ECM and E^2CM performance under bursty traffic

Cyriel Minkenberg & Mitch Gusat
IBM Research GmbH, Zurich
April 26, 2007

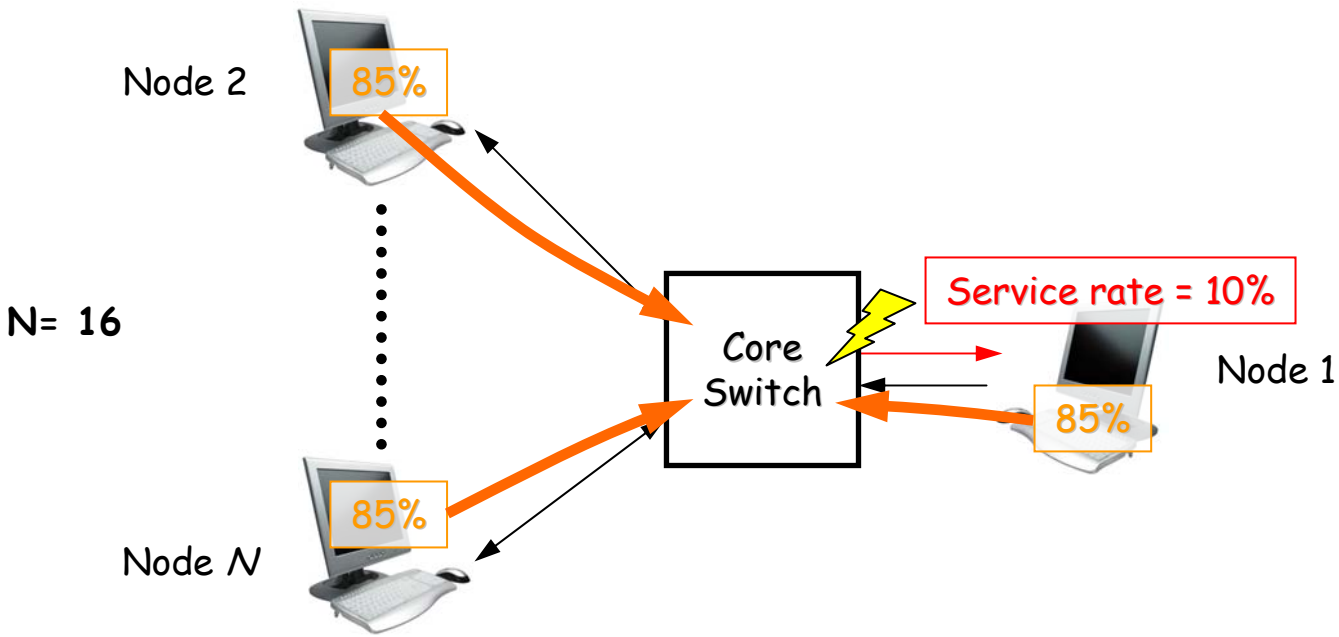
Target

- Study Output-Generated (OG) single hop congestion with bursty injection processes

Conditions, parameters, simulation environment

- Traffic
  - Non-Pareto temporal source injection burstiness:
    - i.i.d. bursty arrivals
    - **geometrically** distributed burst size around mean **B = [1.2, 12, 48, 120]** us
- LL-FC: runs with and w/o PAUSE
- CM: none, ECM and E^2CM
- Metrics: TP_{aggr}, TP_{hot}, Q_{hot}, frame drops
    - for details see the "fine print" page

Output-Generated Single-Hop Hotspot



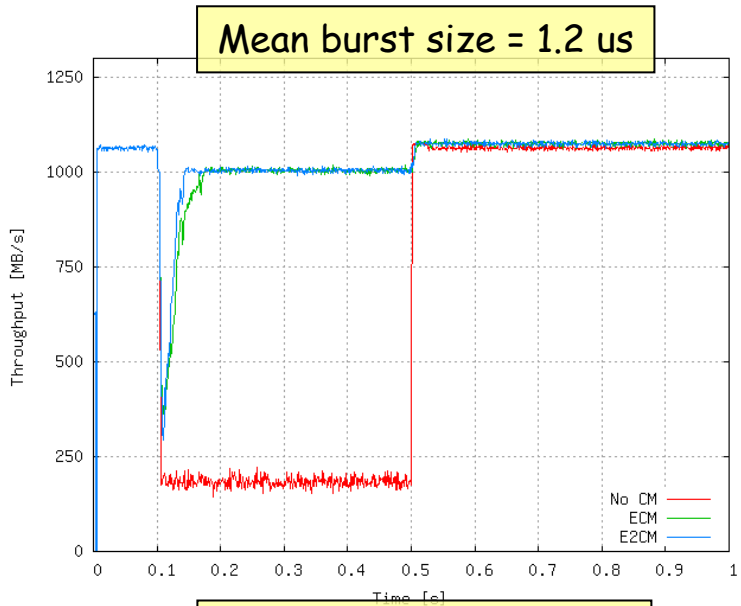
- All nodes: Uniform destination distribution, load = 85% (8.5 Gb/s)
- Node 1 service rate = 10%
- One congestion point
  - Hotspot degree = N-1
  - All flows affected

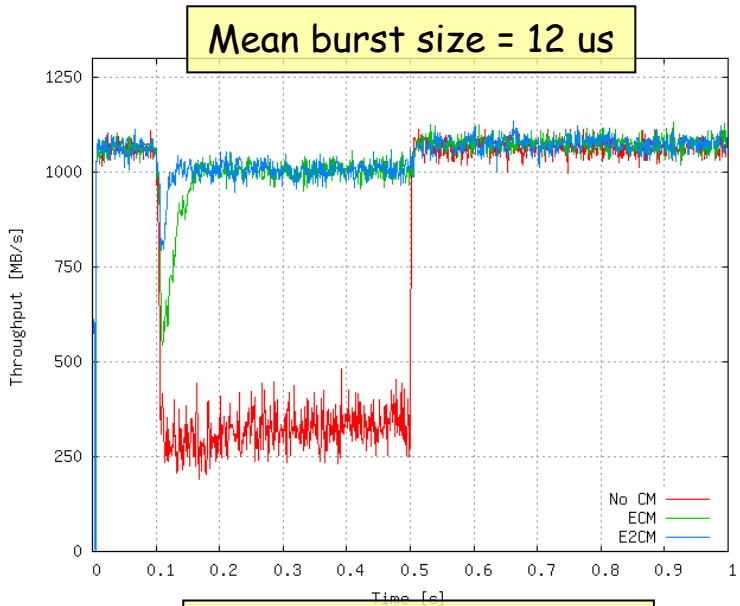
Simulation Setup & Parameters

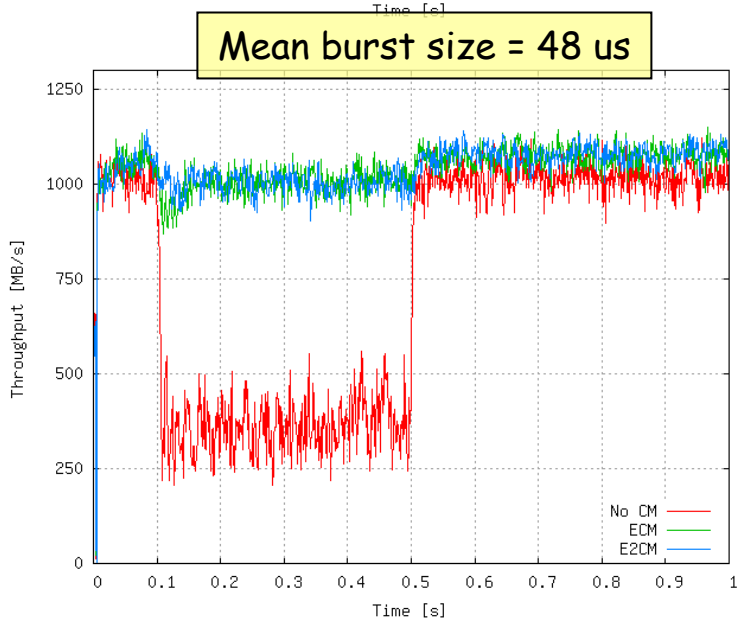
The "fine print"

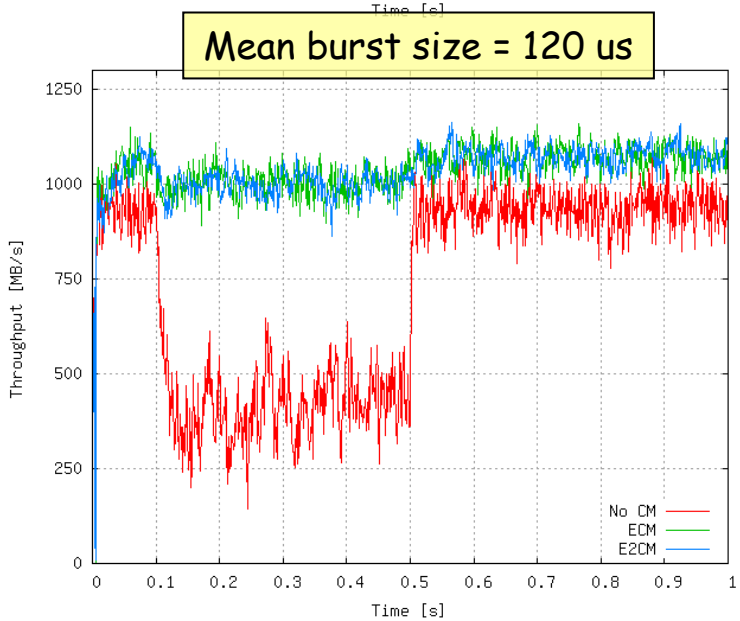
- Traffic
  - I.i.d. Bursty arrivals, geometrically distributed burst size around mean *B*
  - **B = [1.2, 12, 48, 120] us**
  - Uniform destination distribution (to all nodes except self)
  - Fixed frame size = 1500 B
- Scenario
  1. Single-hop output-generated hotspot
- Switch
  - N = 16
  - M = 300 KB/port
  - Partitioned memory per input, shared among all outputs
  - No limit on per-output memory usage
  - PAUSE enabled or disabled
    - Applied on a per input basis based on local high/low watermarks
    - $watermark_{high}$ = 260 KB
    - $watermark_{low}$ = 230 KB
    - If disabled, frames dropped when input partition full
- Adapter
  - Per-node virtual output queuing, round-robin scheduling
  - No limit on number of rate limiters
  - Ingress buffer size = 1500 KB, partitioned across VOQs, per-flow selective source quench used when VOQ full, round-robin VOQ service
  - Egress buffer size = 150 KB
  - PAUSE enabled
    - $watermark_{high}$ = 150 – rtt*bw KB
    - $watermark_{low}$ = $watermark_{high}$ - 10 KB
- ECM
  - W = 2.0
  - Q_{eq} = 75 KB (= M/4)
  - G_d = 0.5 / ((2*W+1)*Q_{eq})
  - G_{i0} = (R_{link} / R_{unit}) * ((2*W+1)*Q_{eq})
  - G_i = 0.1 * G_{i0}
  - P_{sample} = 2% (on average 1 sample every 75 KB
  - R_{unit} = R_{min} = 1 Mb/s
  - BCN_MAX enabled, threshold = 260 KB
  - No BCN(0,0), no self-increase
- E^2CM (per-flow)
  - W = 2.0
  - $Q_{eq,flow}$ = 15 KB
  - $G_{d,\ flow}$ = 0.5 / ((2*W+1)*$Q_{eq,flow}$)
  - $G_{i,\ flow}$ = 0.005 * (R_{link} / R_{unit}) / ((2*W+1)*$Q_{eq,flow}$)
  - P_{sample} = 2% (on average 1 sample every 75 KB)
  - R_{unit} = R_{min} = 1 Mb/s
  - BCN_MAX enabled, threshold = 52 KB

Aggregate throughput - PAUSE disabled

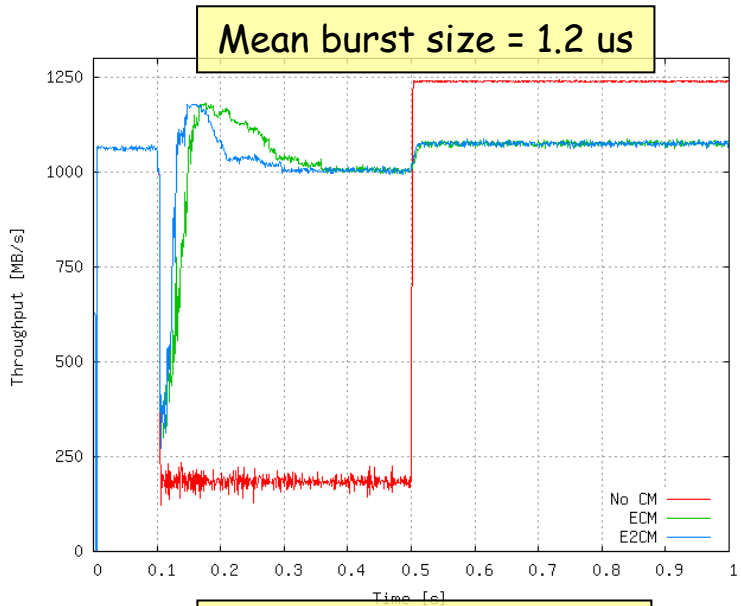


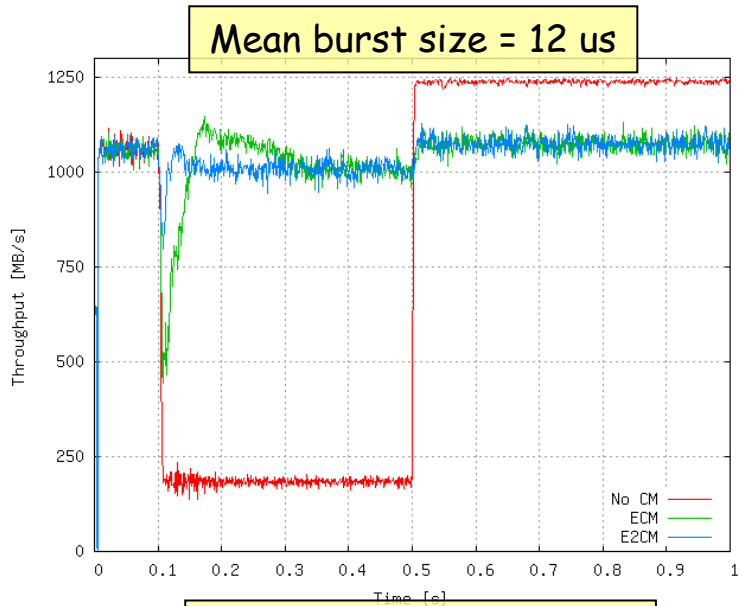


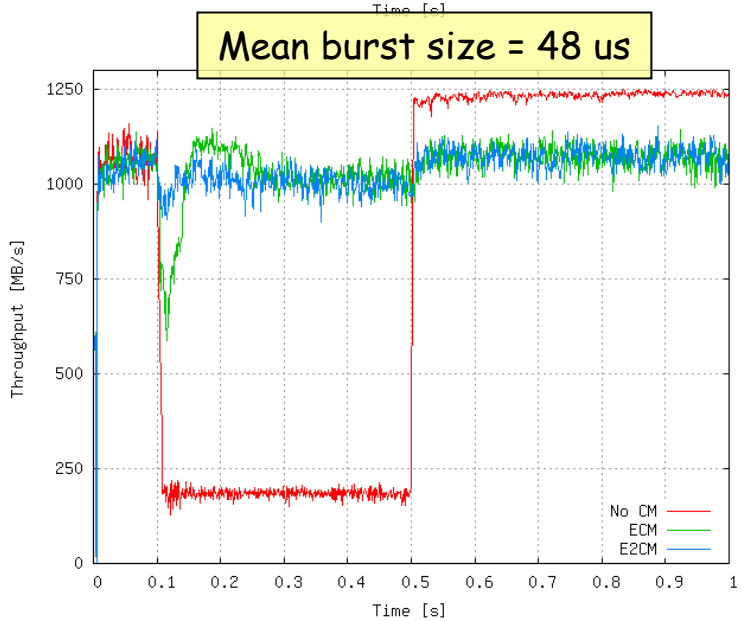


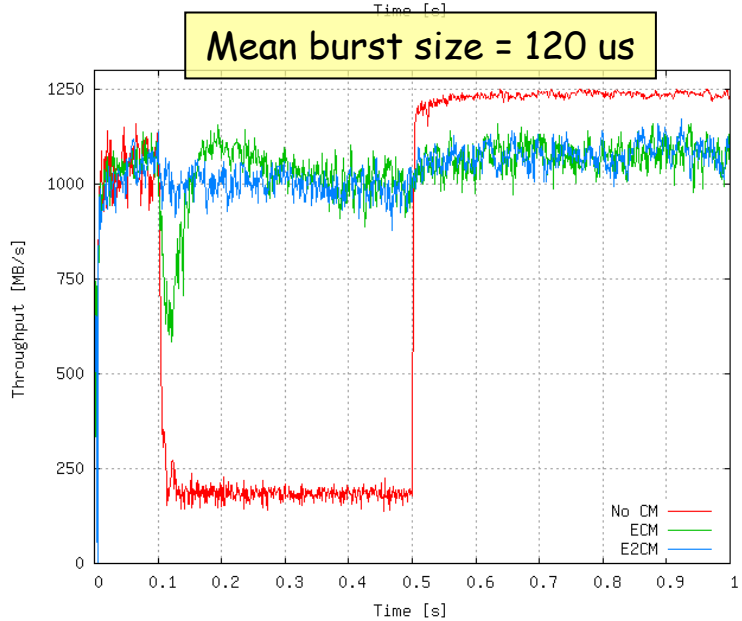


Aggregate throughput – PAUSE enabled

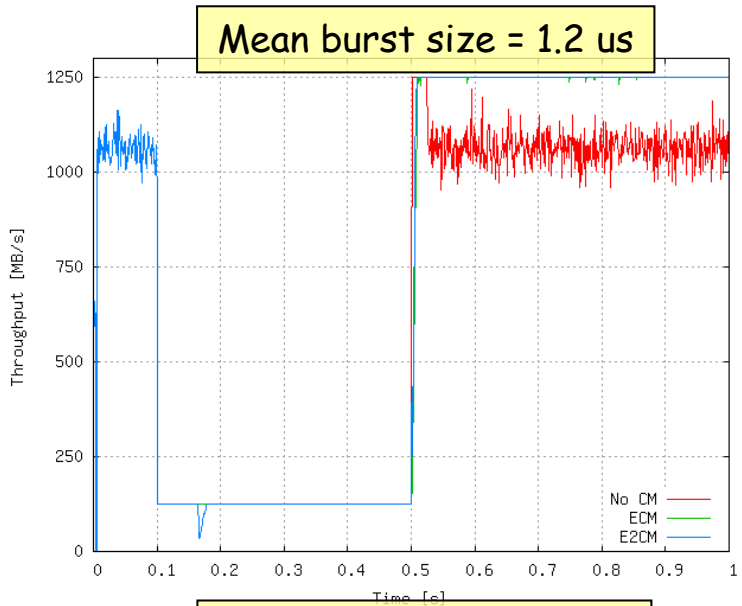


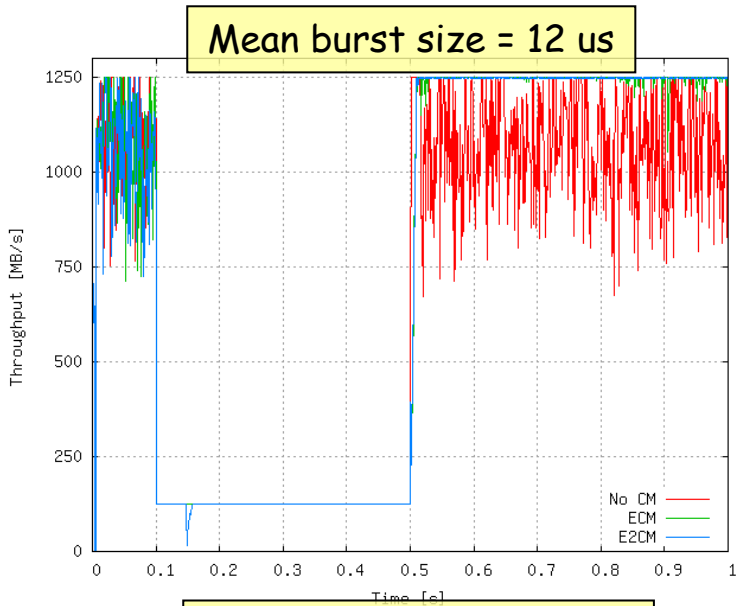


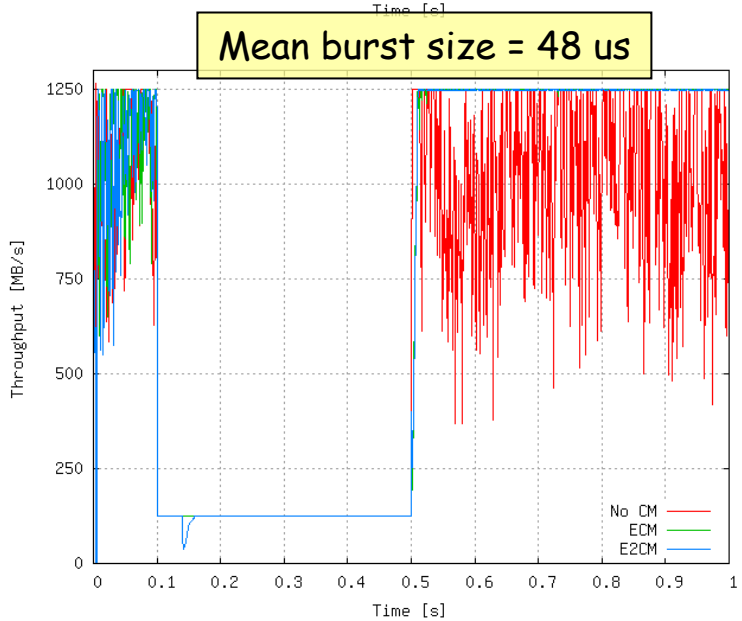


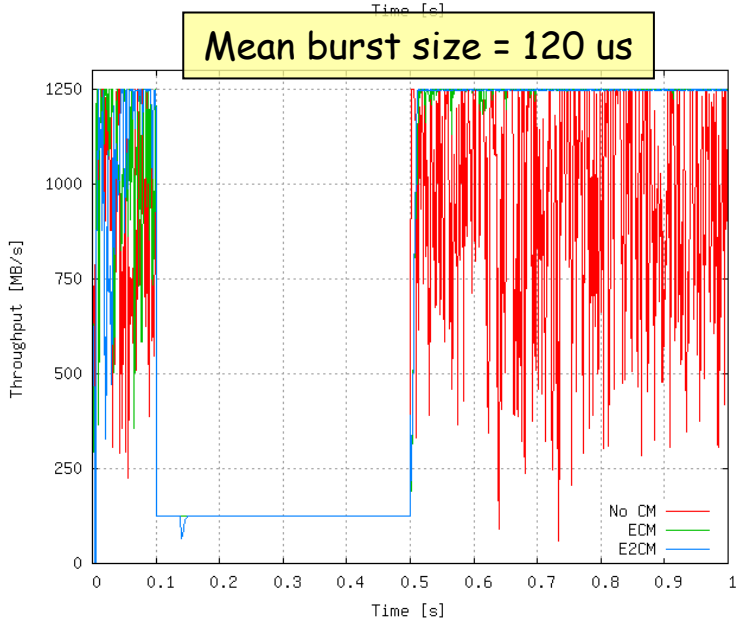


Hot port throughput - PAUSE disabled

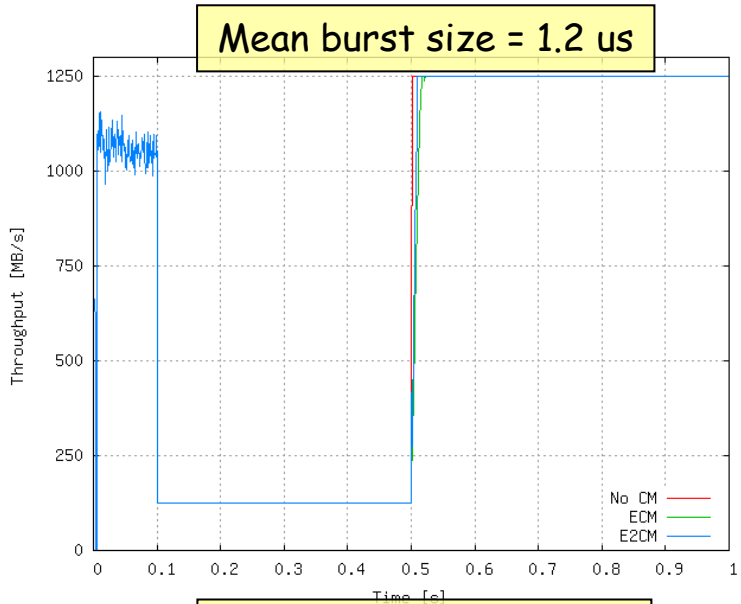


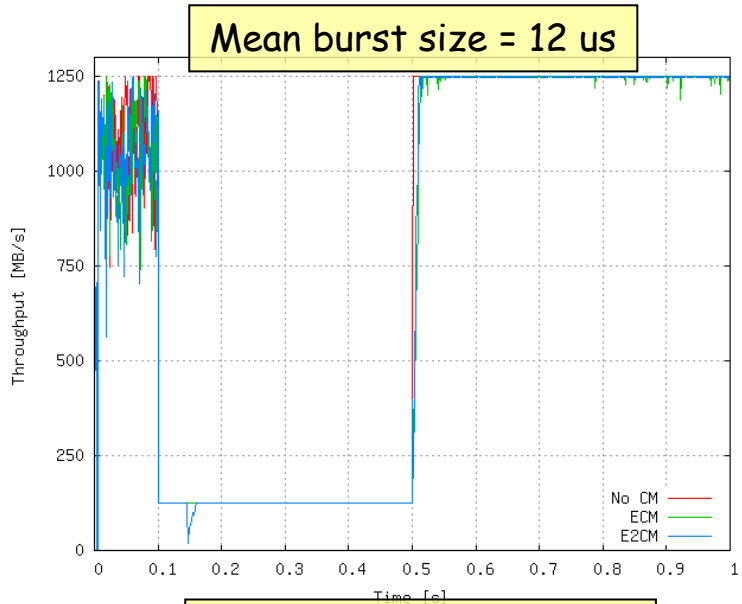


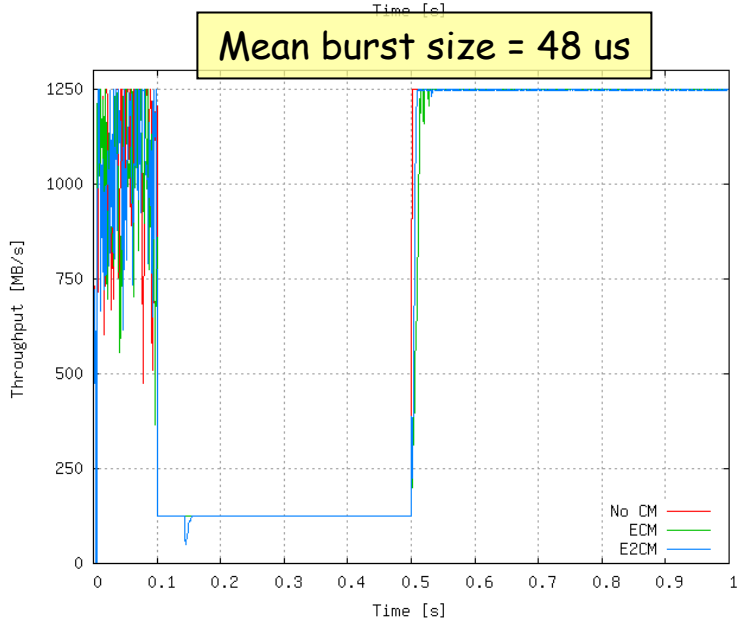


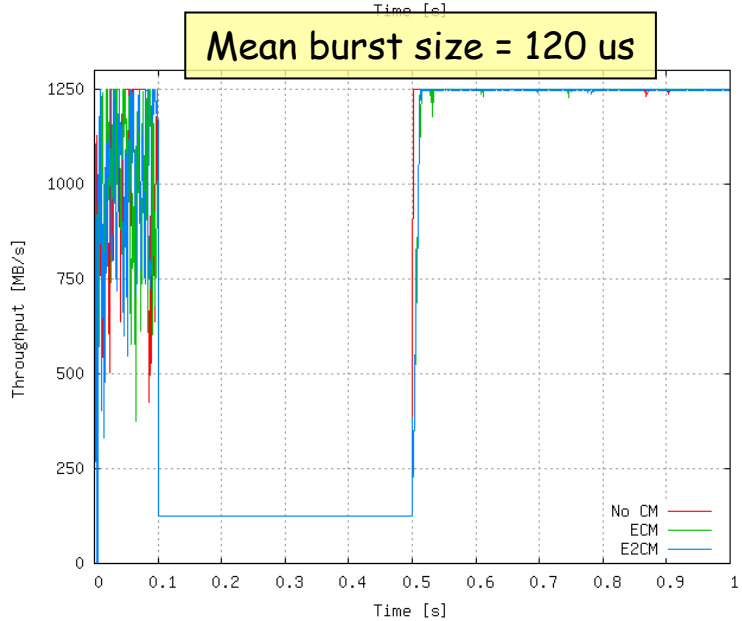


Hot port throughput – PAUSE enabled

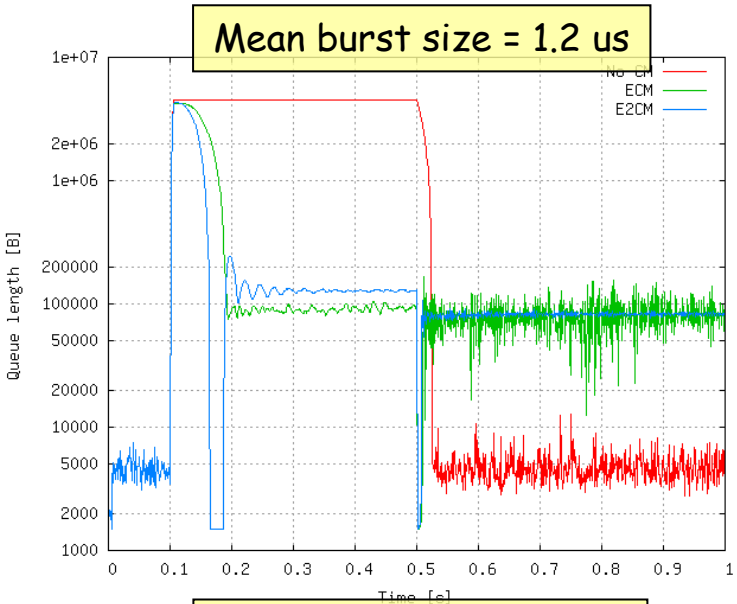


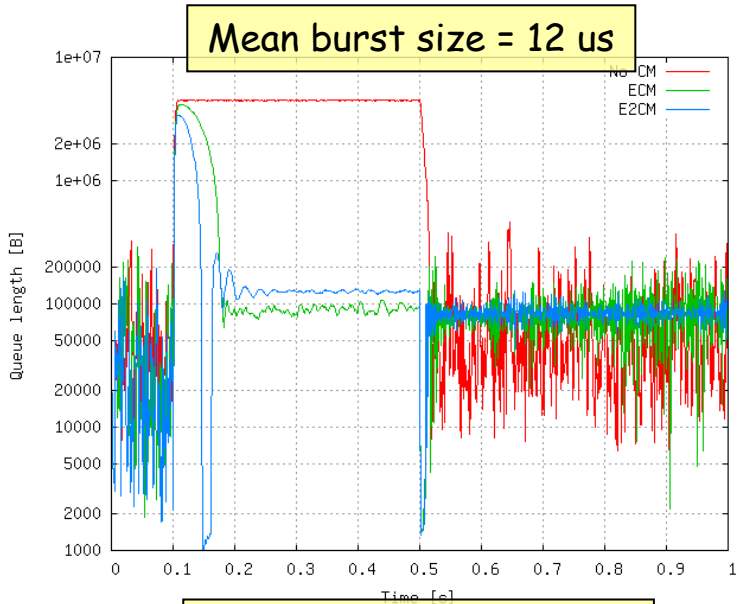


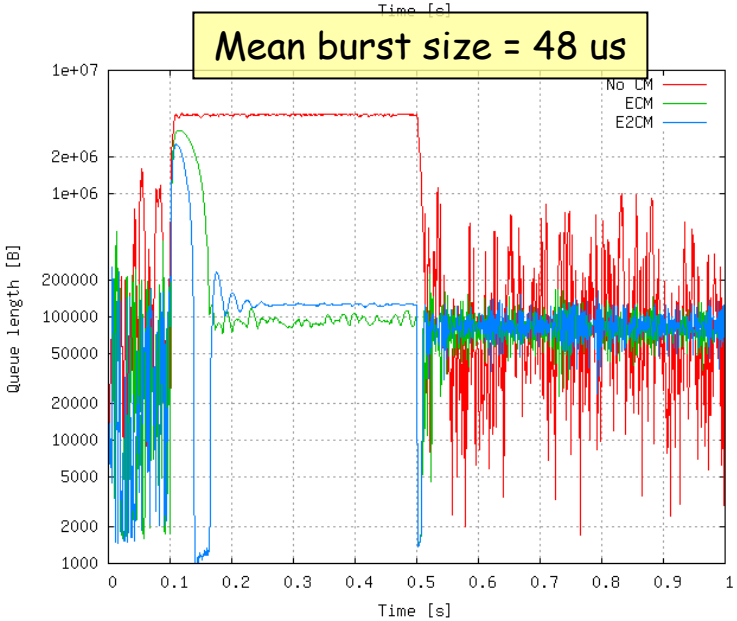


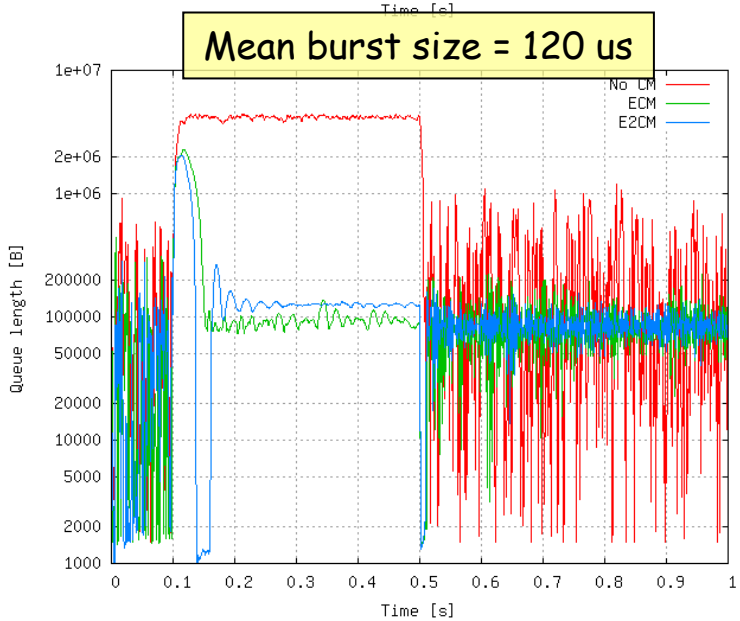


Hot queue length - PAUSE disabled

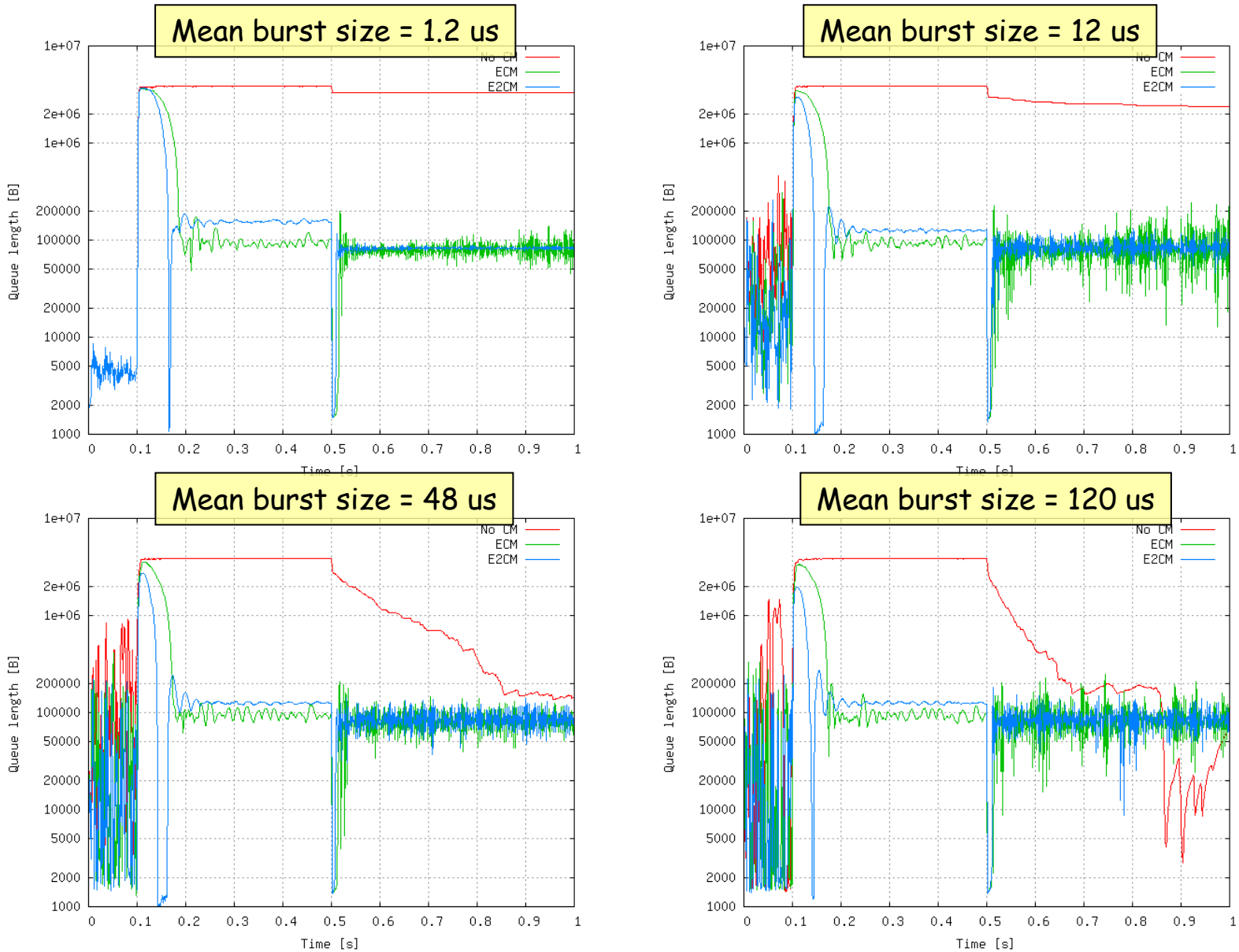




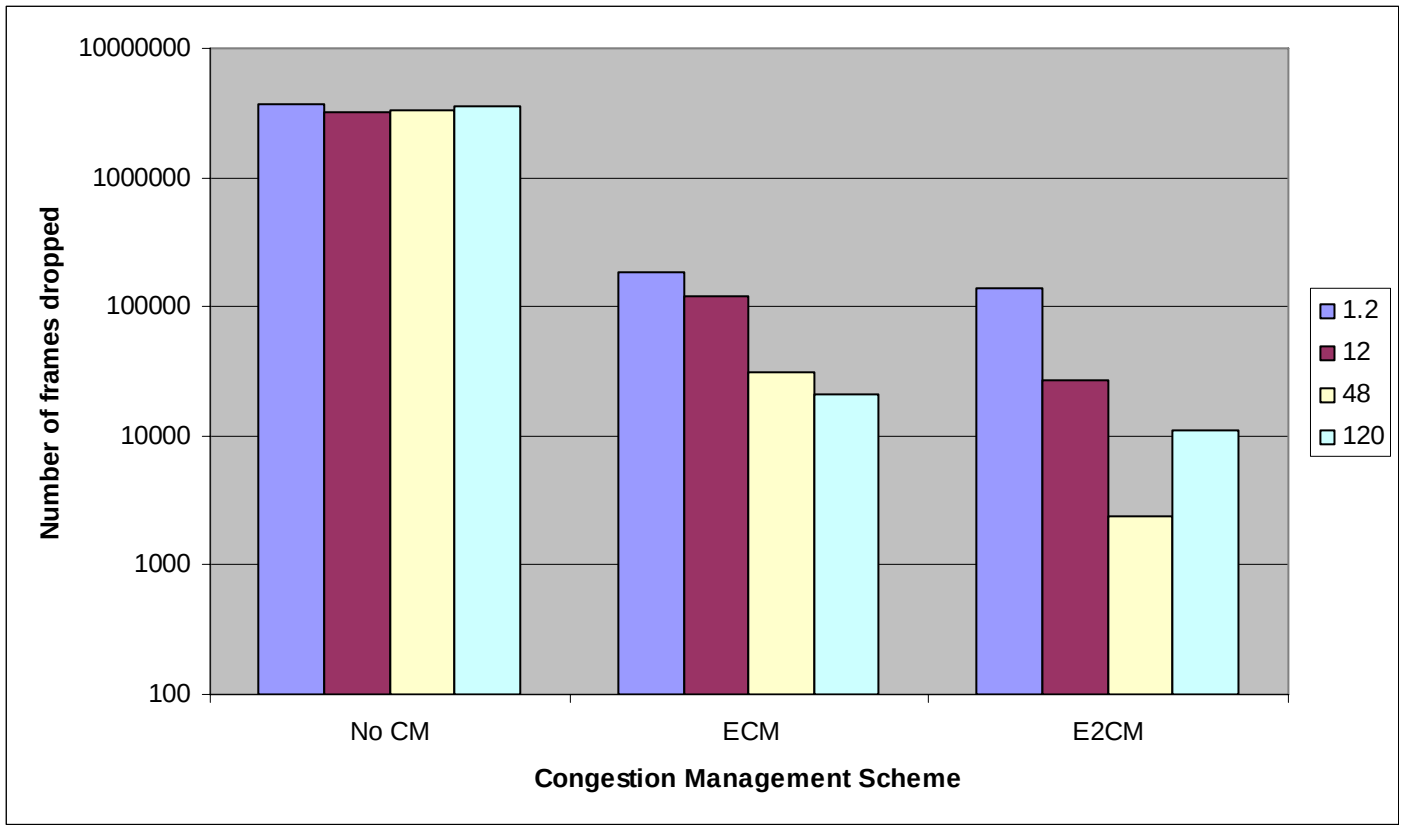




Hot queue length – PAUSE enabled



Frame drops (PAUSE disabled)



Conclusions to Bursty OG

- For high burstiness CM improves aggregate throughput even w/o hotspot (no PAUSE)
- Difficulty (of control) is proportional to 1/B
  - As mean burst size increases
    - Aggregate throughput recovers faster
    - Queue stabilizes more quickly (1st overshoot)
    - Frame drops are fewer (w/o PAUSE)
      - except a sweet-spot anomaly at b=48 for E2CM
- Future work: FCT metric
  - Not trivial to generate standard workload and use standard measurements...
  - Using trace-based simulation?

ECM and E^{2}CM performance in large switch configurations

Single-Hop High Degree Hotspot

Cyriel Minkenberg & Mitch Gusat
IBM Research GmbH, Zurich
April 26, 2007

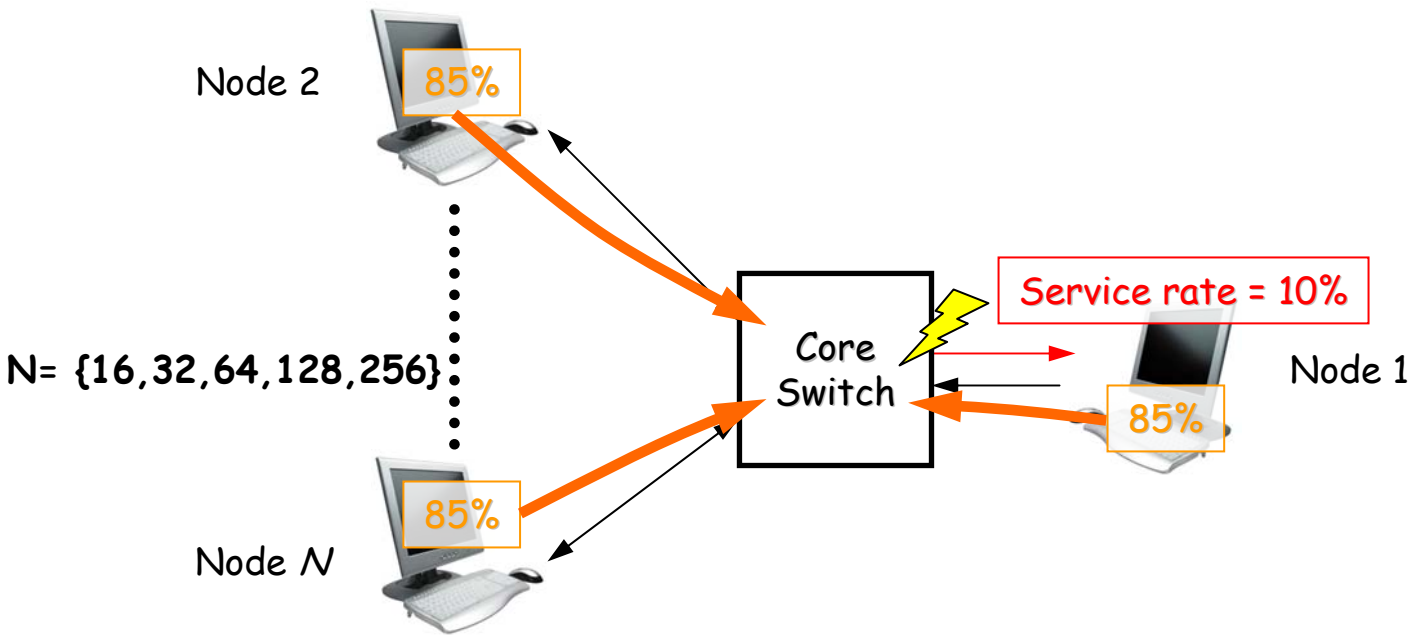
Targets

1. Study Output-Generated (OG) single-hop scenario with **high hotspot degree** (HSD) congestion
2. First look at E²CM with continuous probing (Pat's suggestion in sim adhoc call April 12th)

Conditions, parameters, simulation environment

- Traffic
  - i.i.d. Bernoulli arrivals
- LL-FC: runs with and w/o PAUSE
- CM: No CM, ECM, E²CM, E²CM-CP
- Metrics: TP_{aggr}, TP_{hot}, Q_{hot}, frame drops
  - for details see the "fine print" page

Output-Generated Single-Hop High HSD



- All nodes: Uniform destination distribution, load = 85% (8.5 Gb/s)
- Node 1 service rate = 10%
- One congestion point
  - Hotspot degree = N-1
  - All flows affected

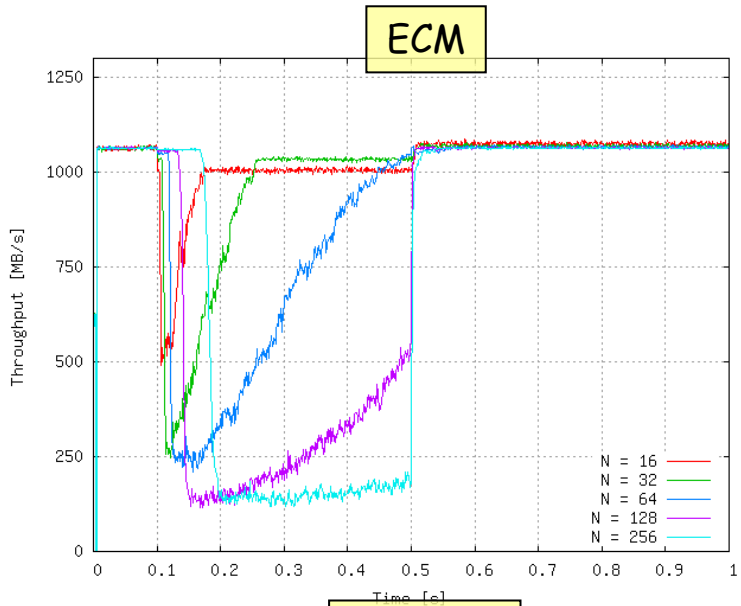
Simulation Setup & Parameters (same as before)

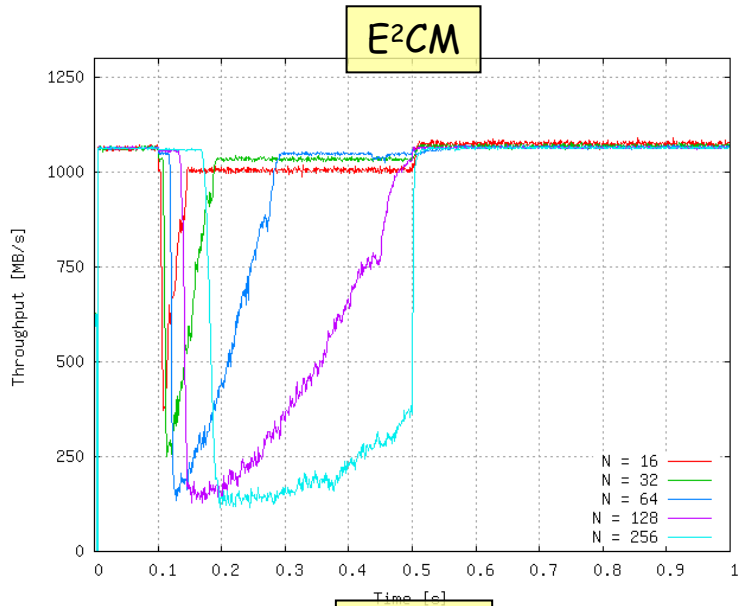
- Traffic
  - I.i.d. Bernoulli arrivals, geometrically distributed burst size around mean *B*
  - Uniform destination distribution (to all nodes except self)
  - Fixed frame size = 1500 B
- Scenario
  1. Single-hop output-generated hotspot
- Switch
  - **Radix N = [16, 32, 64, 128, 256]**
  - M = 300 KB/port
  - Partitioned memory per input, shared among all outputs
  - No limit on per-output memory usage
  - PAUSE enabled or disabled
    - Applied on a per input basis based on local high/low watermarks
    - $watermark_{high}$ = 260 KB
    - $watermark_{low}$ = 230 KB
    - If disabled, frames dropped when input partition full

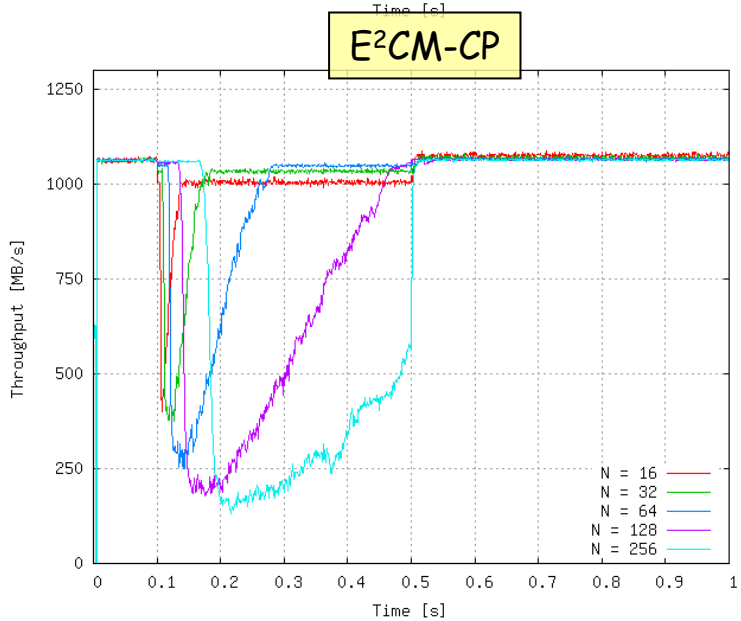
E^2CM-CP = E^2CM with continuous probing, i.e., probing is always active

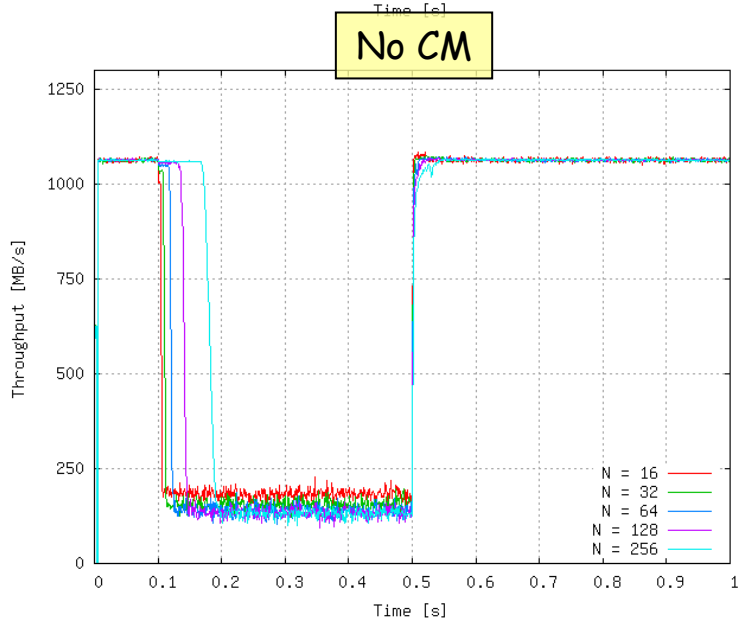
- Adapter
  - Per-node virtual output queuing, round-robin scheduling
  - No limit on number of rate limiters
  - Ingress buffer size = 1500 KB, partitioned across VOQs, per-flow selective source quench used when VOQ full, round-robin VOQ service
  - Egress buffer size = 150 KB
  - PAUSE enabled
    - $watermark_{high}$ = 150 – rtt*bw KB
    - $watermark_{low}$ = $watermark_{high}$ - 10 KB
- ECM
  - W = 2.0
  - Q_{eq} = 75 KB (= M/4)
  - G_d = 0.5 / ((2*W+1)*Q_{eq})
  - G_{i0} = (R_{link} / R_{unit}) * ((2*W+1)*Q_{eq})
  - G_i = 0.1 * G_{i0}
  - P_{sample} = 2% (on average 1 sample every 75 KB
  - R_{unit} = R_{min} = 1 Mb/s
  - BCN_MAX enabled, threshold = 260 KB
  - No BCN(0,0), no self-increase
- E^2CM (per-flow)
  - W = 2.0
  - $Q_{eq,flow}$ = 15 KB
  - $G_{d,\ flow}$ = 0.5 / ((2*W+1)*$Q_{eq,flow}$)
  - $G_{i,\ flow}$ = 0.005 * (R_{link} / R_{unit}) / ((2*W+1)*$Q_{eq,flow}$)
  - P_{sample} = 2% (on average 1 sample every 75 KB)
  - R_{unit} = R_{min} = 1 Mb/s
  - BCN_MAX enabled, threshold = 52 KB

Aggregate throughput - PAUSE disabled

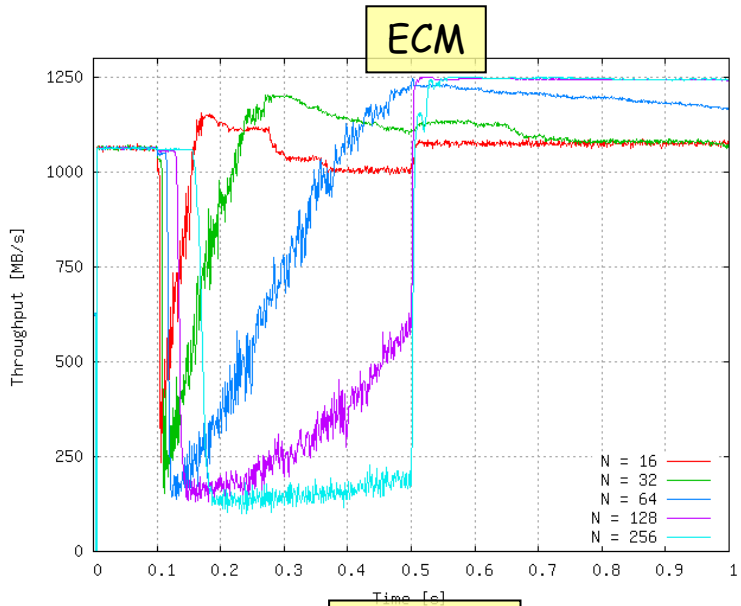


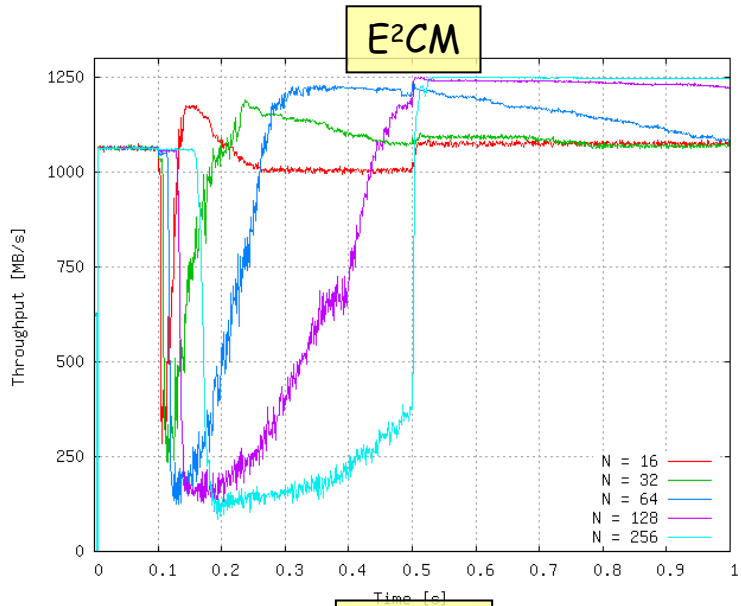


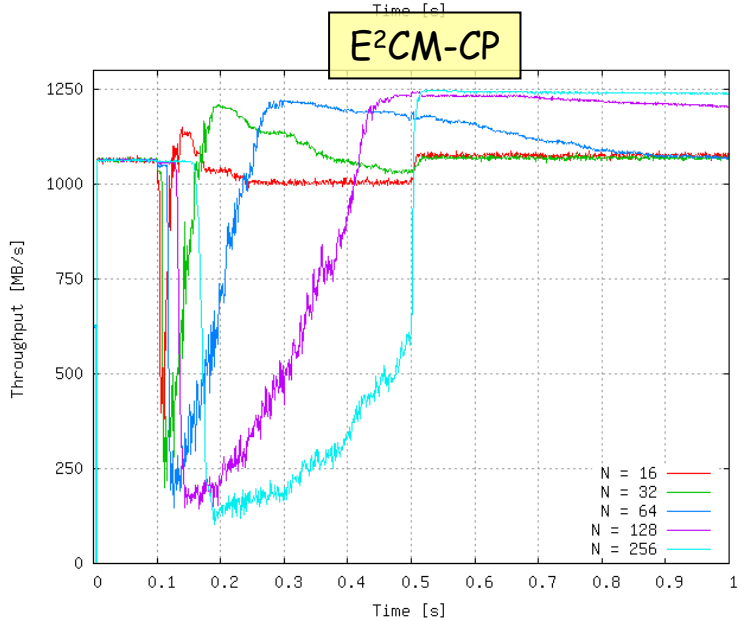


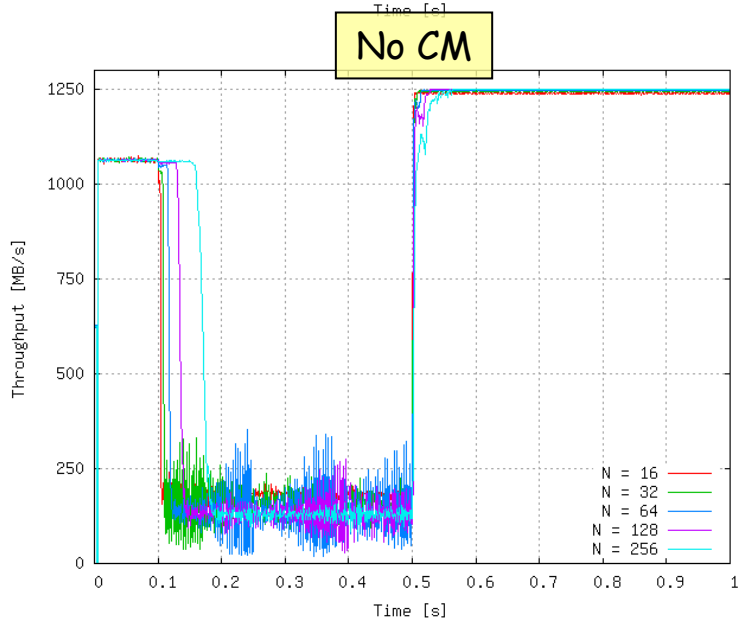


Aggregate throughput – PAUSE enabled

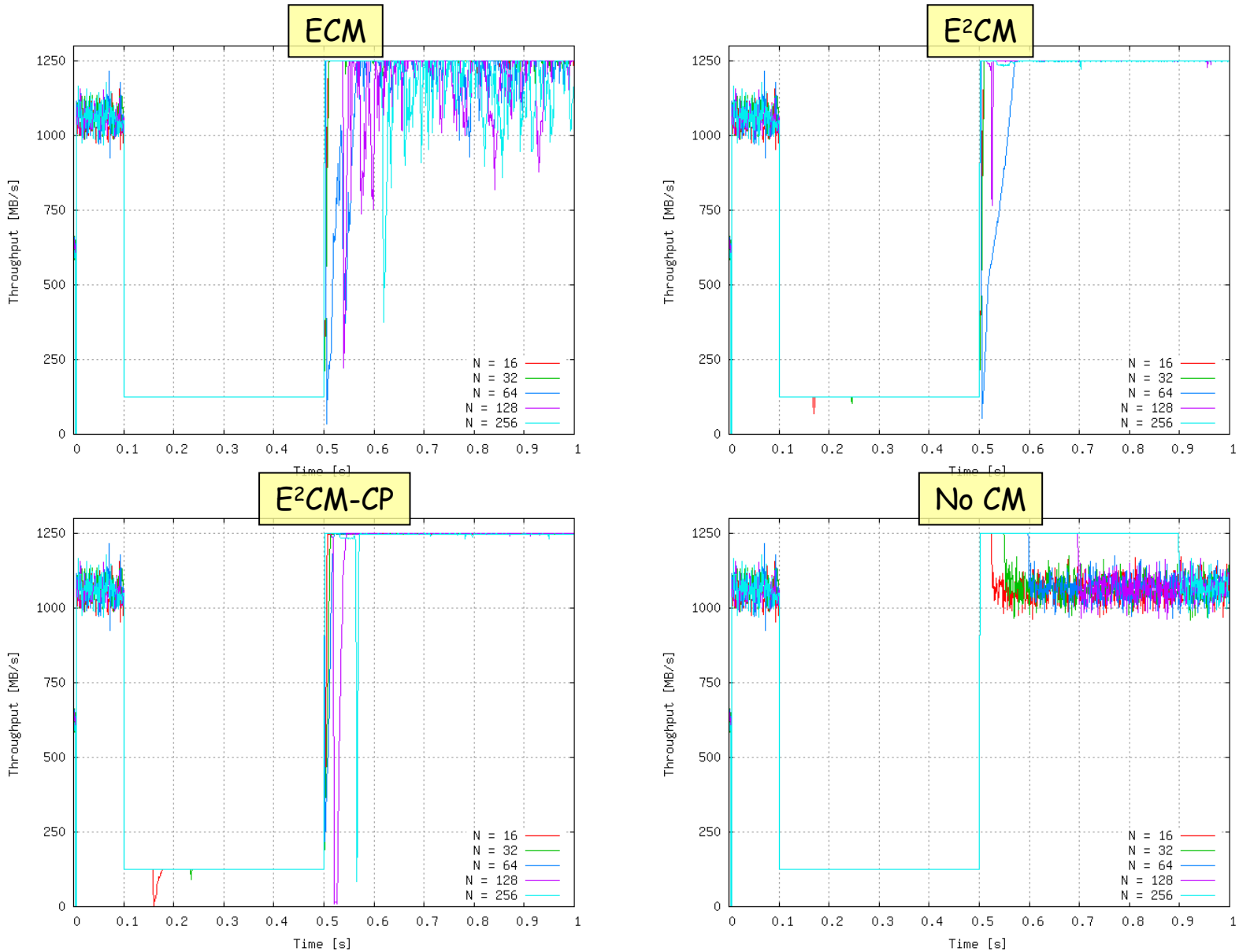




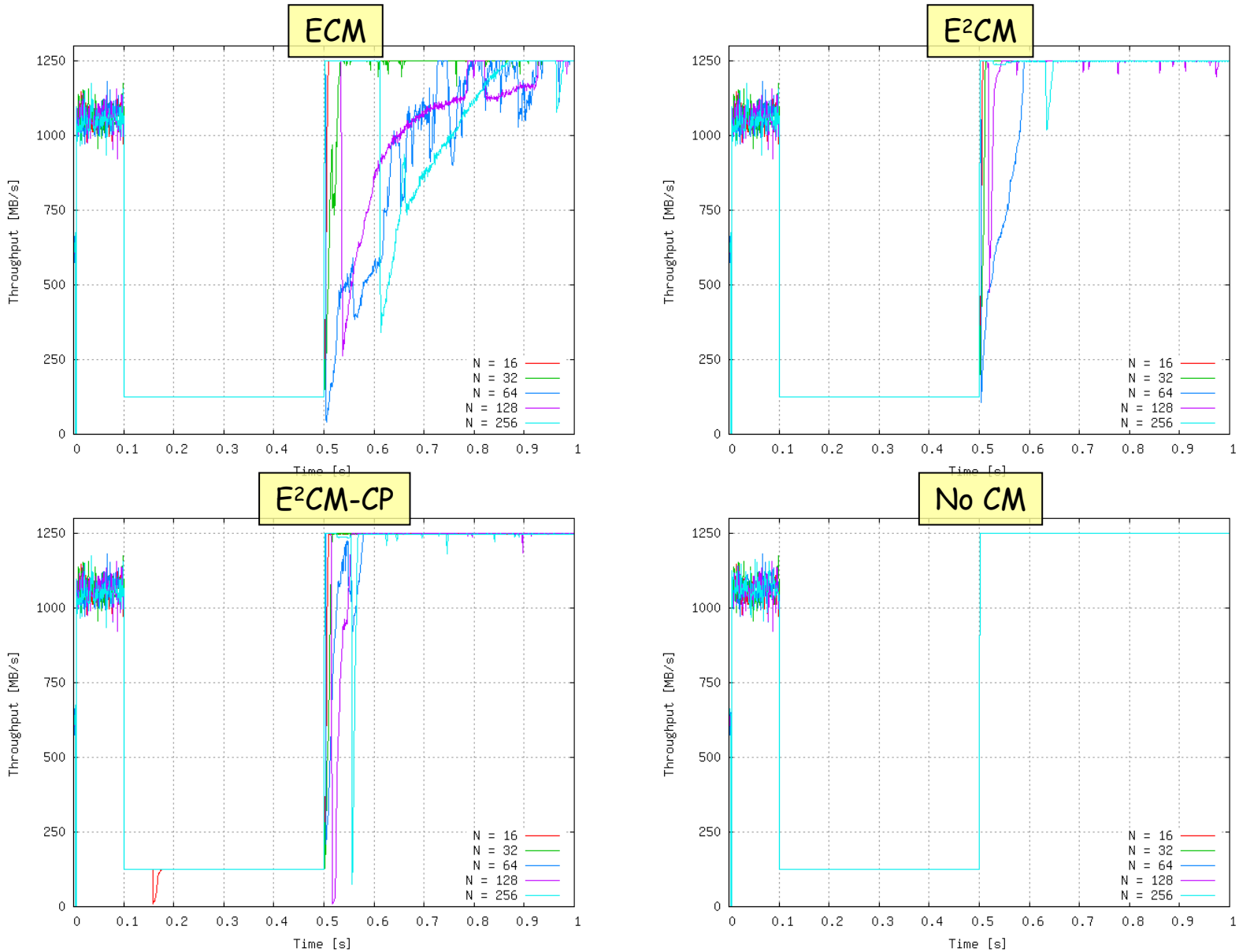




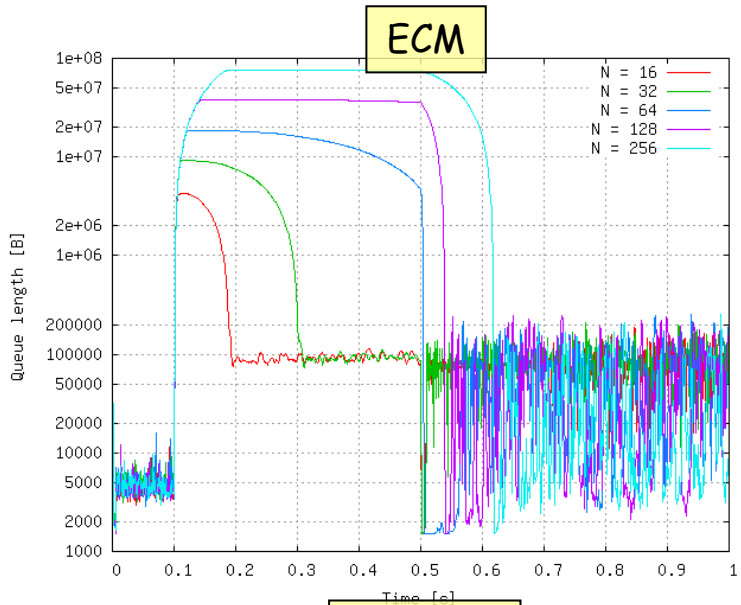
Hot port throughput - PAUSE disabled

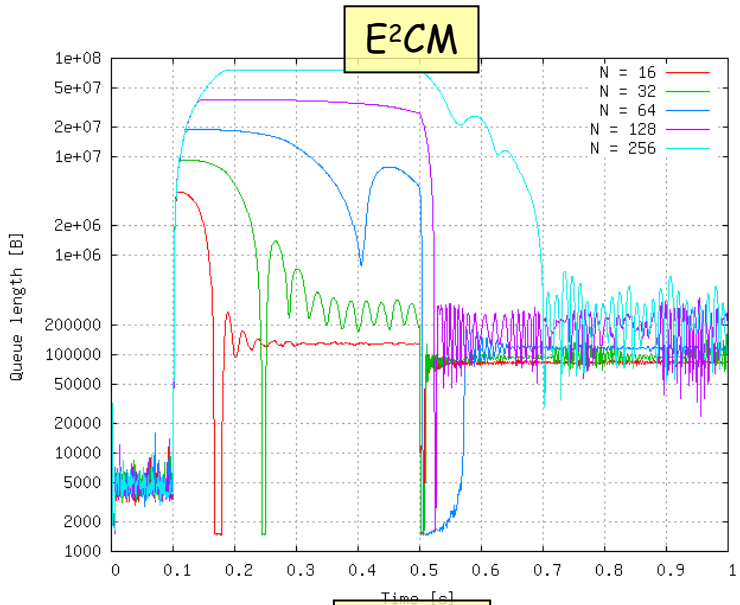


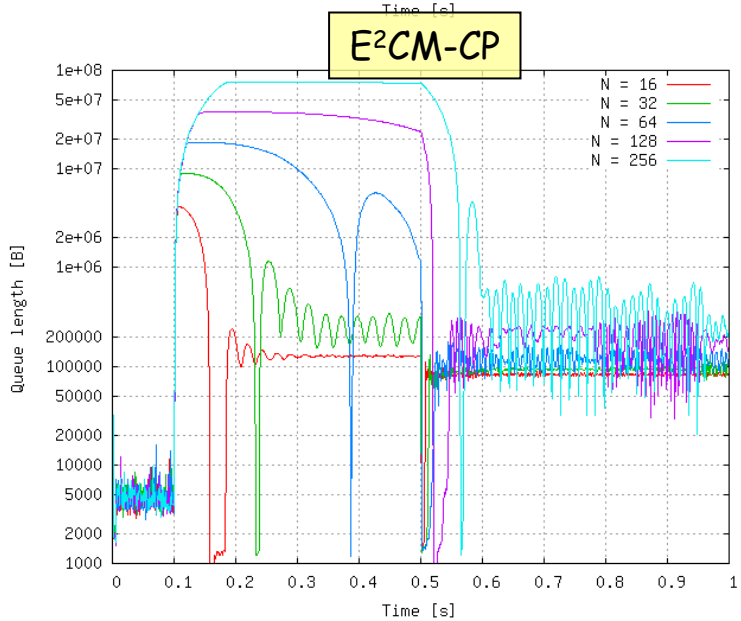
Hot port throughput – PAUSE enabled

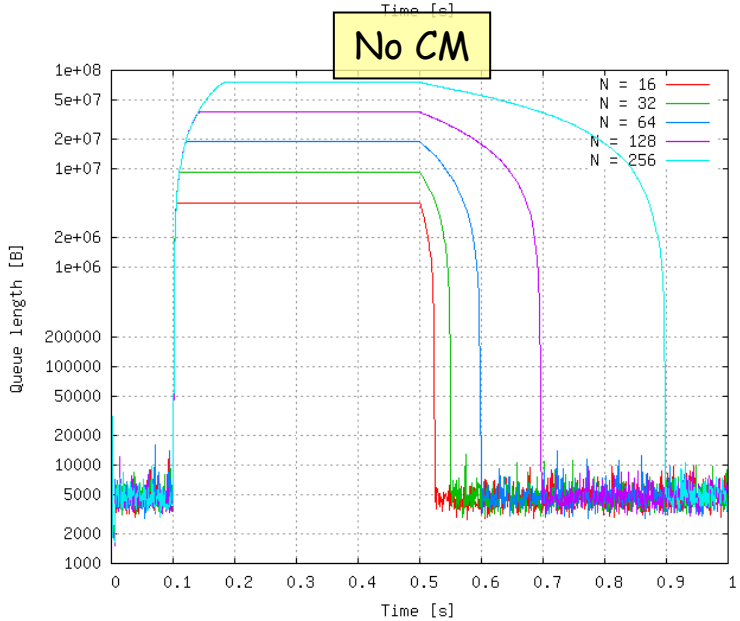


Hot queue length - PAUSE disabled

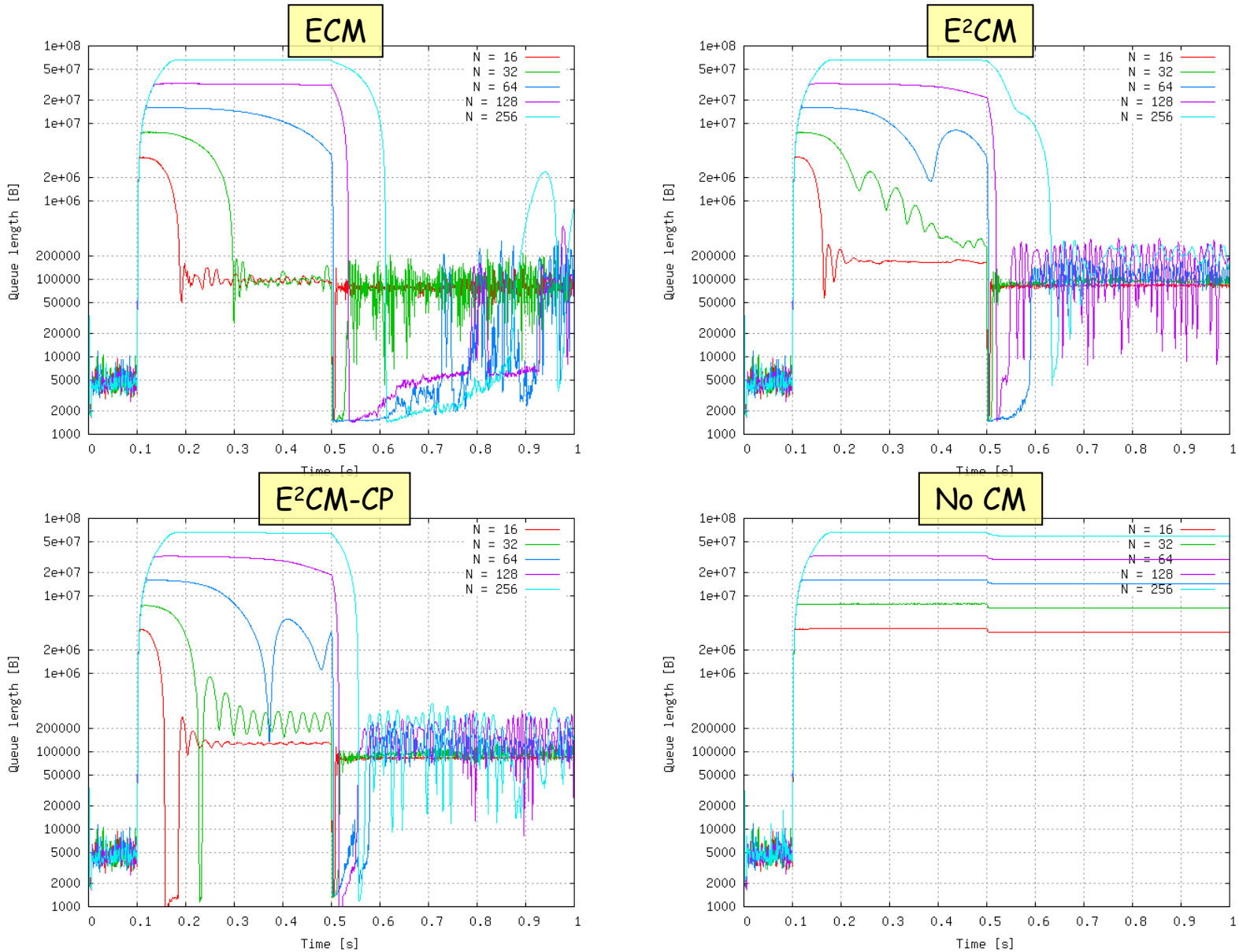




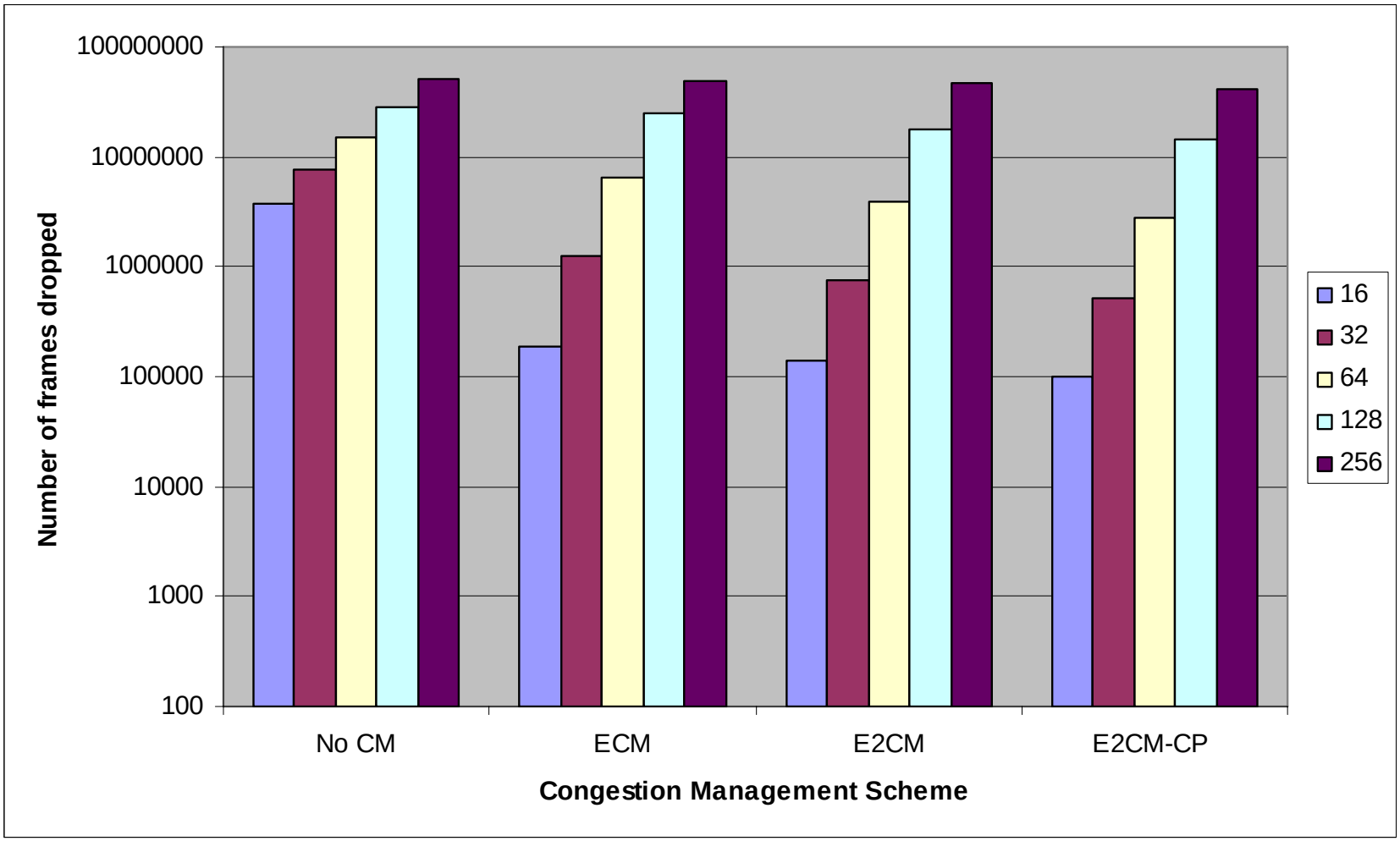




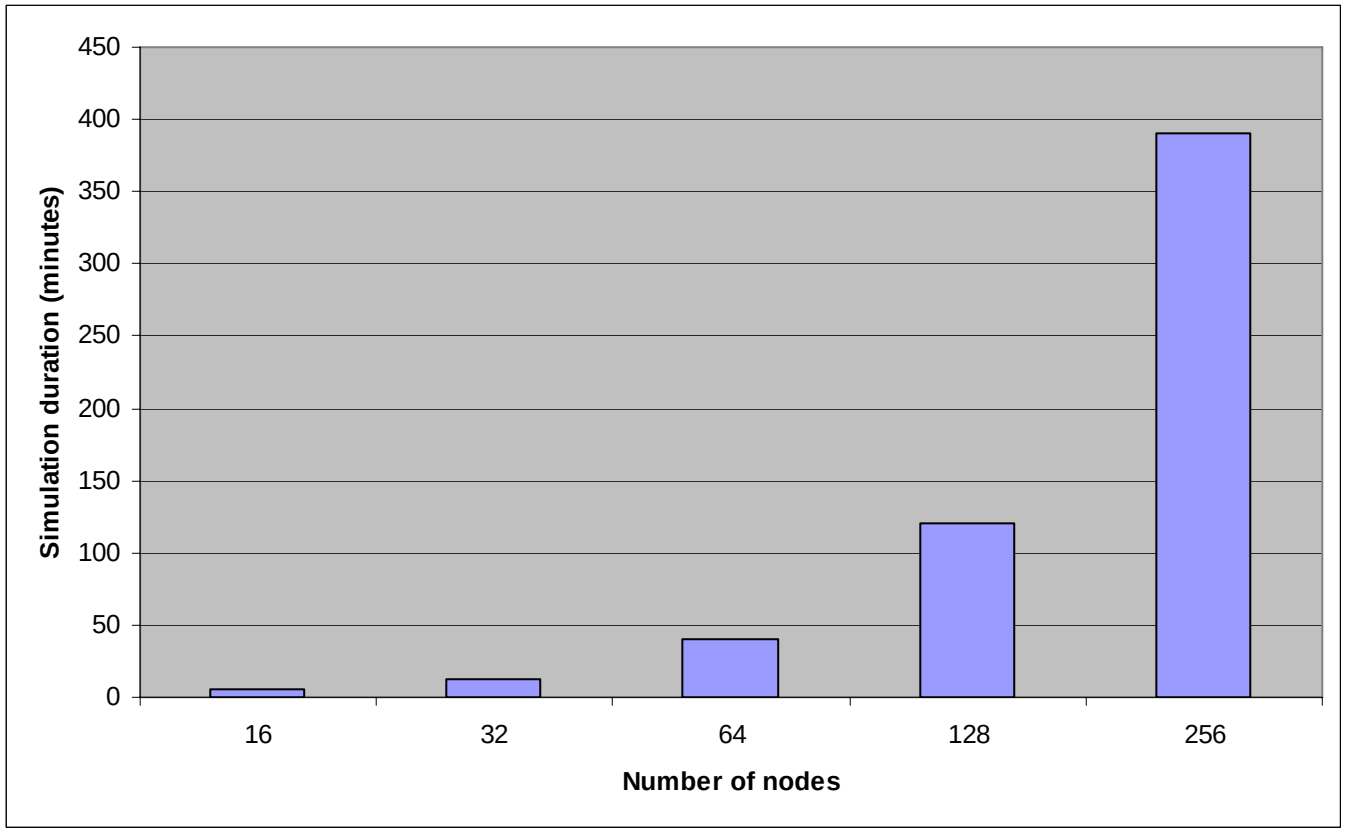
Hot queue length – PAUSE enabled



Frame drops (PAUSE disabled)



Simulation duration per run



- Number of nodes doubles ➔ simulation time triples

Conclusions on High-HSD OG: A Corner Case?

- Recovery duration drastically increases with HSD
  - With 256 nodes, recovery exceeds hotspot duration (400 ms) in all cases
  - PAUSE makes no substantial difference, except that accumulated backlog for cold ports causes overshoot when used
  - E^2CM with continuous probing performs (for this scenario) better than both baselines

- Persistent high HSD requires parameter tuning
  - Is this really a common case to be worried about or rather a "corner case"?
  - Higher decrease gains?
  - Currently also testing use of BCN(0,0), as BCN_MAX does not result in sufficiently fast throttling