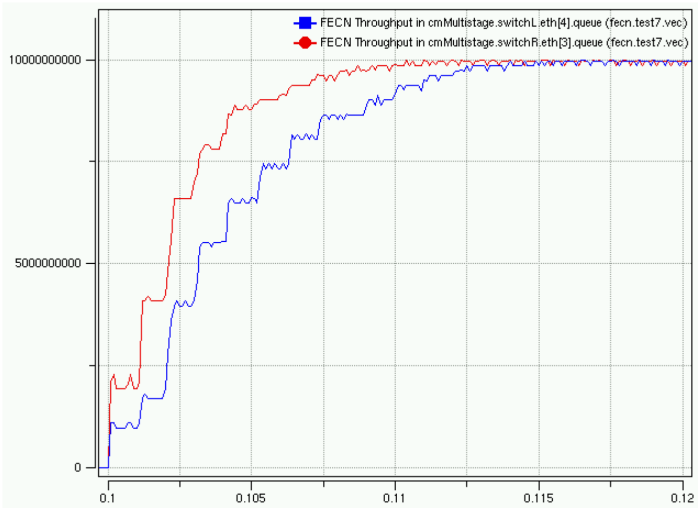
Test 7: Throughput

ECM

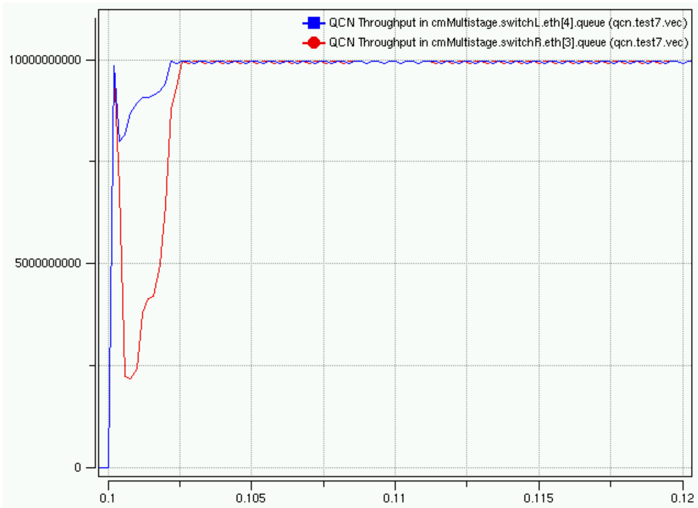


E2CM

FECN

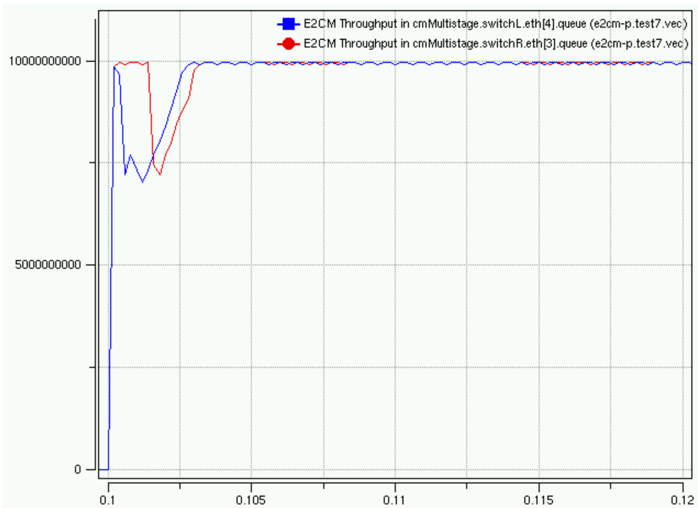


QCN

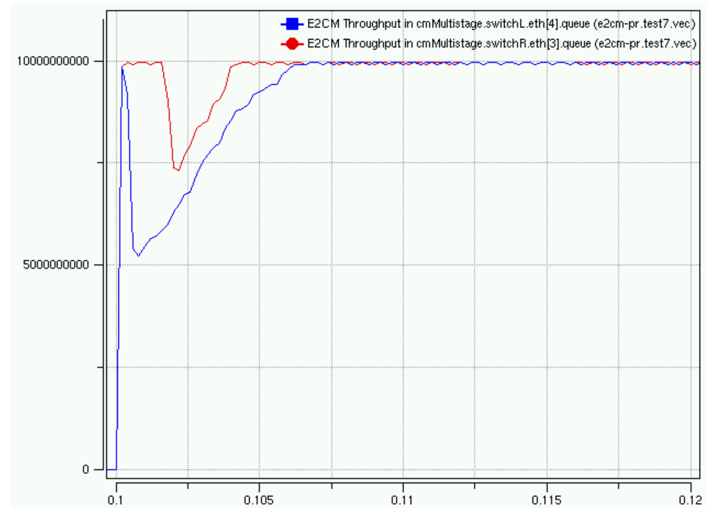


Test 7: Throughput

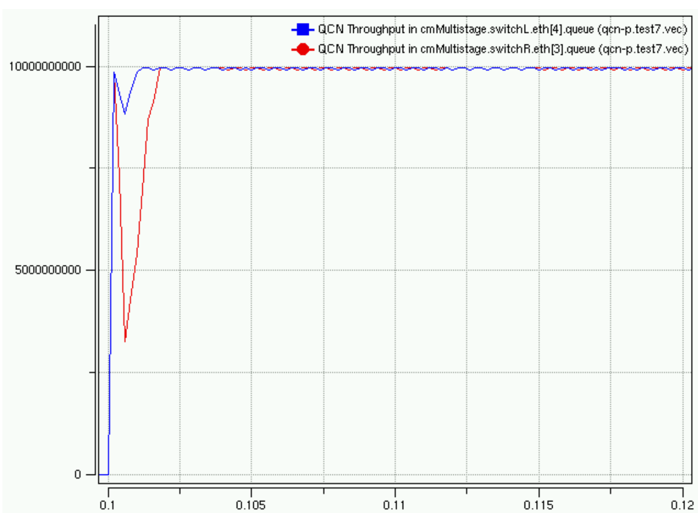
E2CM-P



E2CM-PR



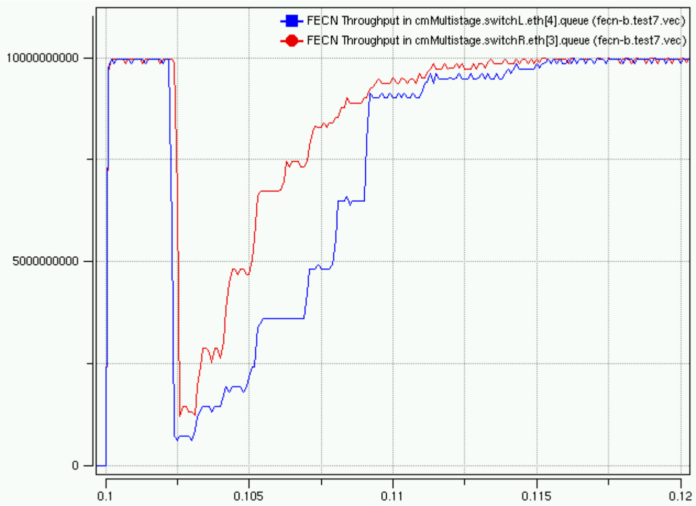
QCN-P



QCN-PR

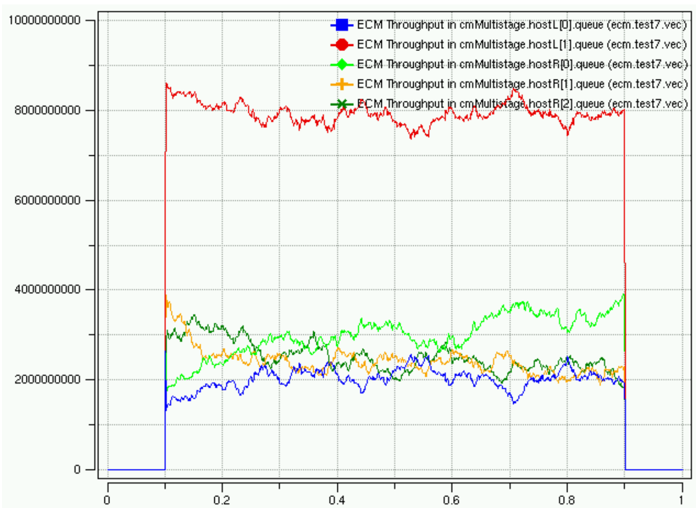
Test 7: Throughput

FECN-B

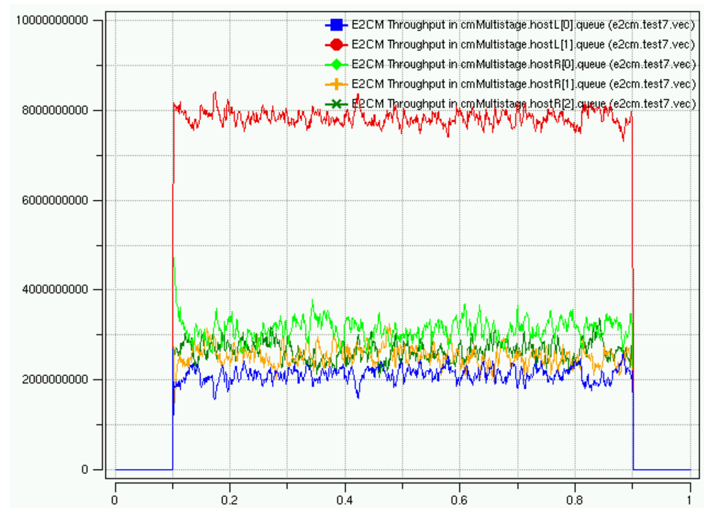


Test 7: Fairness

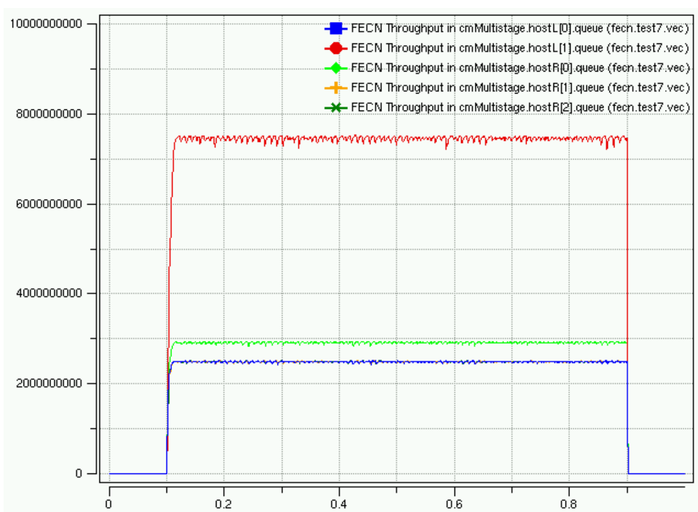
ECM



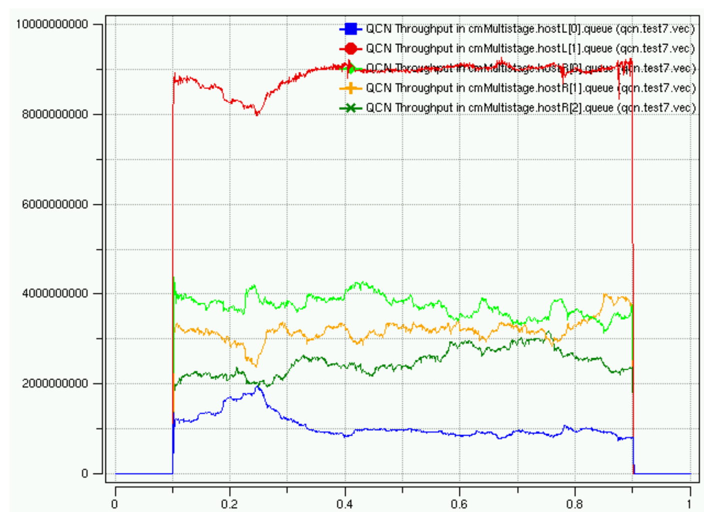
E2CM



FECN

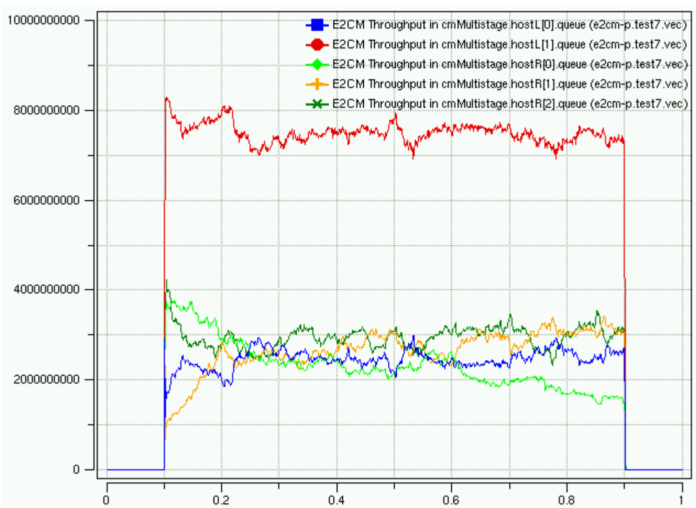


QCN

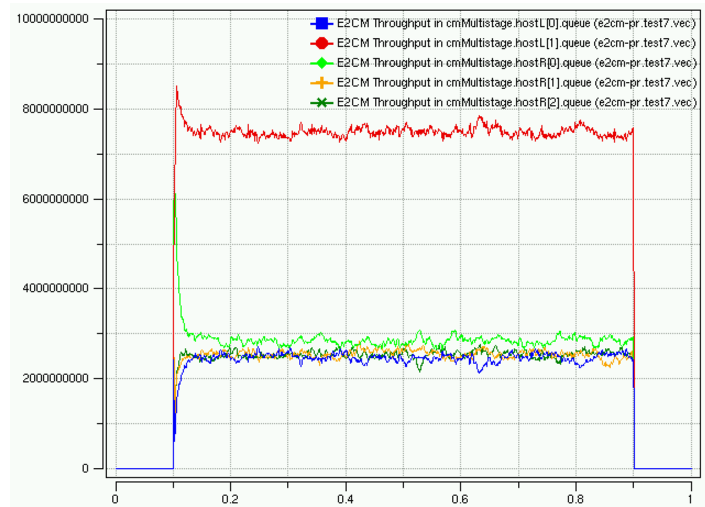


Test 7: Fairness

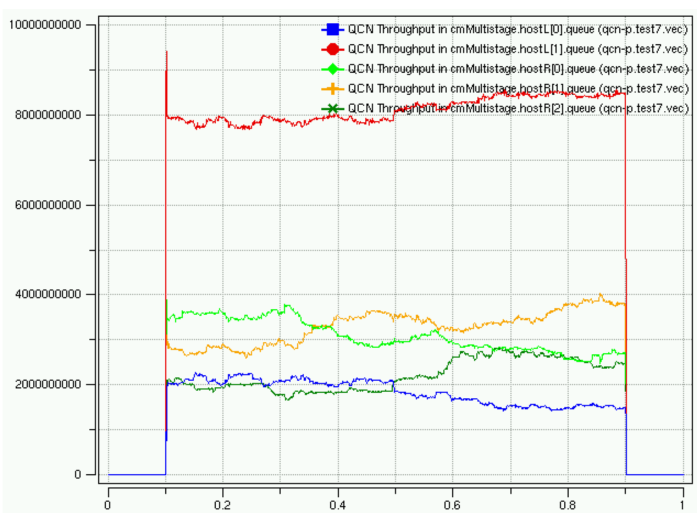
E2CM-P



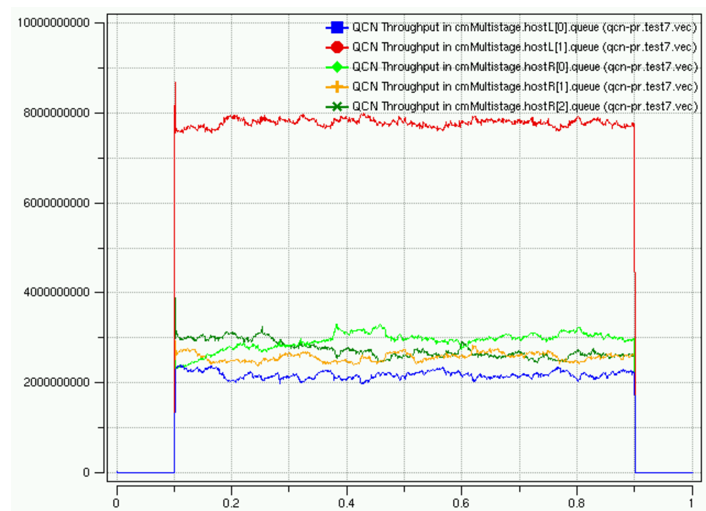
E2CM-PR



QCN-P

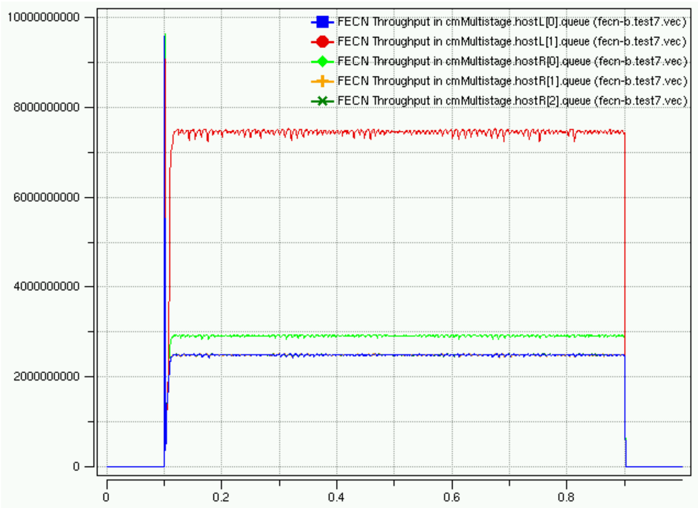


QCN-PR

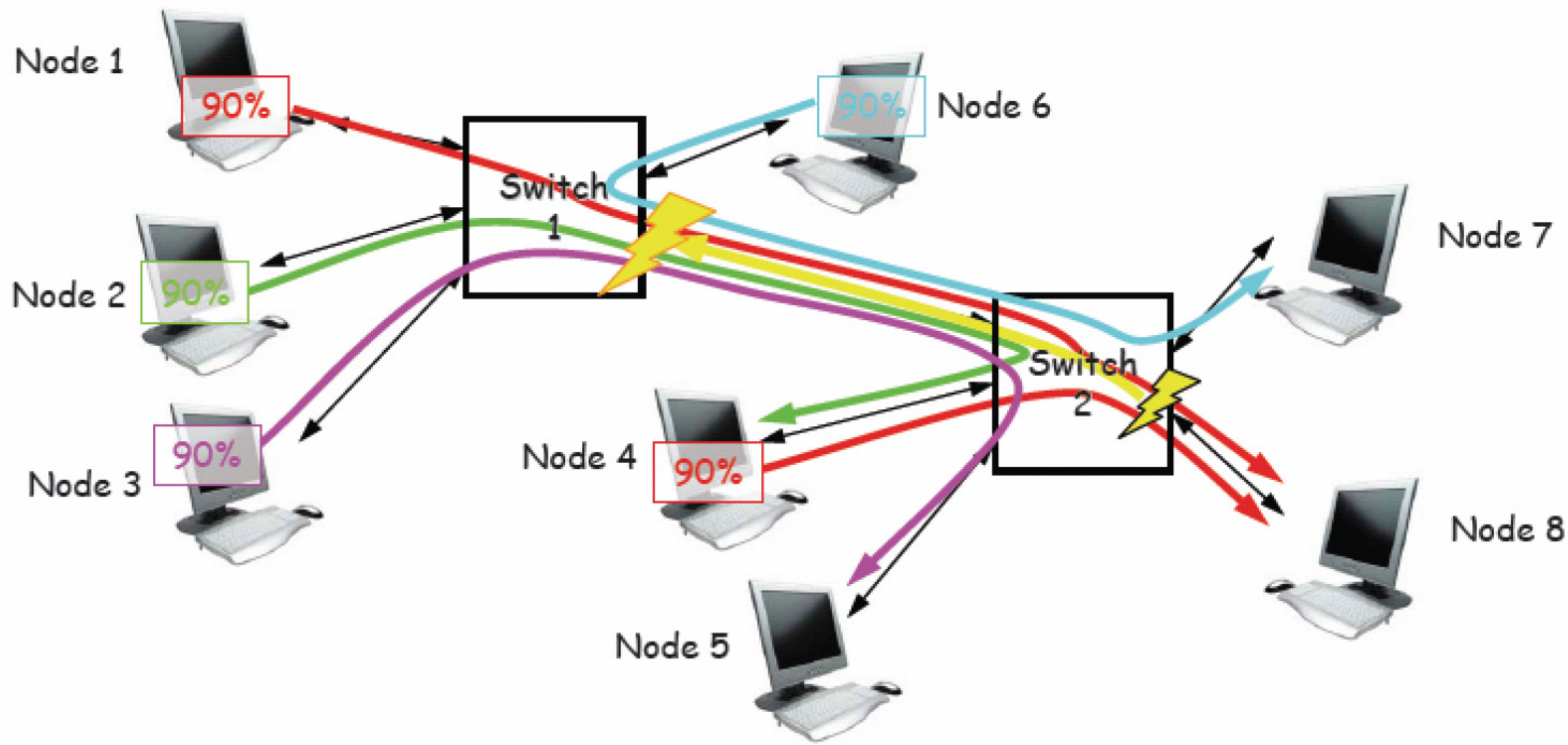


Test 7: Fairness

FECN-B



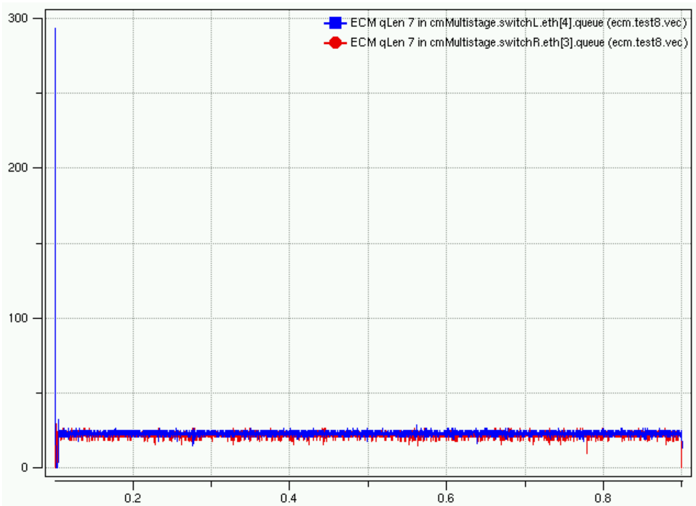
Test 8: Multi-stage Dual Hotspot (Heavy & Light)



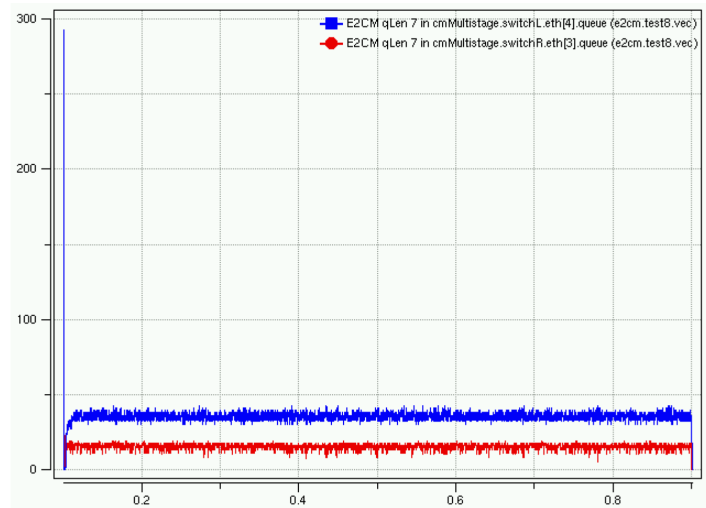
- Two switches, all links 10 Gbps, no background traffic
- Two flows of 9 Gbps each from nodes 1 and 4 to node 8
- Three flows of 9 Gbps each from node 2 to node 4, 3 to 5, and 6 to 7
- Two congestion points
  - Port from switch 1 to switch 2
  - Port from switch 2 to node 8
- Fair allocation should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

Test 8: Queue Length

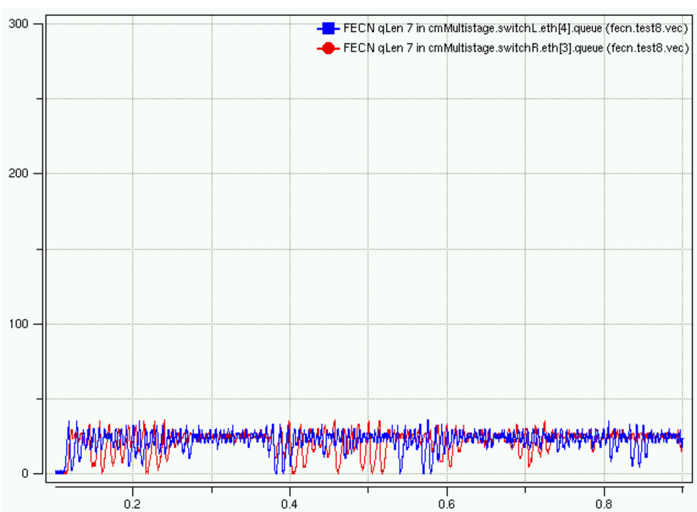
ECM



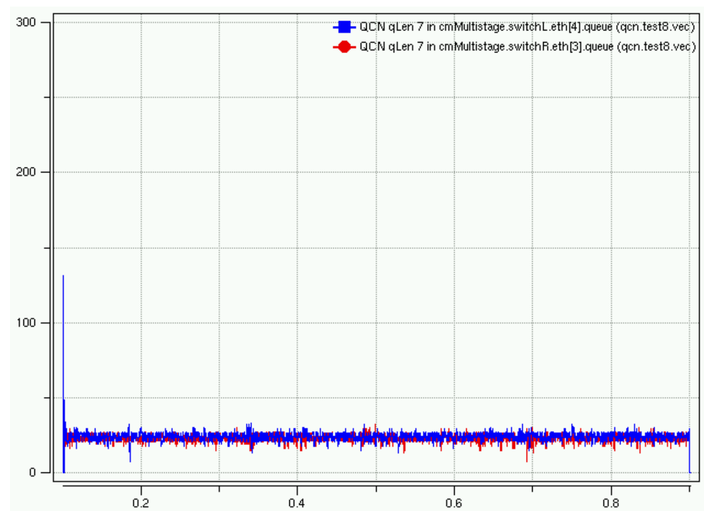
E2CM



FECN

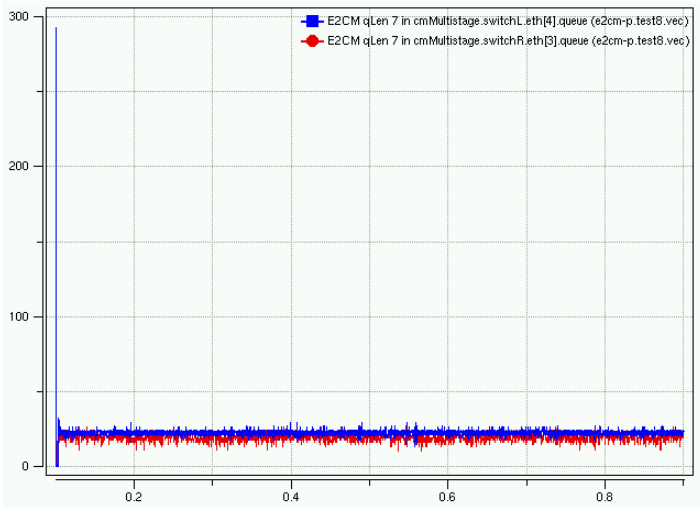


QCN

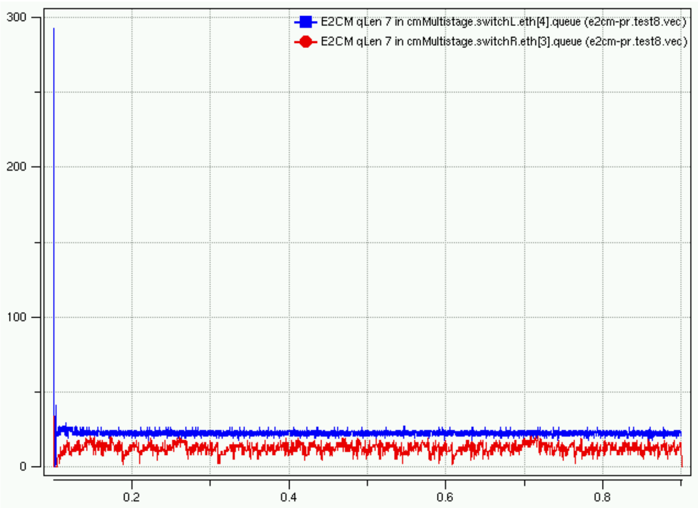


Test 8: Queue Length

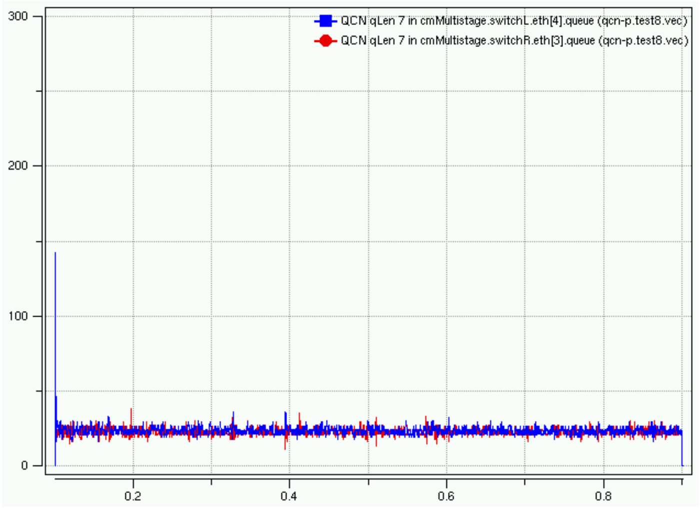
E2CM-P



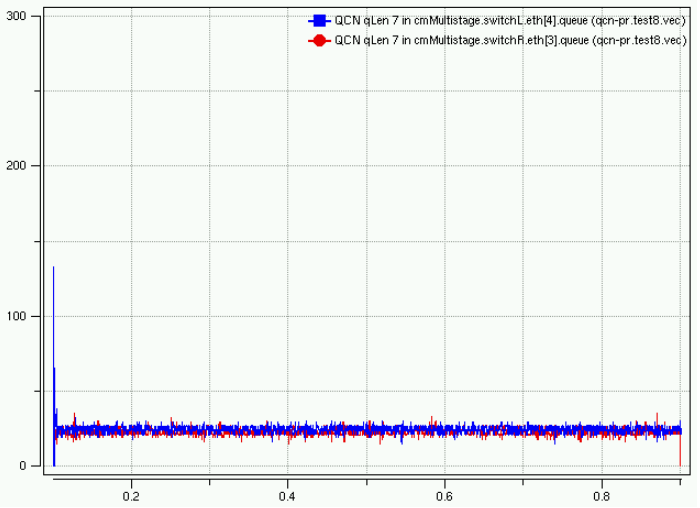
E2CM-PR



QCN-P

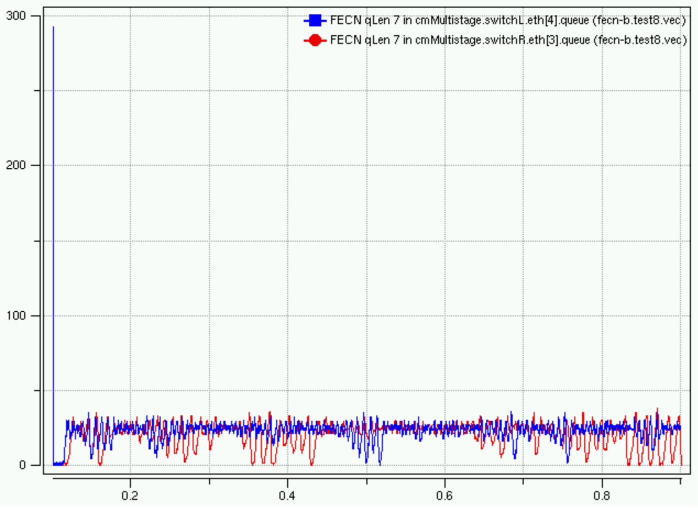


QCN-PR



Test 8: Queue Length

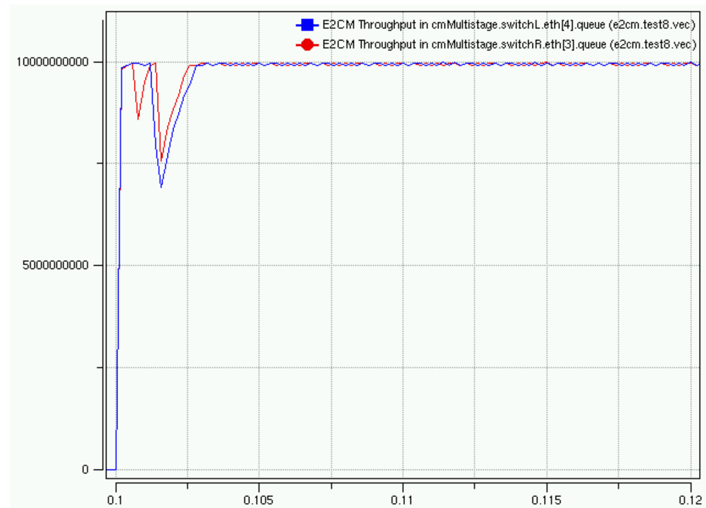
FECN-B



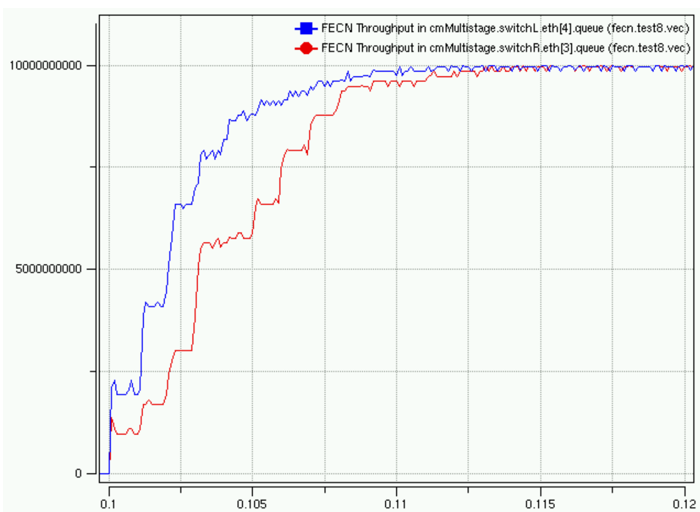
Test 8: Throughput

ECM

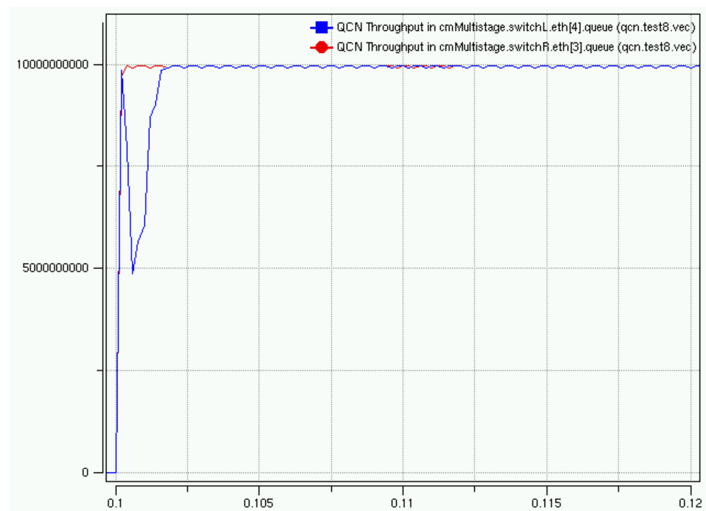
E2CM



FECN

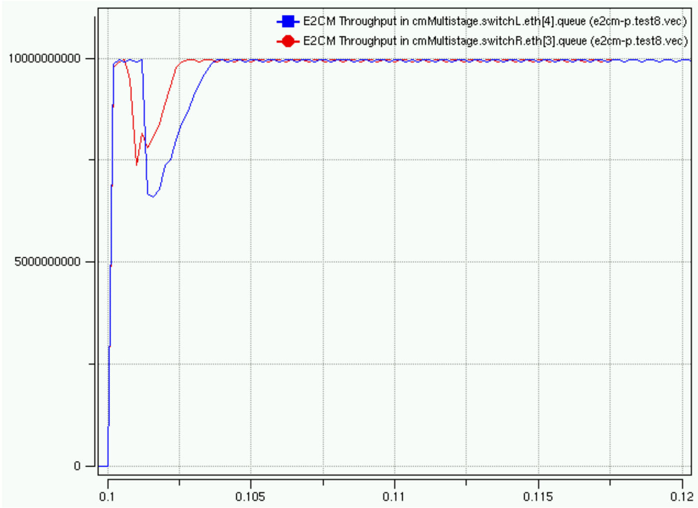


QCN

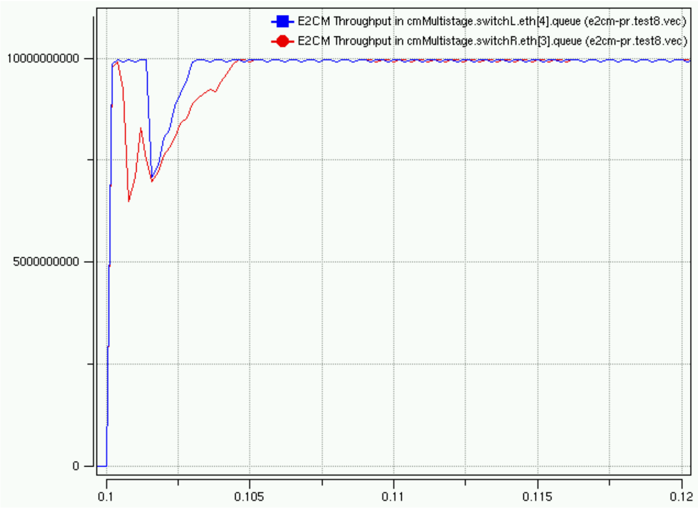


Test 8: Throughput

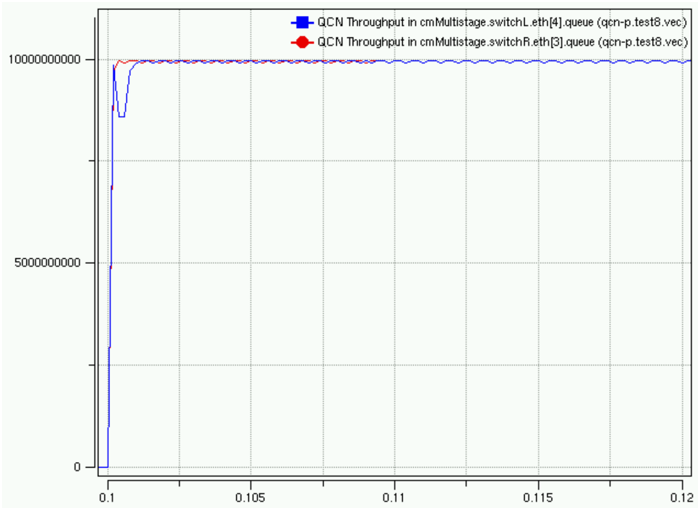
E2CM-P



E2CM-PR



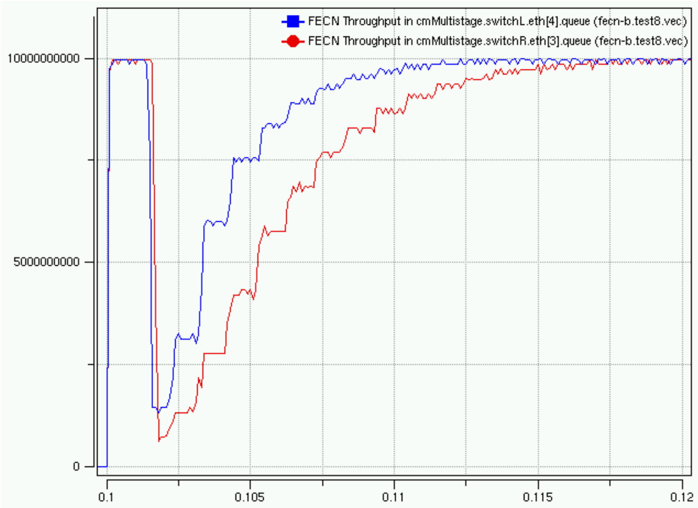
QCN-P



QCN-PR

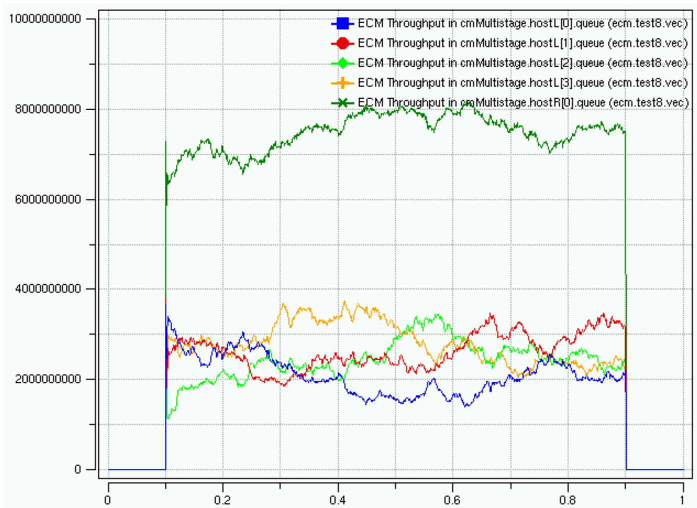
Test 8: Throughput

FECN-B

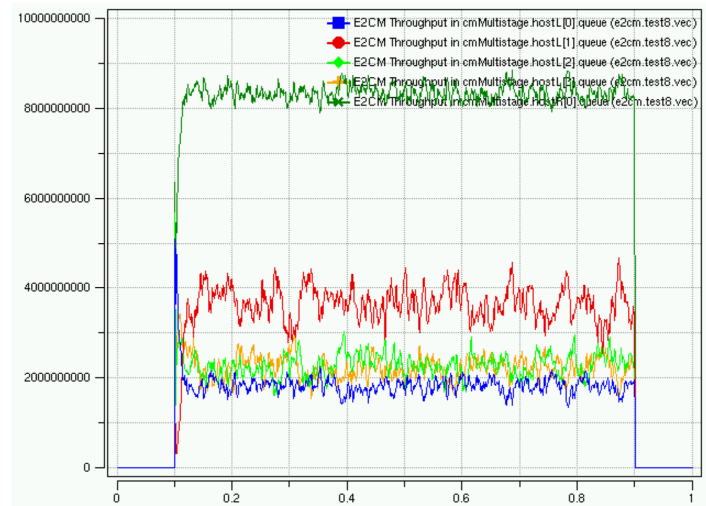


Test 8: Fairness

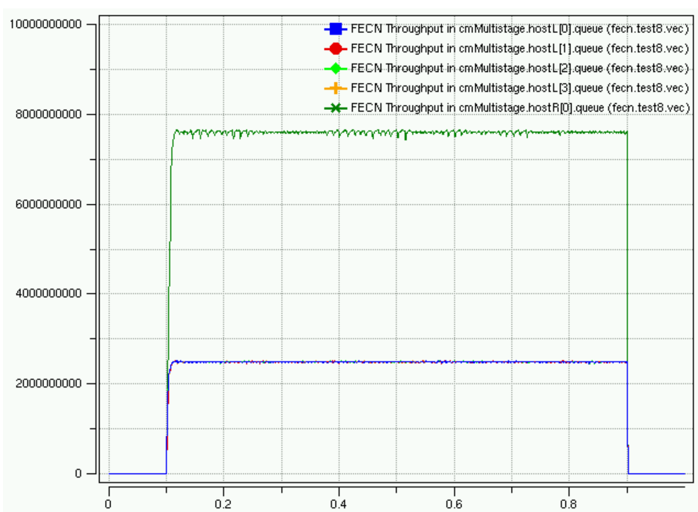
ECM



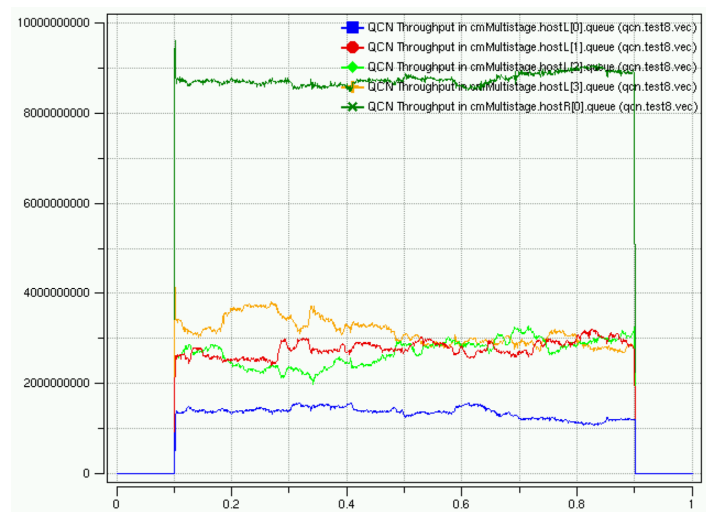
E2CM



FECN

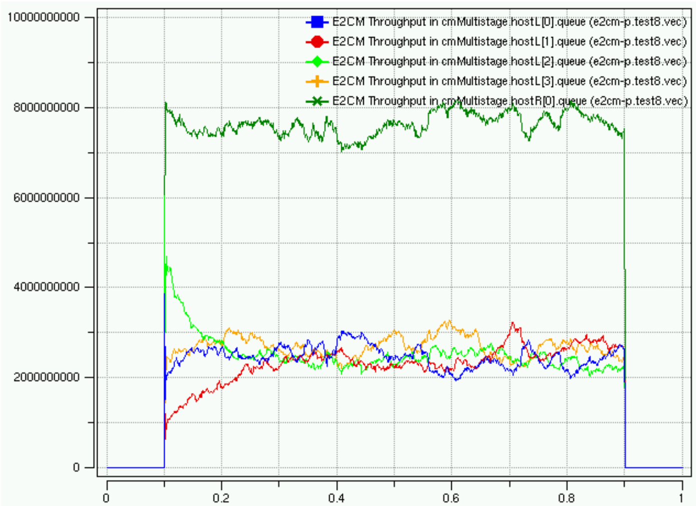


QCN

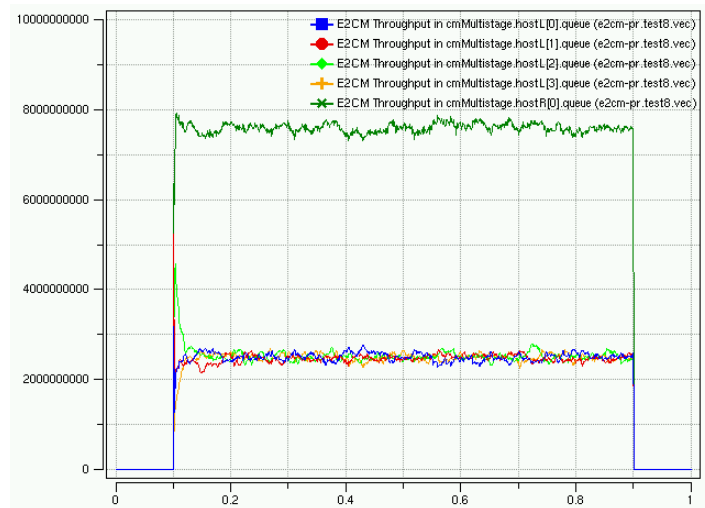


Test 8: Fairness

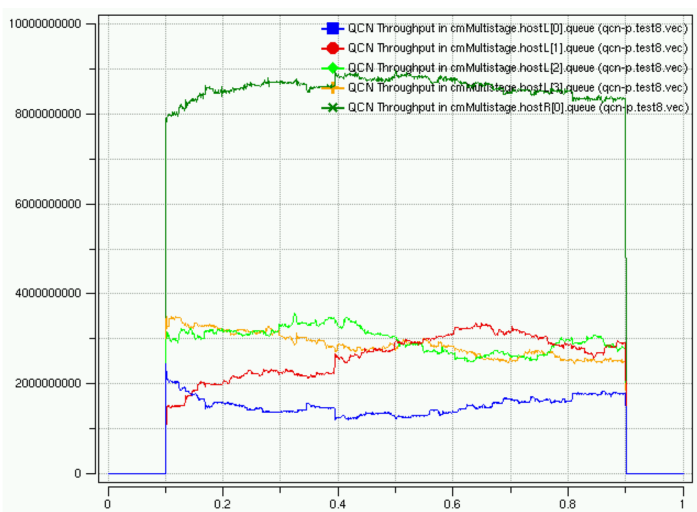
E2CM-P



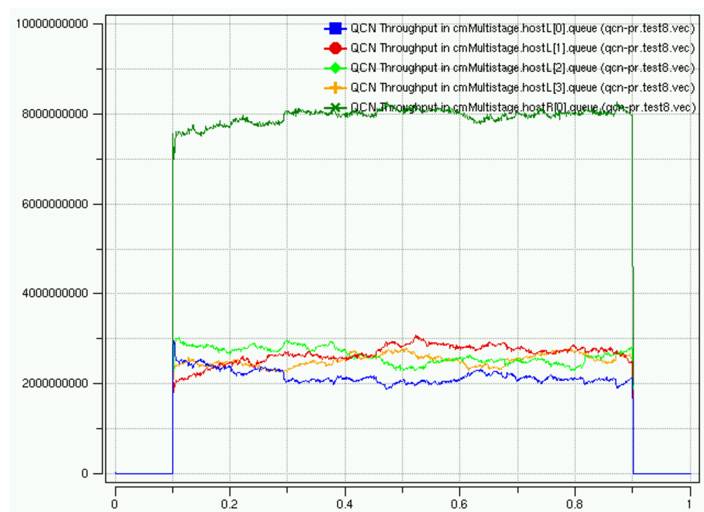
E2CM-PR



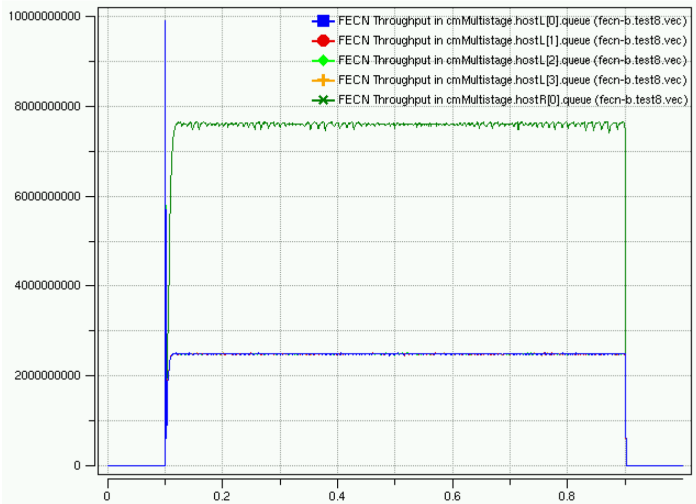
QCN-P



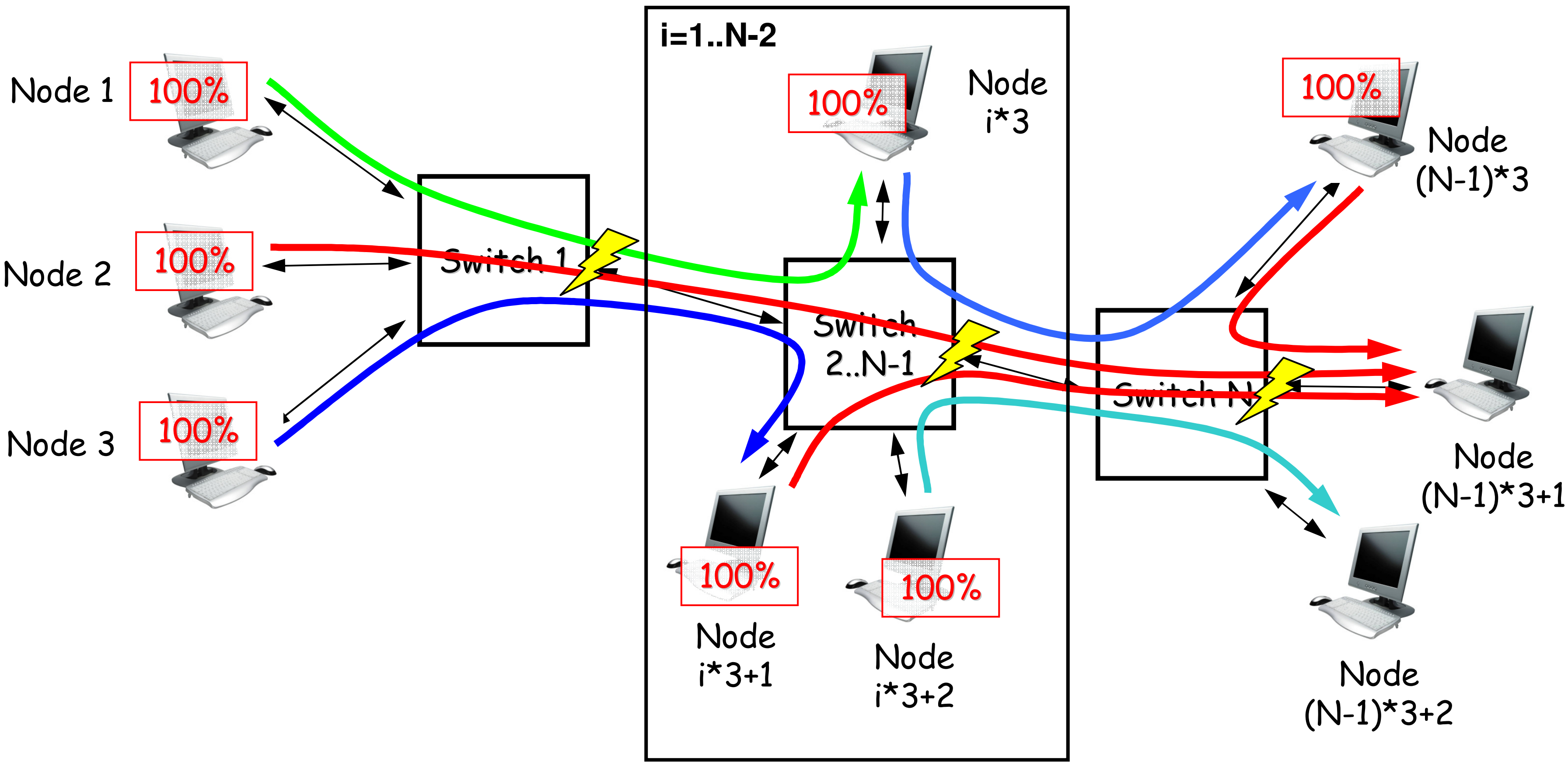
QCN-PR



FECN-B


FECN Throughput in cmMultistage.hostL[0].queue (fecn-b.test8.vec)
FECN Throughput in cmMultistage.hostL[1].queue (fecn-b.test8.vec)
FECN Throughput in cmMultistage.hostL[2].queue (fecn-b.test8.vec)
FECN Throughput in cmMultistage.hostL[3].queue (fecn-b.test8.vec)
FECN Throughput in cmMultistage.hostR[0].queue (fecn-b.test8.vec)
10000000000
8000000000
6000000000
4000000000
2000000000
0
0
0.2
0.4
0.6
0.8

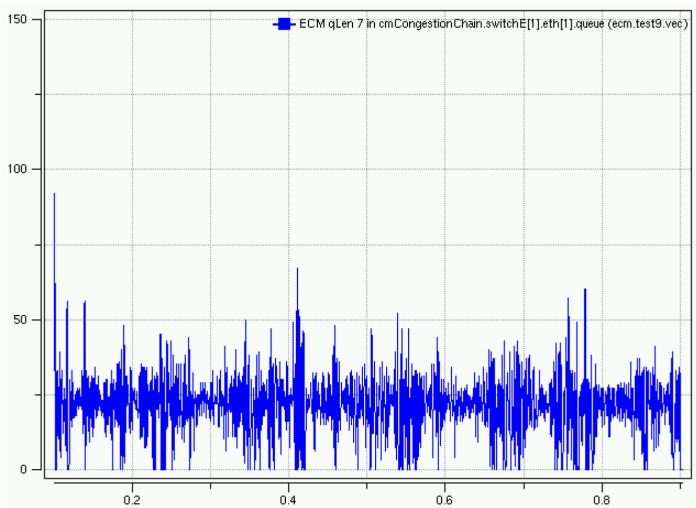
Test 9: 7-stage Hotspot



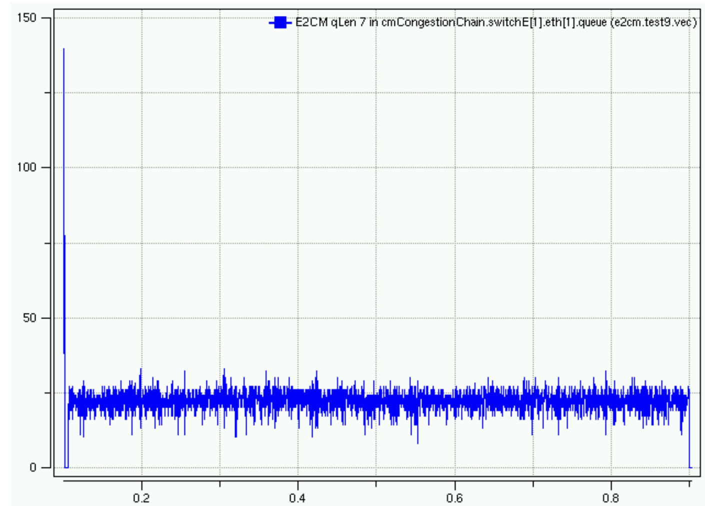
- N=7 switches; 3 hosts per switch
- Node <i> sends to node <i+3>; Node <i+1> sends to node (N-1)*3+1; node <i+2> sends to node <i+4>
- 100% load from all nodes
- Node (N-1)*3+1 receives traffic from <N> sources
- N hotspots

Test 9: Queue Length

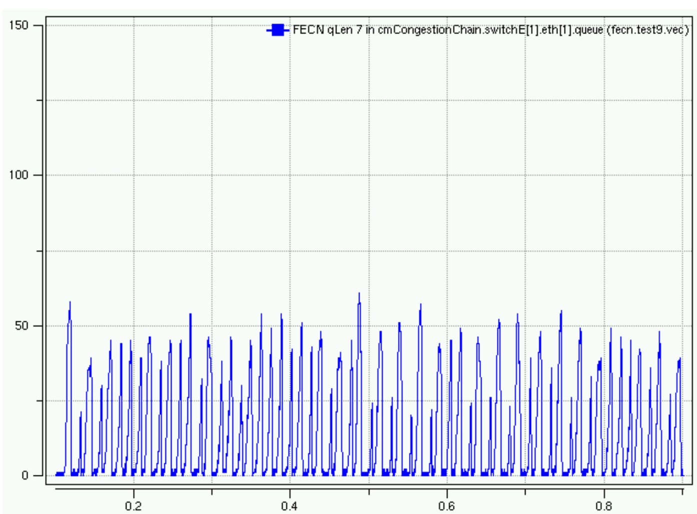
ECM



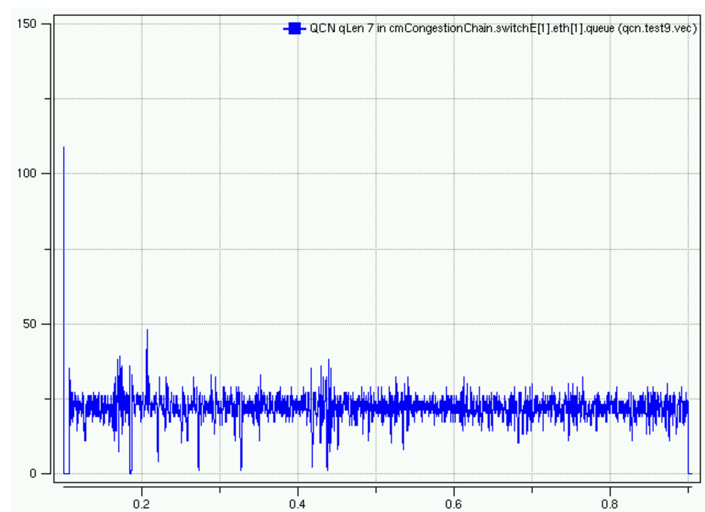
E2CM



FECN

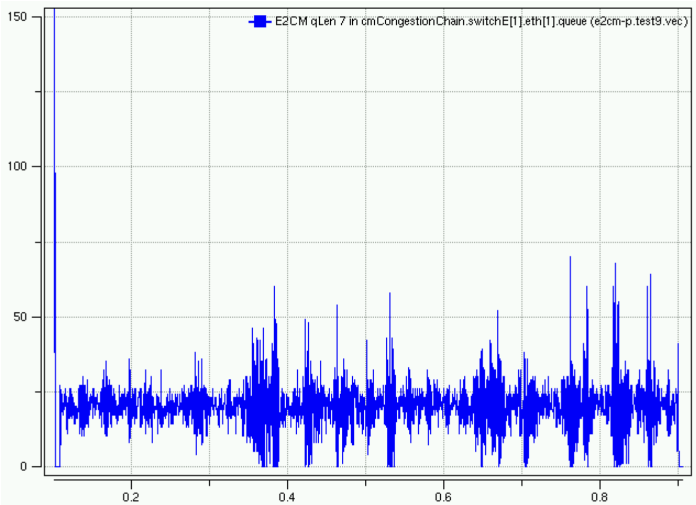


QCN

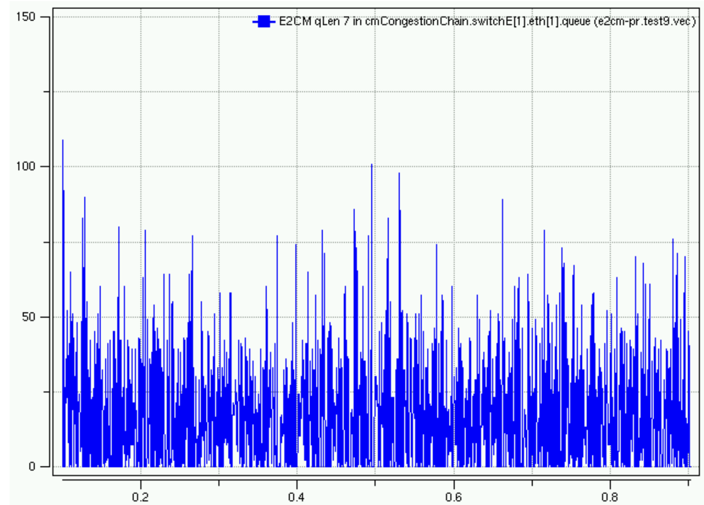


Test 9: Queue Length

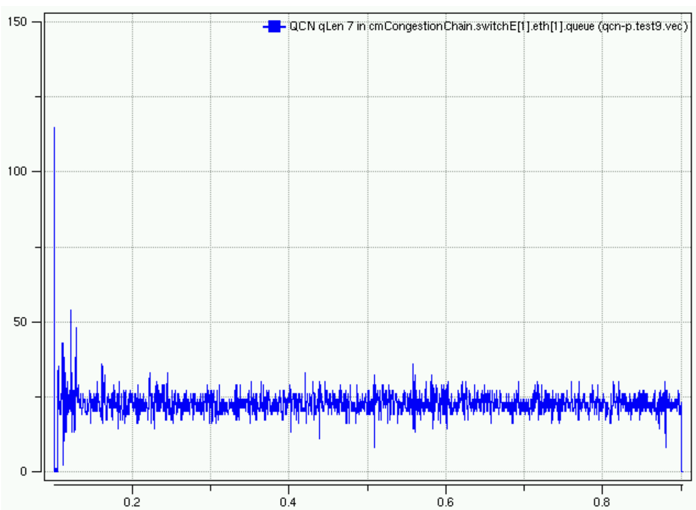
E2CM-P



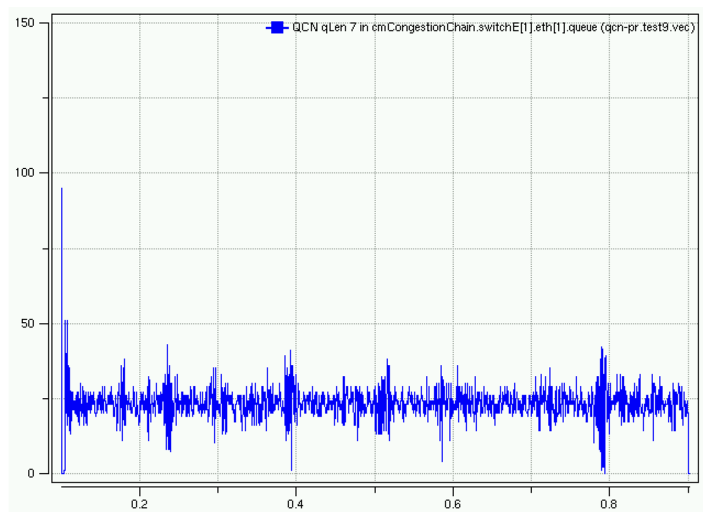
E2CM-PR



QCN-P

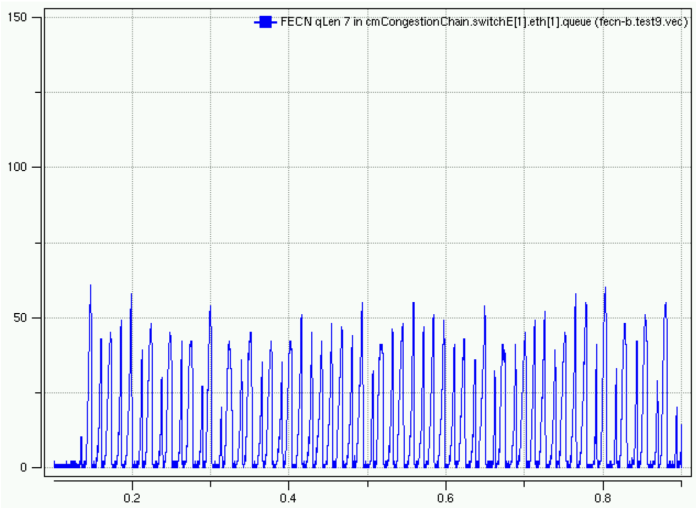


QCN-PR



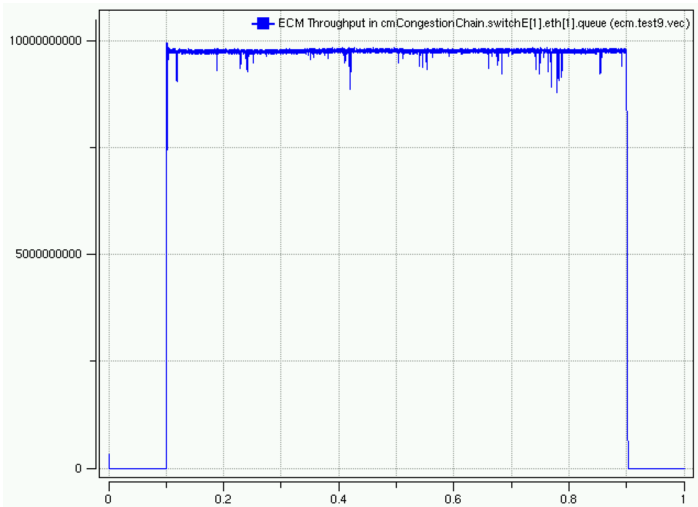
Test 9: Queue Length

FECN-B

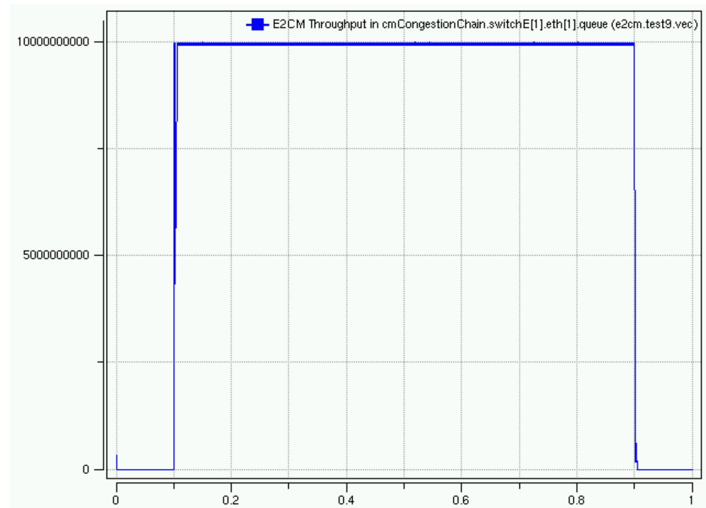


Test 9: Throughput

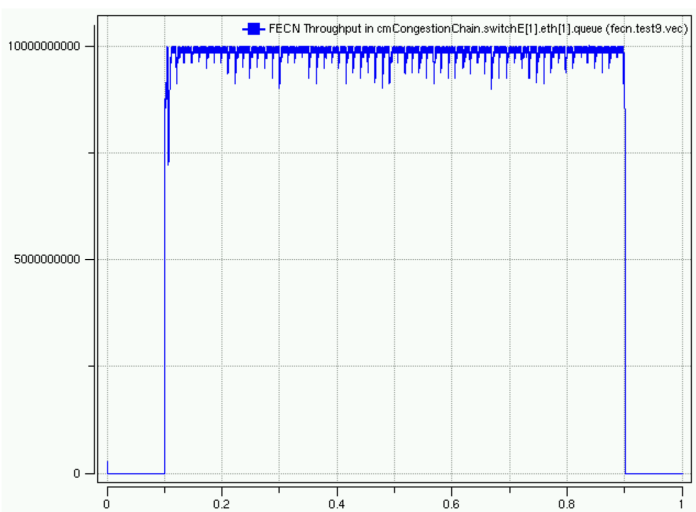
ECM



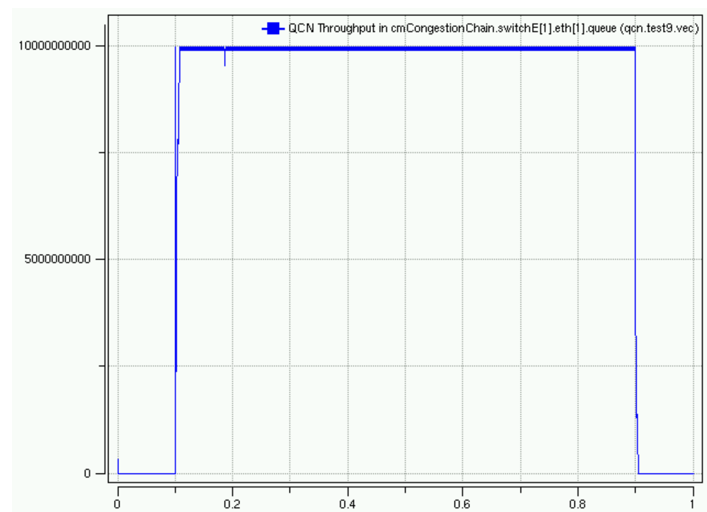
E2CM



FECN

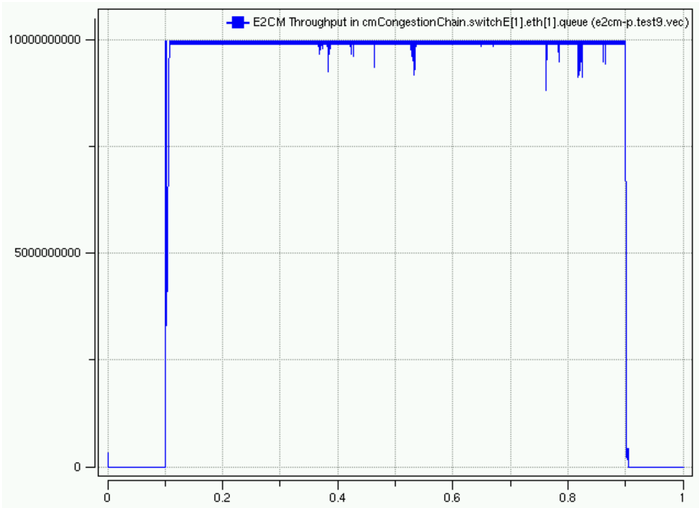


QCN

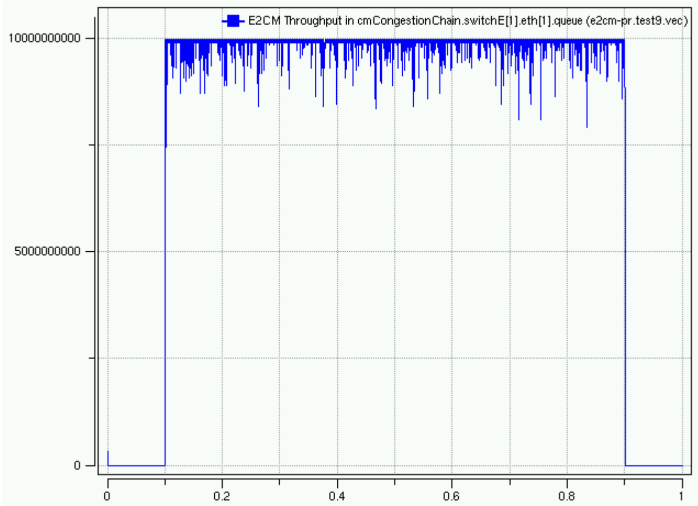


Test 9: Throughput

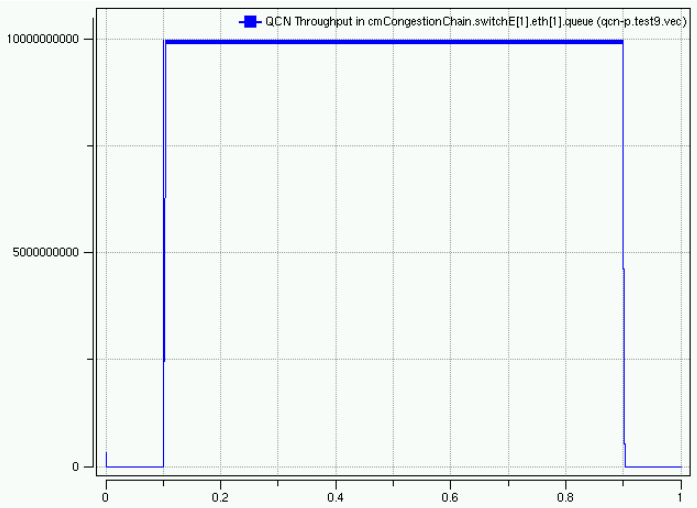
E2CM-P



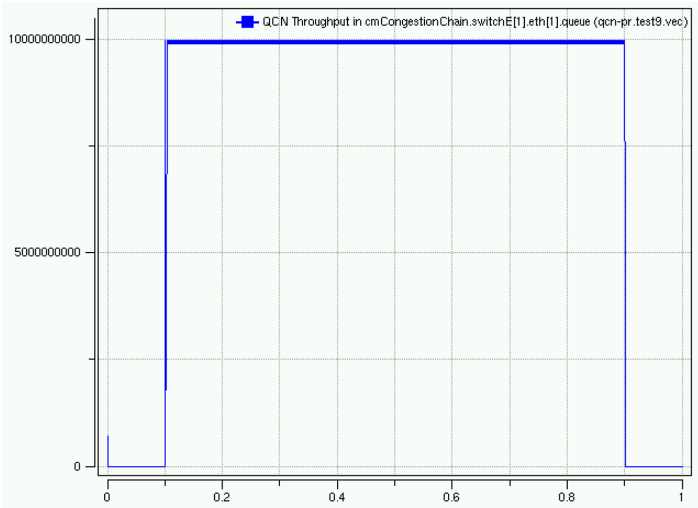
E2CM-PR



QCN-P

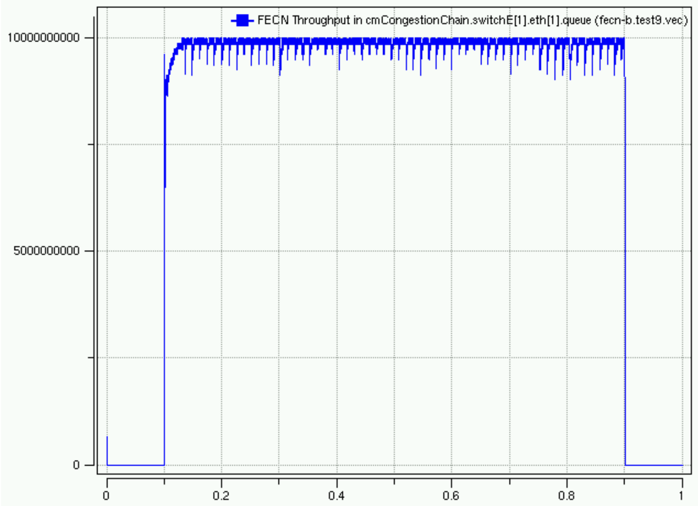


QCN-PR



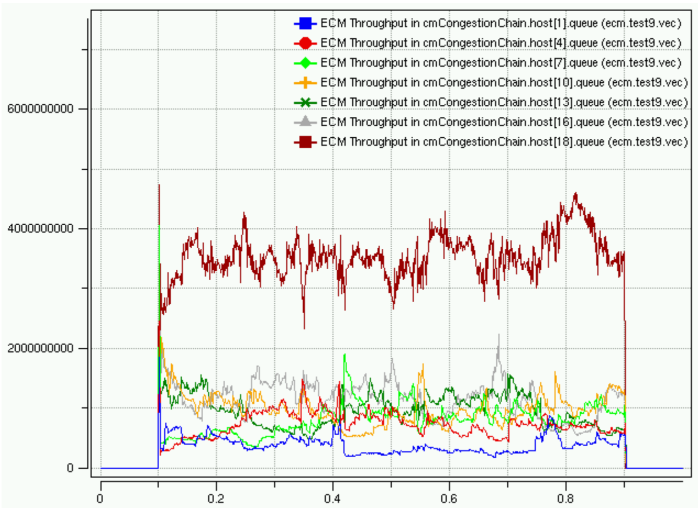
Test 9: Throughput

FECN-B

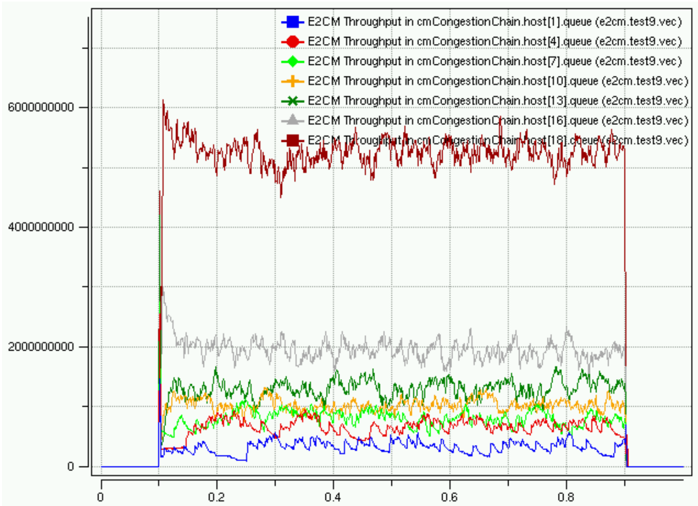


Test 9: Fairness

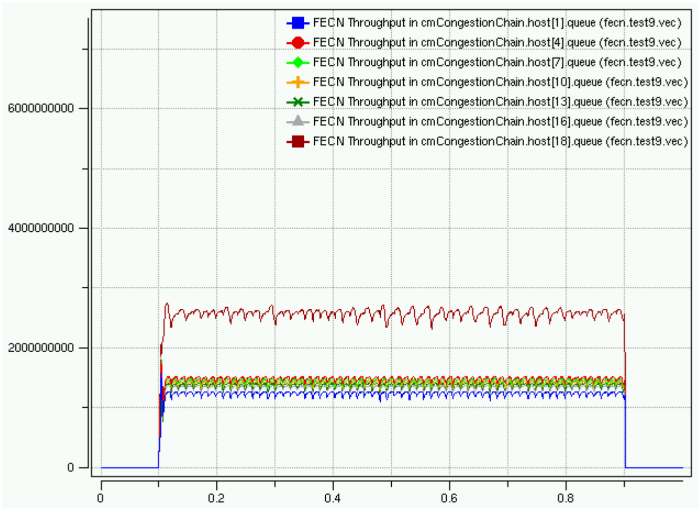
ECM



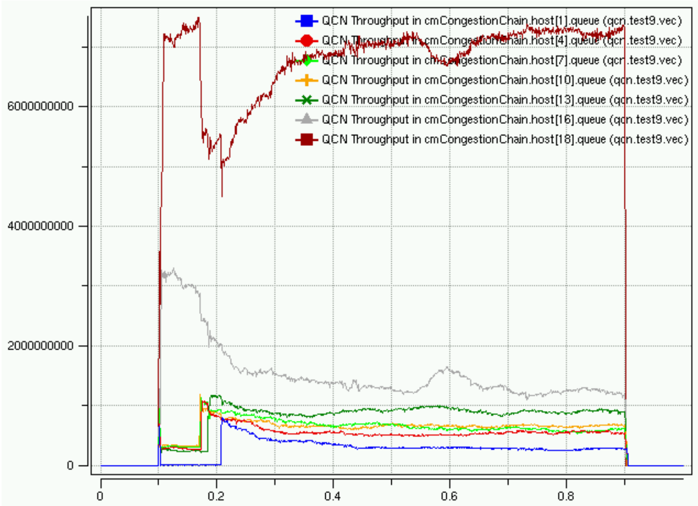
E2CM



FECN

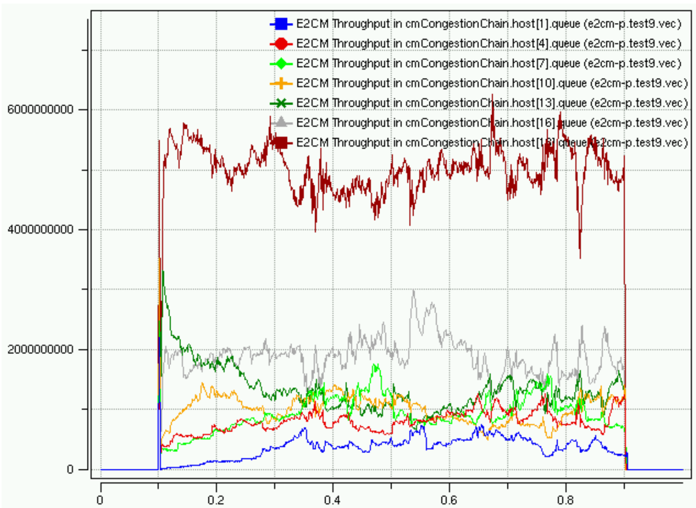


QCN

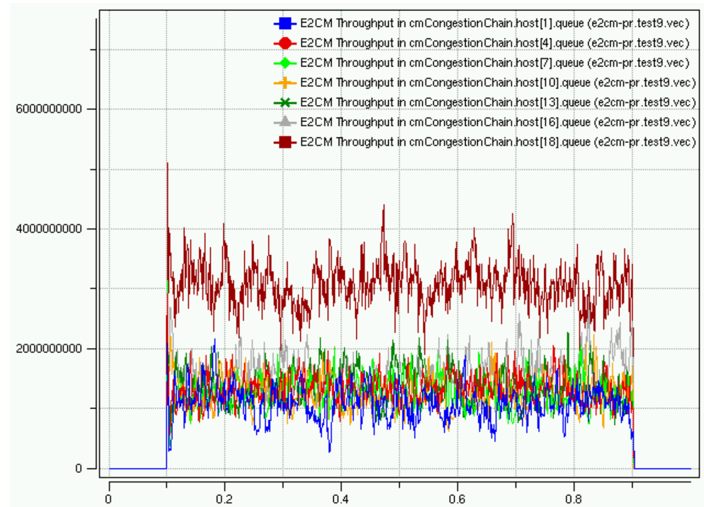


Test 9: Fairness

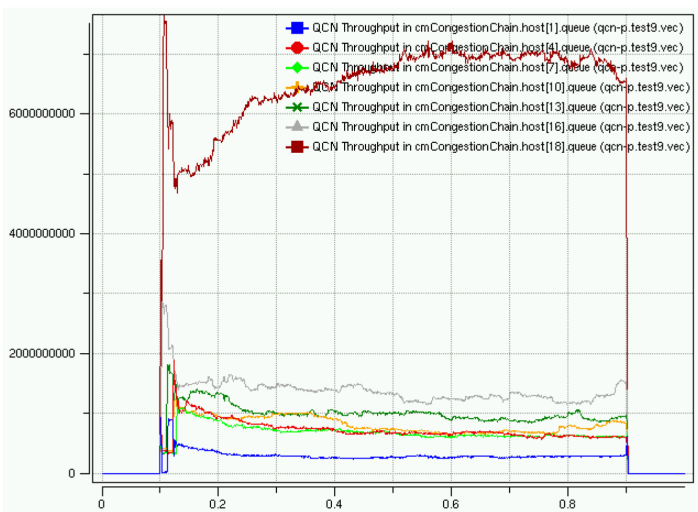
E2CM-P



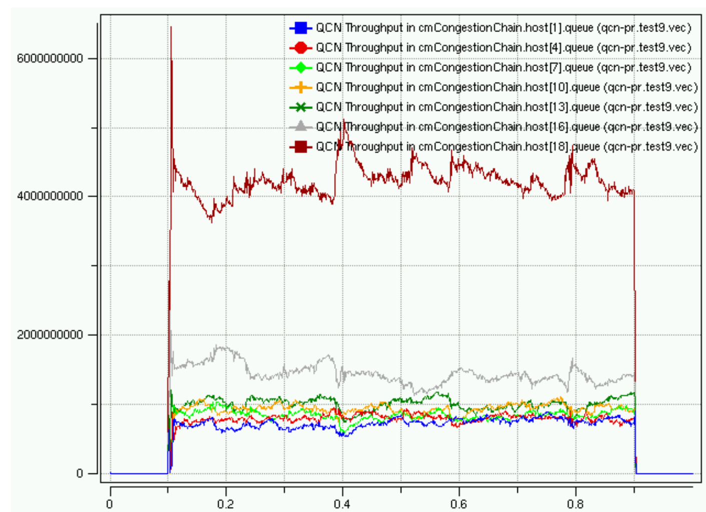
E2CM-PR



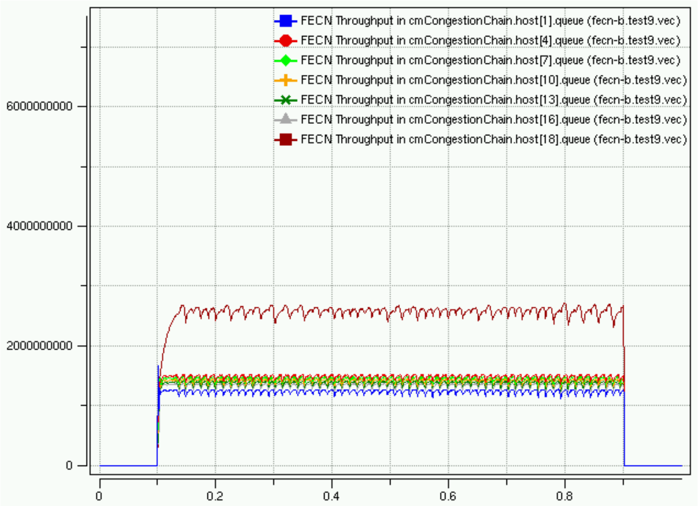
QCN-P



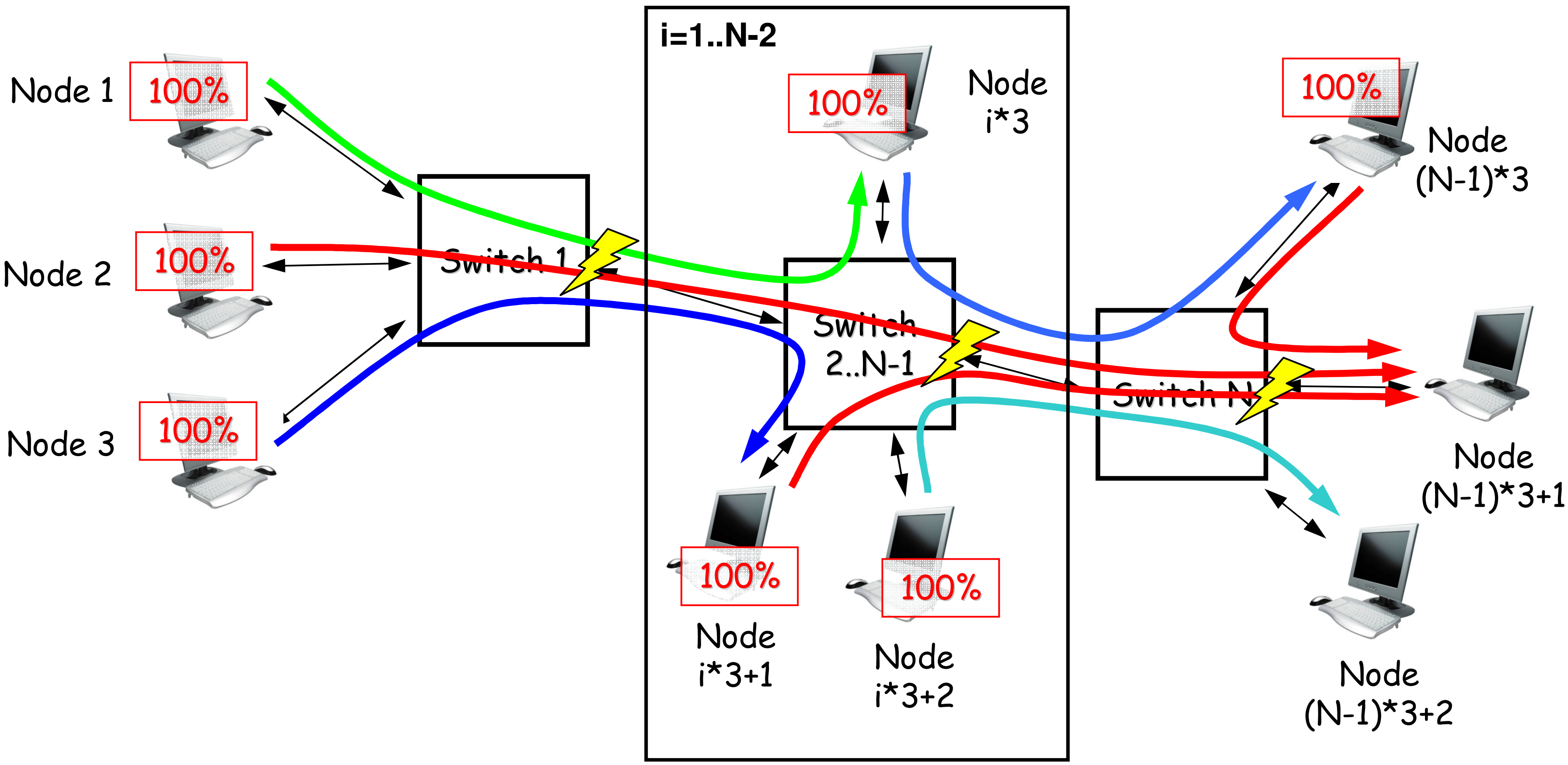
QCN-PR



FECN-B


FECN Throughput in cmCongestionChain.host[1].queue (fecn-b.test9.vec)
FECN Throughput in cmCongestionChain.host[4].queue (fecn-b.test9.vec)
FECN Throughput in cmCongestionChain.host[7].queue (fecn-b.test9.vec)
FECN Throughput in cmCongestionChain.host[10].queue (fecn-b.test9.vec)
FECN Throughput in cmCongestionChain.host[13].queue (fecn-b.test9.vec)
FECN Throughput in cmCongestionChain.host[16].queue (fecn-b.test9.vec)
FECN Throughput in cmCongestionChain.host[18].queue (fecn-b.test9.vec)
6000000000
4000000000
2000000000
0
0
0.2
0.4
0.6
0.8

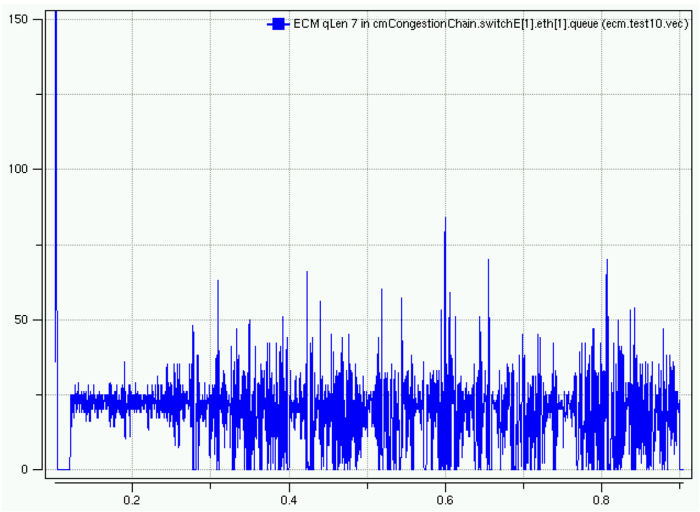
Test 10: 12-stage Hotspot



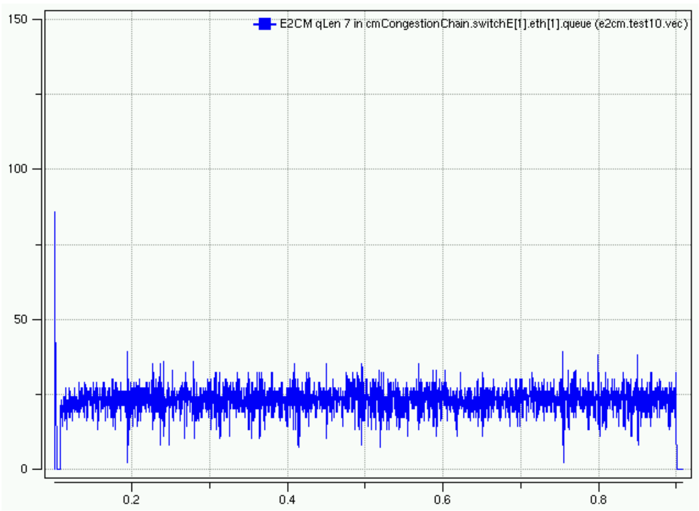
- N=10 switches; 3 hosts per switch
- Node <i> sends to node <i+3>; Node <i+1> sends to node (N-1)*3+1; node <i+2> sends to node <i+4>
- 100% load from all nodes
- Node (N-1)*3+1 receives traffic from <N> sources
- N hotspots

Test 10: Queue Length

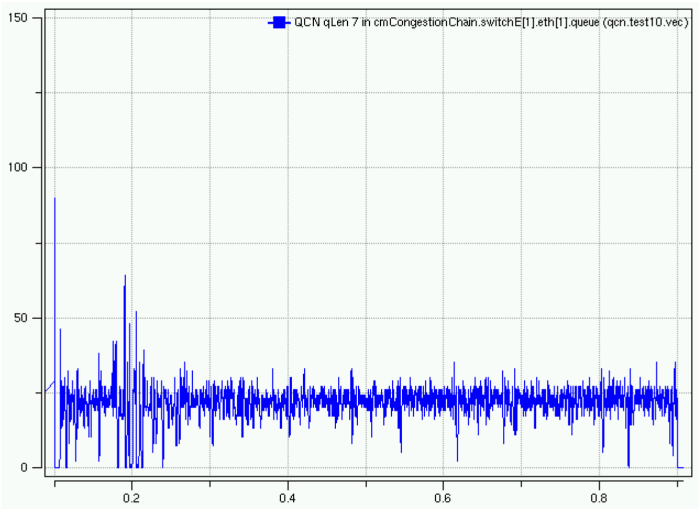
ECM



E2CM

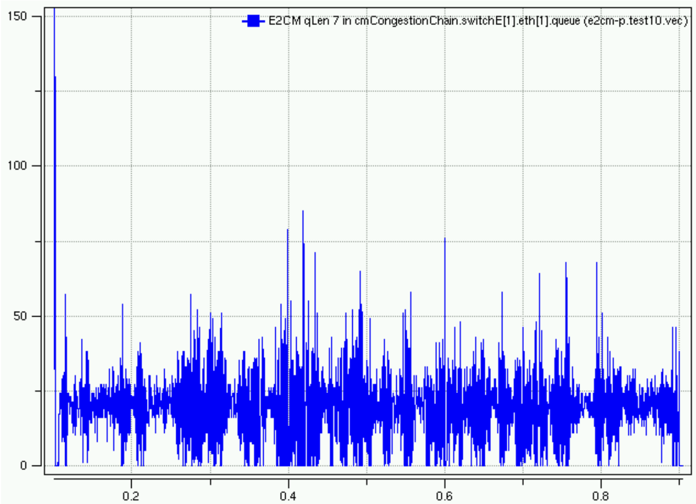


QCN

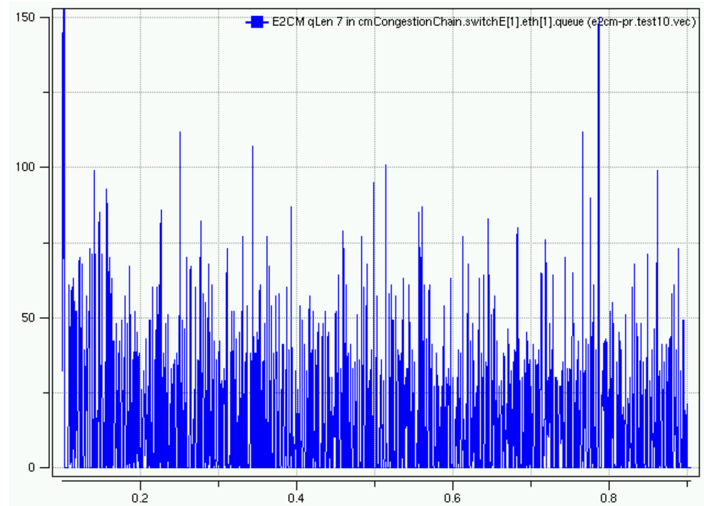


Test 10: Queue Length

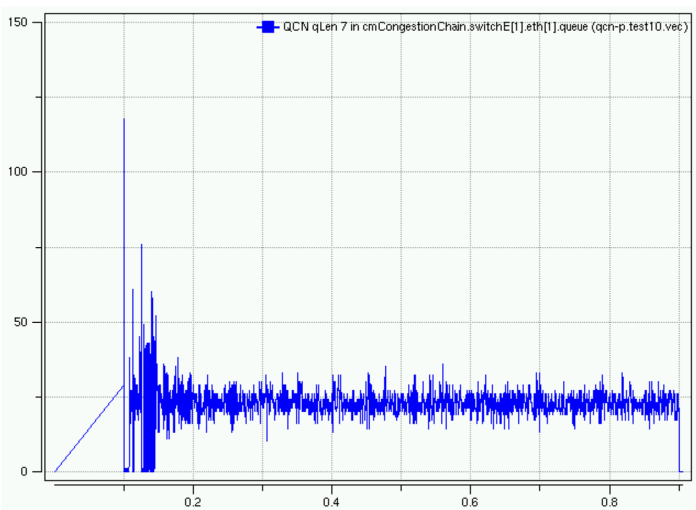
E2CM-P



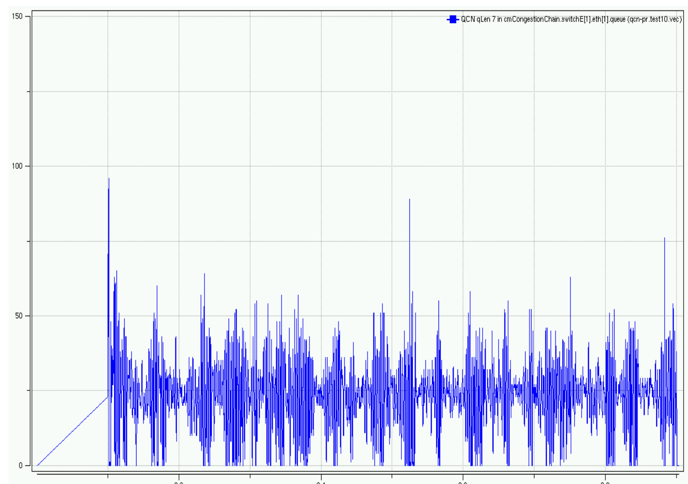
E2CM-PR



QCN-P

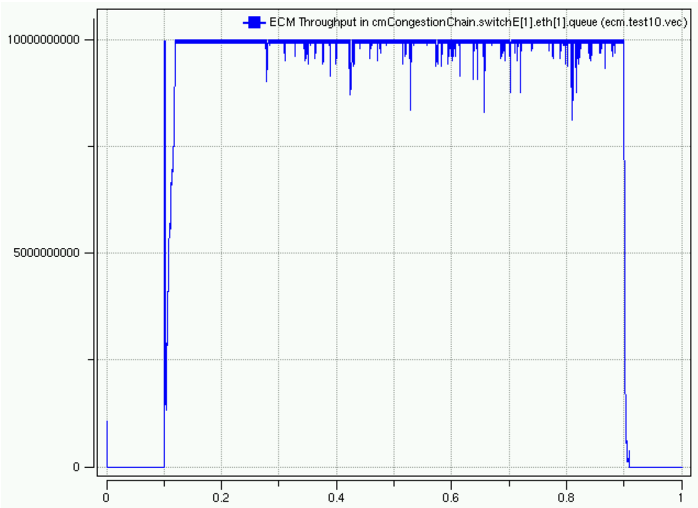


QCN-PR

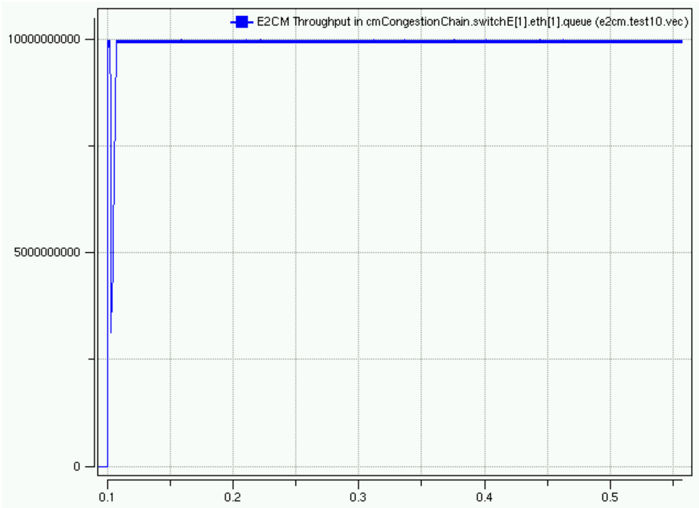


Test 10: Throughput

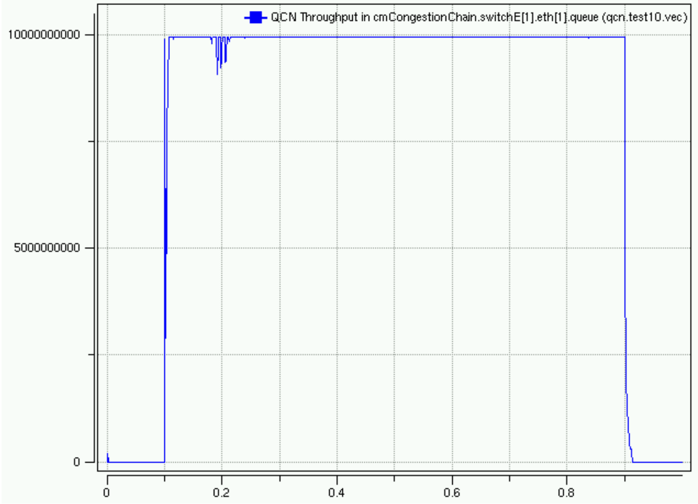
ECM



E2CM

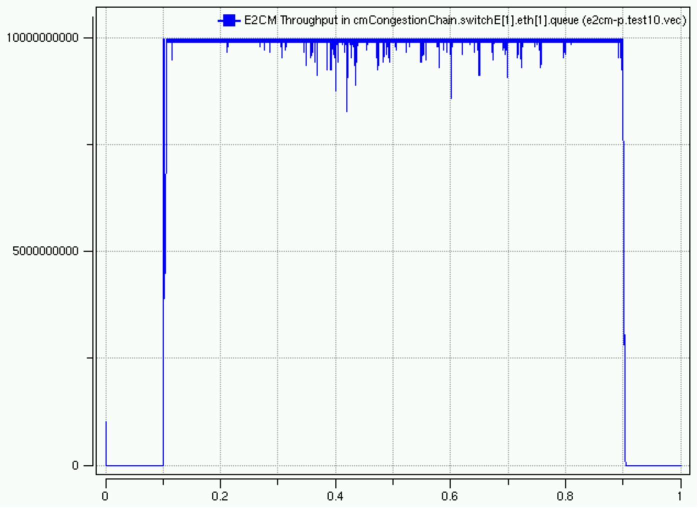


QCN

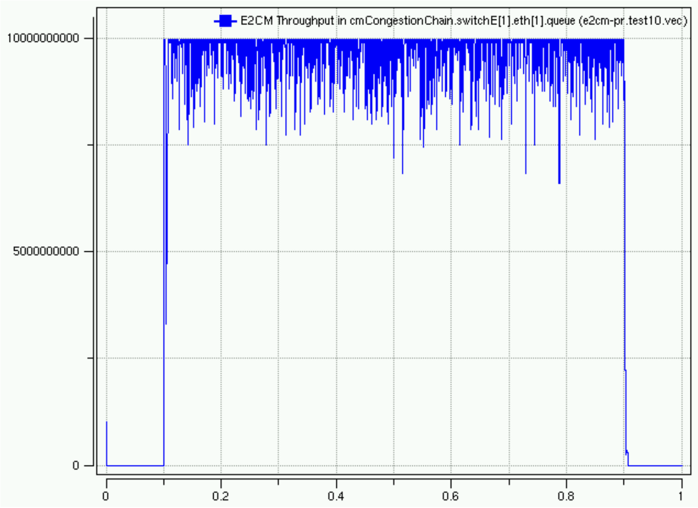


Test 10: Throughput

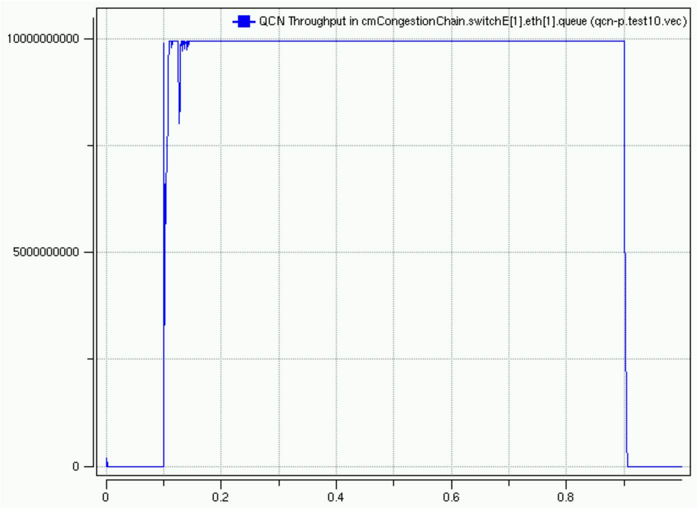
E2CM-P



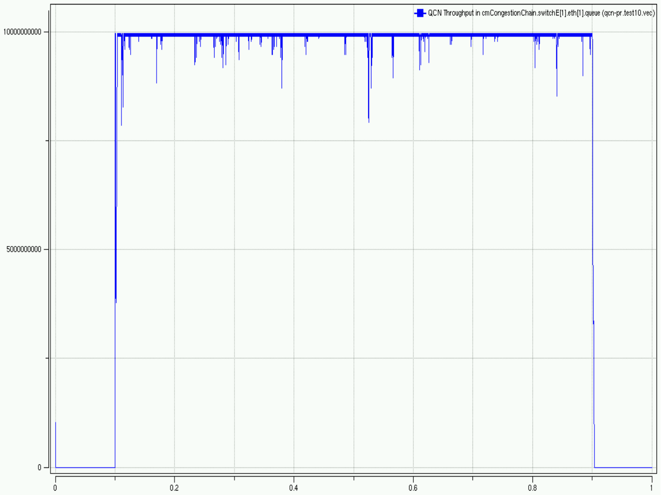
E2CM-PR



QCN-P

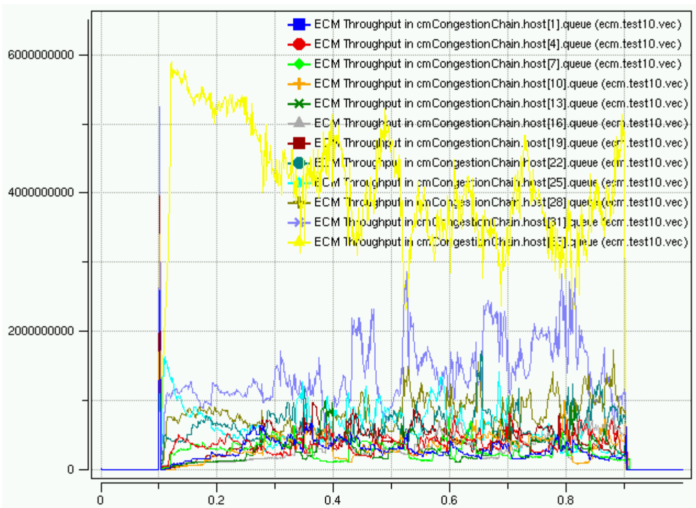


QCN-PR

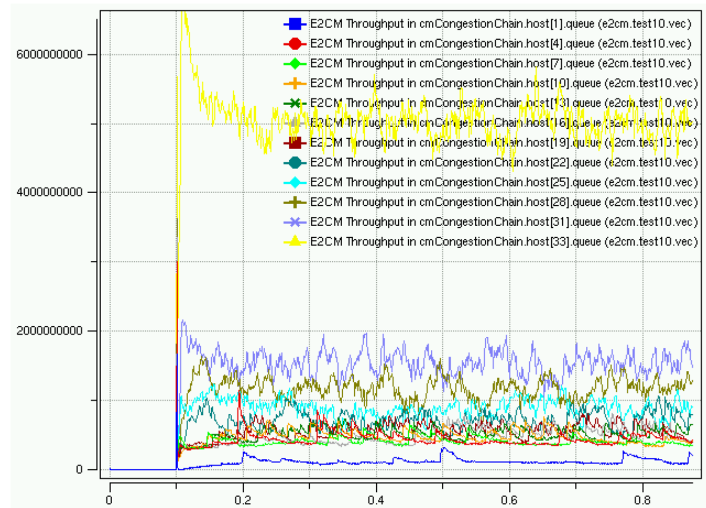


Test 10: Fairness

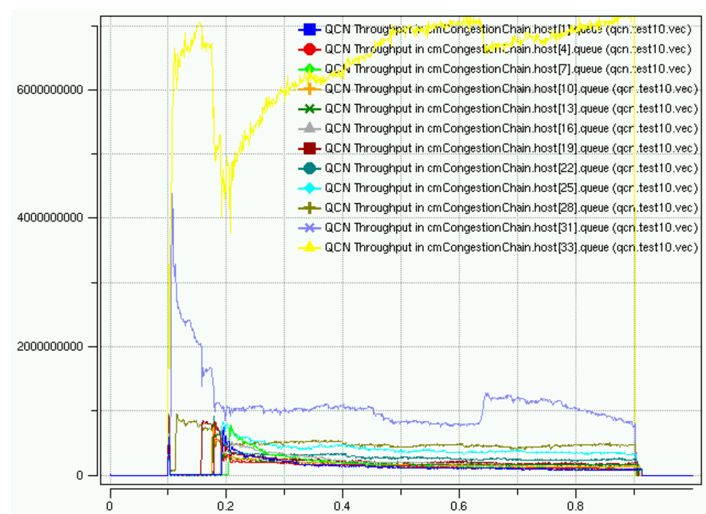
ECM



E2CM

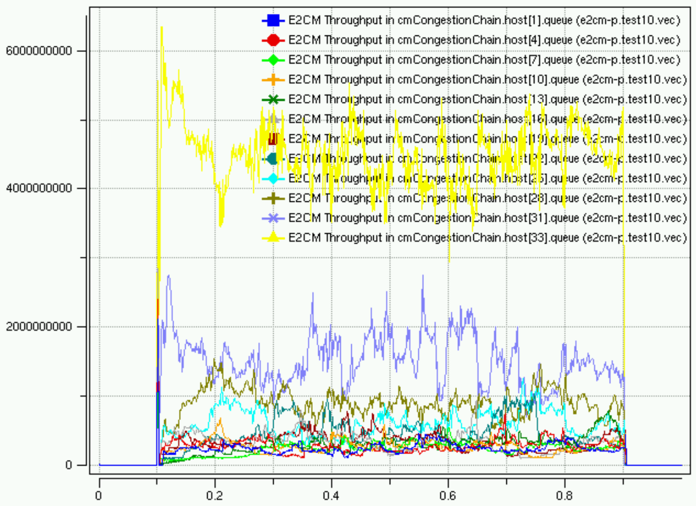


QCN

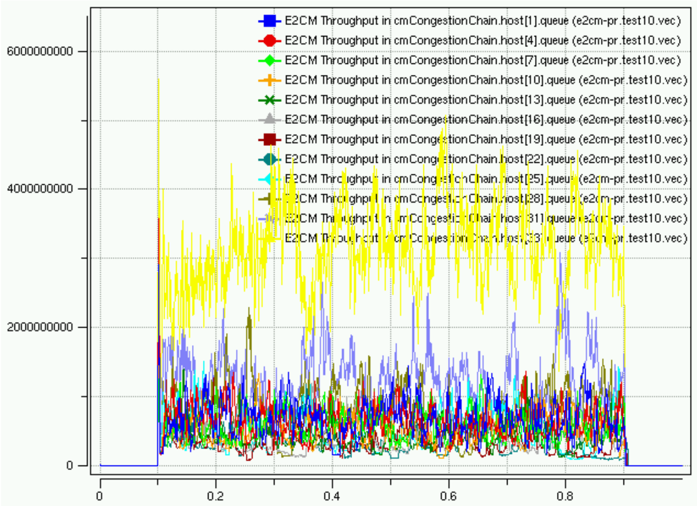


Test 10: Fairness

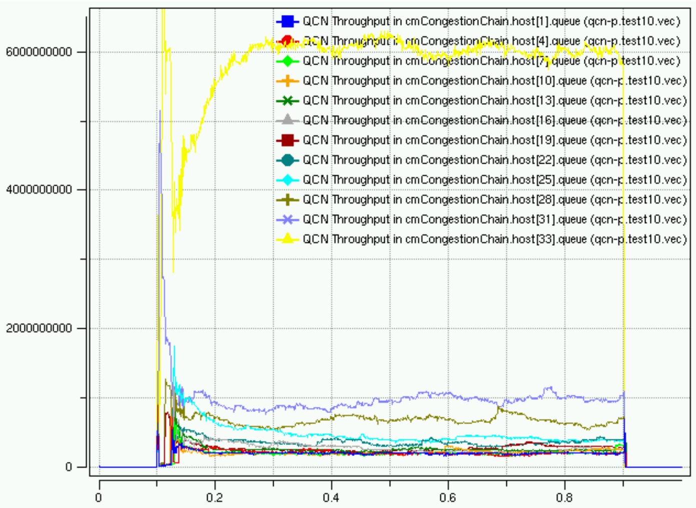
E2CM-P



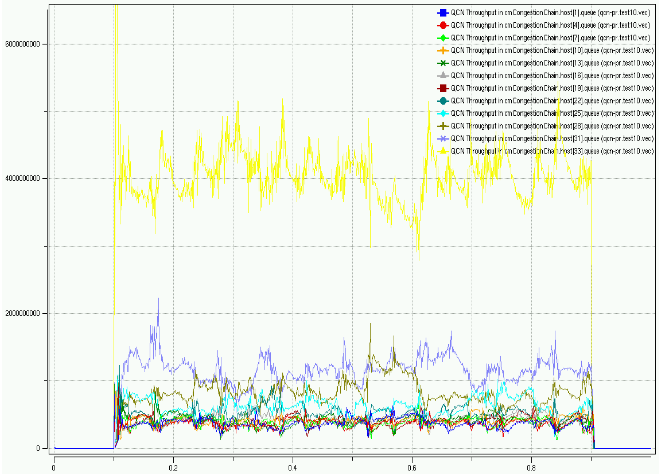
E2CM-PR



QCN-P



QCN-PR



Summary

- ECM
  - Good performance over wide range of conditions
  - Visible throughput impact due to tagging
  - Some oscillation with complex topologies
  - Marginal fairness
- E2CM
  - Good fairness; favors short-distance flows over long distance flows
  - Seems to have problems with output generated hotspots
- FECN
  - Excellent fairness
  - Slow reaction to changed conditions
  - Problems with output generated hotspots and with complex topologies
  - Oscillation with multiple hotspots
- QCN
  - Fast reaction to load increases
  - Slow reaction to load decreases
  - Marginal fairness

Summary – continued

- E2CM-P
  - Pretty much equivalent to ECM
- E2CM-PR
  - Very good fairness
  - Oscillation in complex topologies w/ multiple hotspots
- QCN-P
  - Overall best performance
  - Marginal fairness
- QCN-PR
  - Good fairness
  - Starts oscillating in topology with 12 hotspots

- QCN style ECM packet generation (flexible probability) improves reaction time
  - Might be worthwhile testing it with ECM/E2CM
- Fairness
  - Linear self-increase improves fairness over multiplicative self-increase

    rate += selfIncrease;
  - Even better is proportional self-increase towards maximum rate provided by congested switch

    rate += (switchMaxRate – rate) * selfIncreaseFactor;
  - Rate guidance from switch improves fairness
    - May cause oscillations

Observations - continued

- Avoid negative feedback to probes sent to CP
  - Causes oscillations
- Oscillations observed with pretty much all protocols
  - Especially in topologies with multiple hotspots

To Do

- Test ECM, E2CM etc with flexible ECM rate (QCN style)
- Verify if FECN and E2CM problems with output generated hotspots are caused by the simulation or a real problem
- Verify if observed oscillations are caused by the simulation or a real problem

Conclusions

- Feedback through Endpoint is not a requirement
  - RP ⇔ CP protocol exchange is sufficient
- Tagging is not mandatory for any protocol
  - Can use probes from RP to CP instead
- RP ⇔ CP feedback highly recommended for positive feedback
- Rate guidance feedback helps to achieve fairness

How to proceed

- Use BCN message format for negative feedback
- Use Probes between RP and CP for positive feedback
- Consider adding Bandwidth guideline parameter to information sent from CP to RP
- Also consider including Min/Max rates to allow for more flexible feedback
  - Example: Max rate = CP link speed