ECM and E^{2}CM performance w/ BCN(0,0)

Single-Hop High Degree Hotspot

Cyriel Minkenberg & Mitch Gusat
IBM Research GmbH, Zurich
May 3, 2007

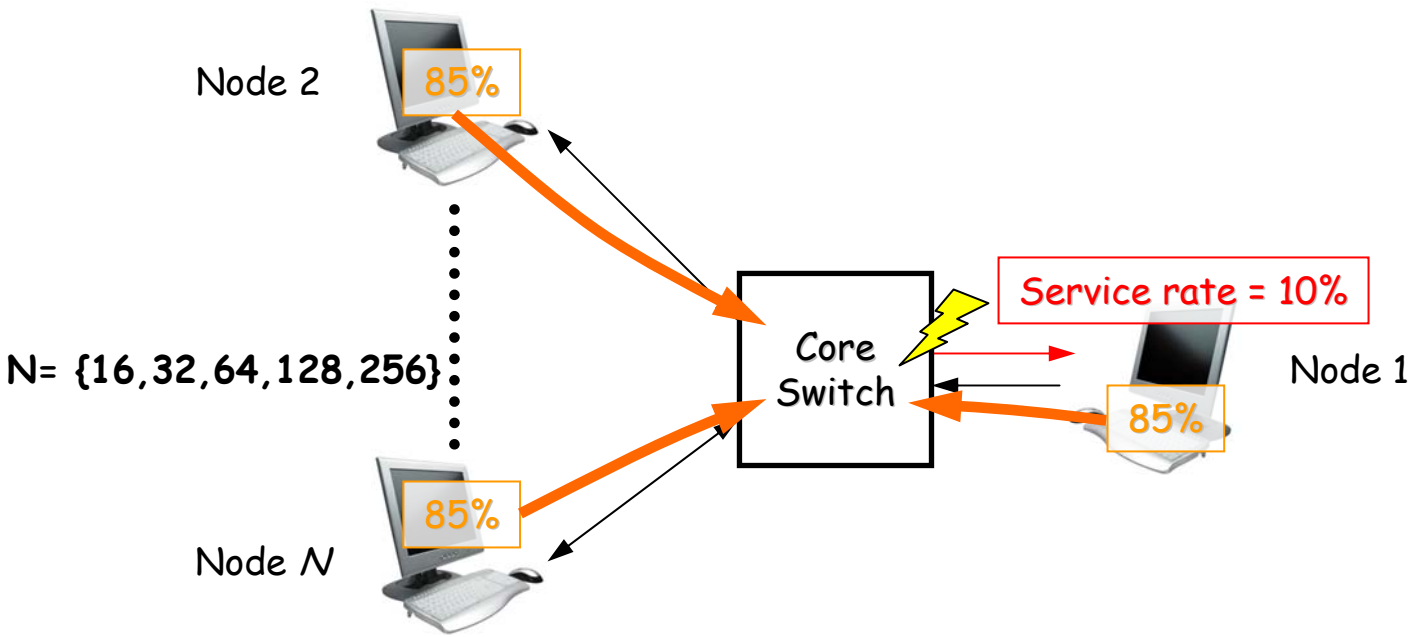
Targets

1. Study Output-Generated (OG) single-hop with **high hotspot degree** (HSD) congestion
2. Test the impact of BCN(0,0)

Conditions, parameters, simulation environment

- Traffic: i.i.d. Bernoulli arrivals
- LL-FC: runs with and w/o PAUSE
- CM: No CM, ECM, E^2CM, E^2CM-CP
  - With/without BCN(0,0)
- 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

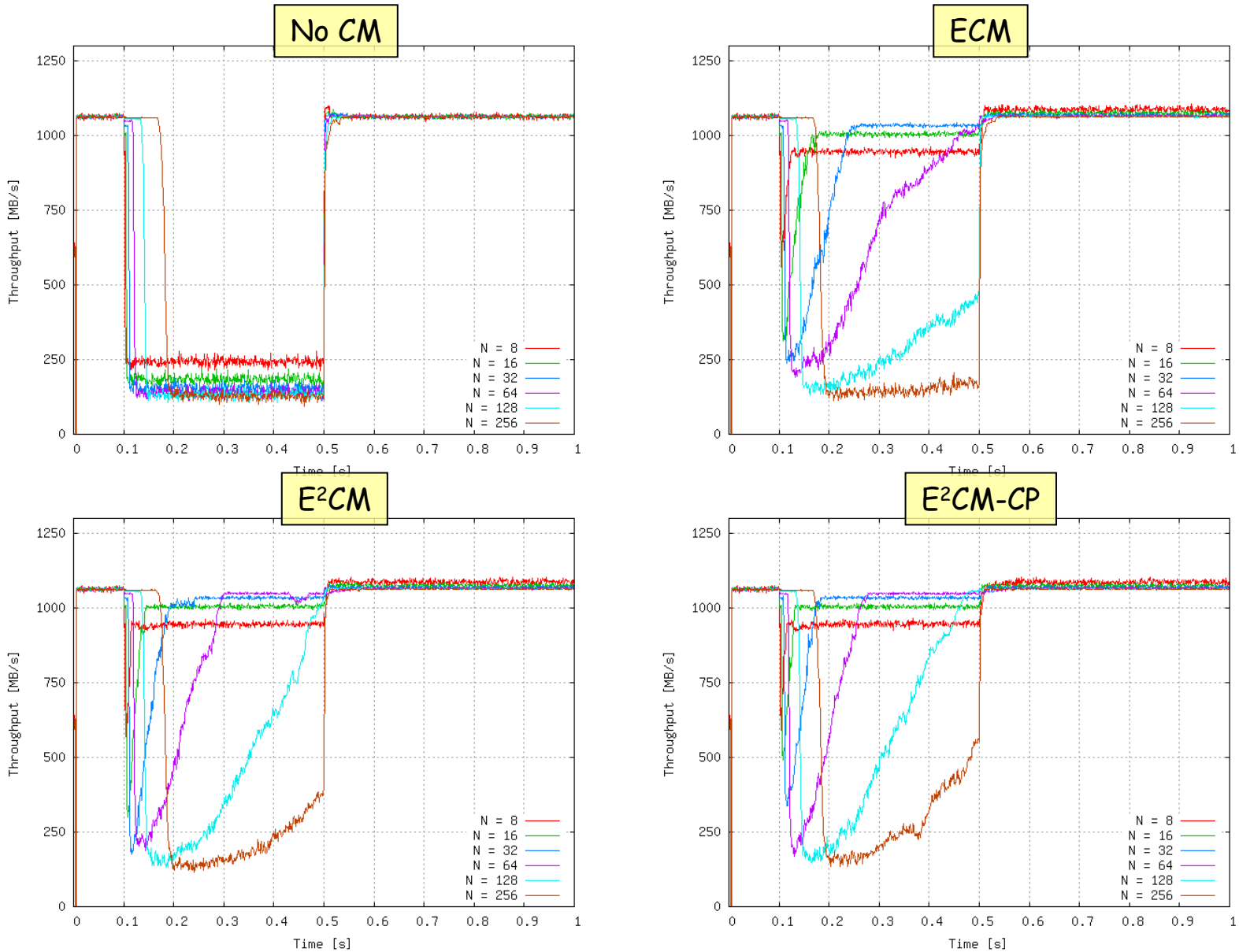
Simulation Setup & Parameters (same as before)

- 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 = [8, 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

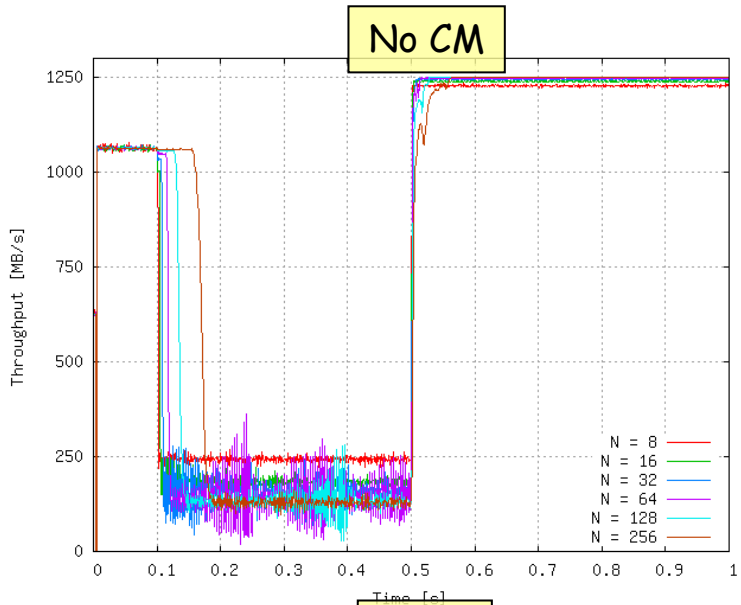
E²CM-CP = E²CM with continuous probing, i.e., probing is always active

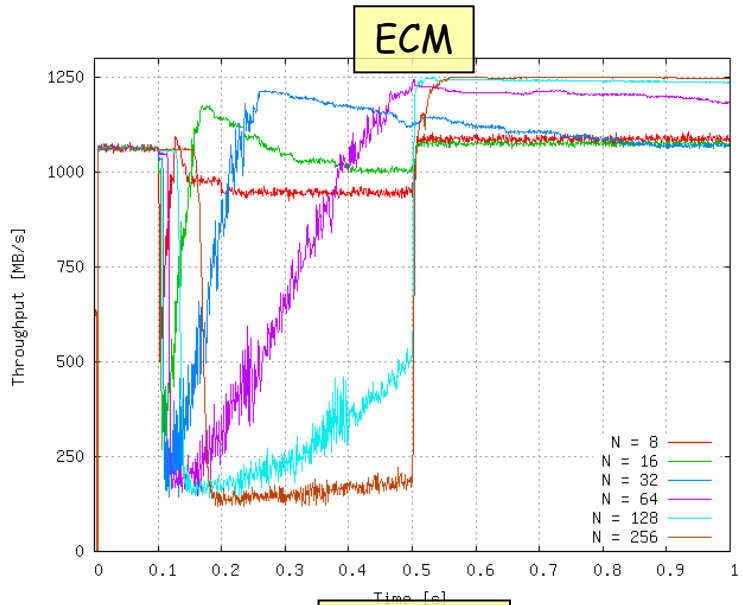
- Adapter
  - Per-node virtual output queuing, round-robin scheduling
  - No limit on number of rate limiters
  - Ingress buffer size = 1500 KB, partitioned across VOQs, per-flow selective source quench used when VOQ full, round-robin VOQ service
  - Egress buffer size = 150 KB
  - PAUSE enabled
    - $watermark_{high}$ = 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
  - BCN(0,0) dis/enabled, threshold = 1040 KB
- E²CM (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
  - BCN(0,0) dis/enabled, threshold = 208 KB

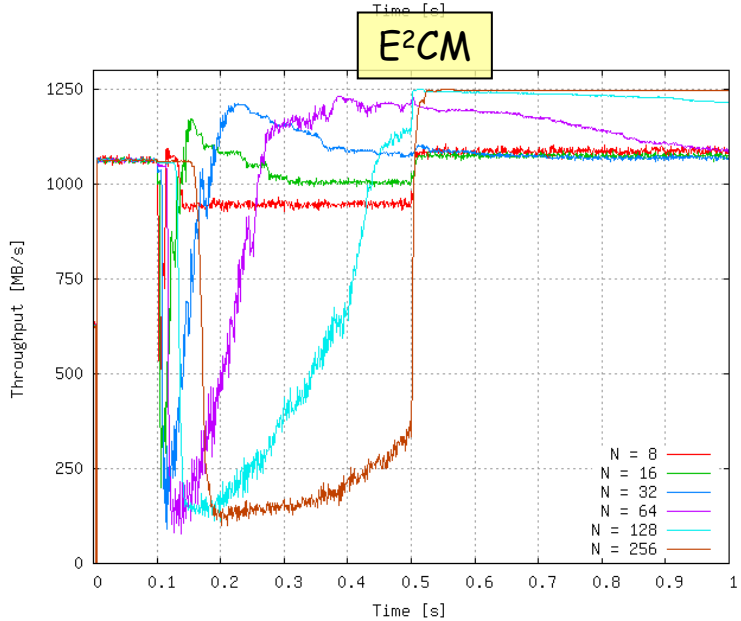
Aggregate throughput: no BCN(0,0), PAUSE disabled

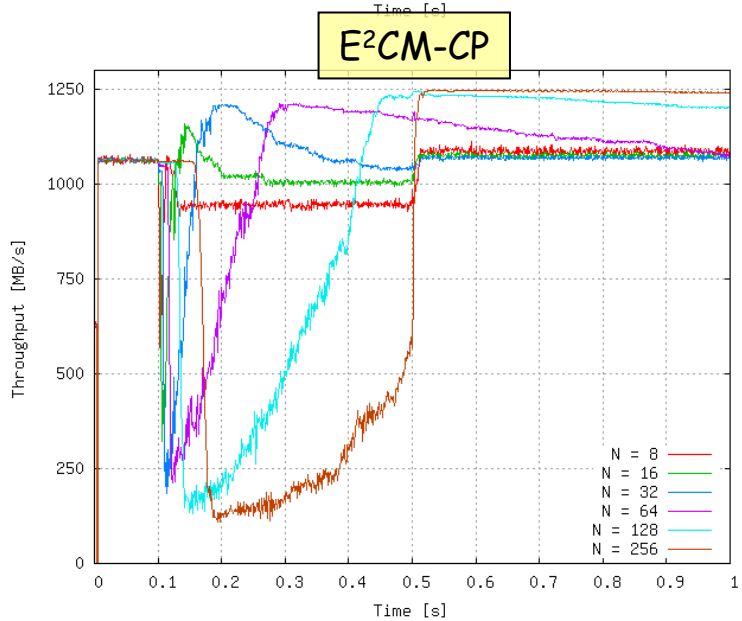


Aggregate throughput: no BCN(0,0), PAUSE enabled

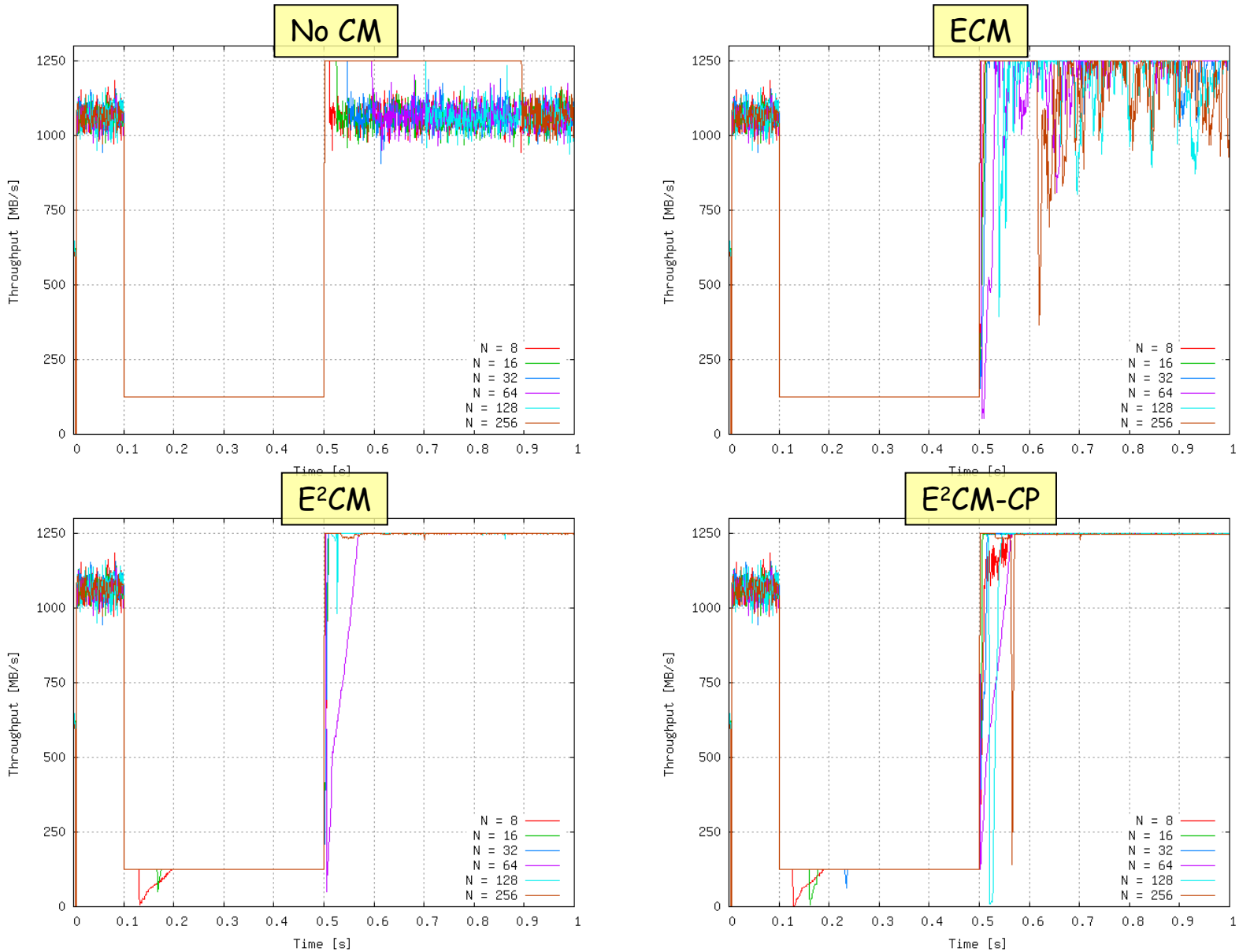




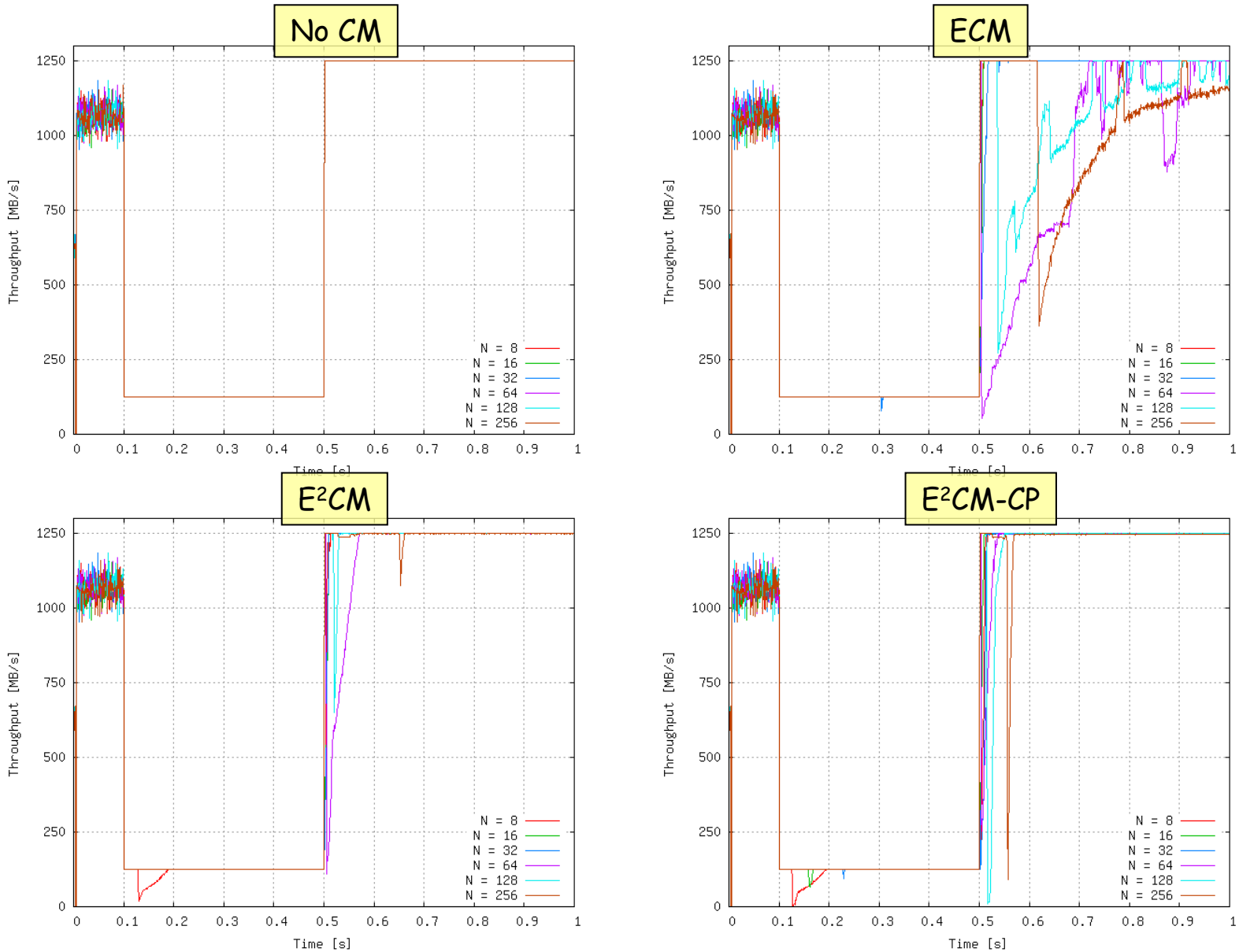




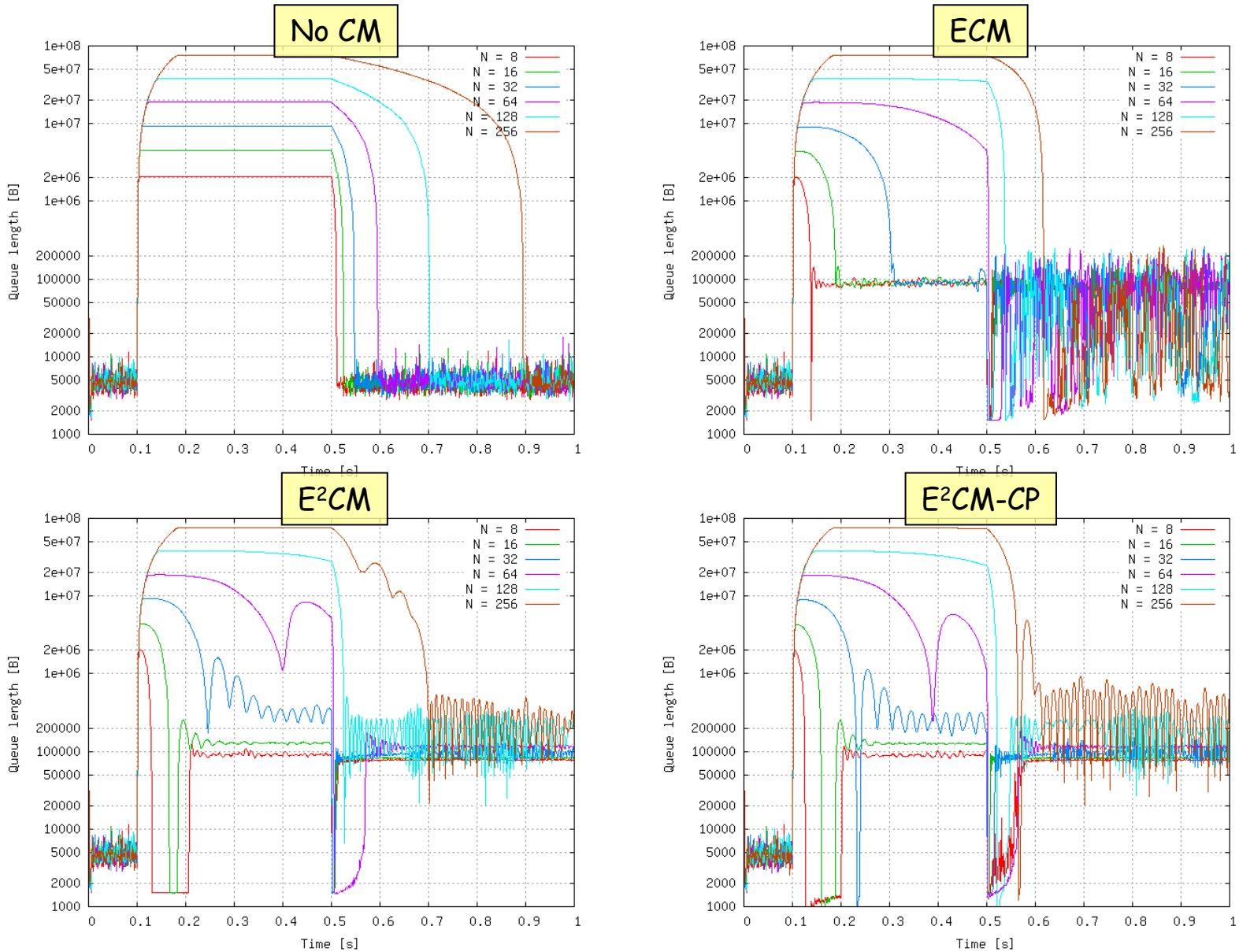
Hot port throughput: no BCN(0,0), PAUSE disabled



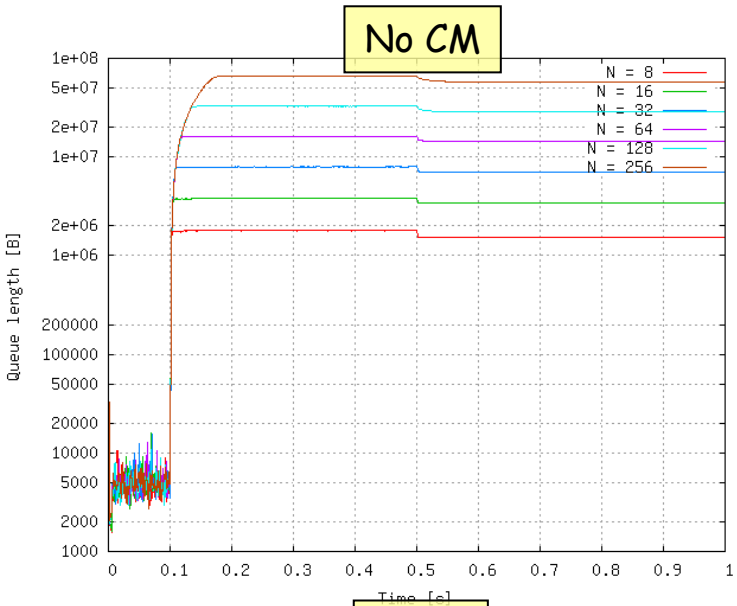
Hot port throughput: no BCN(0,0), PAUSE enabled

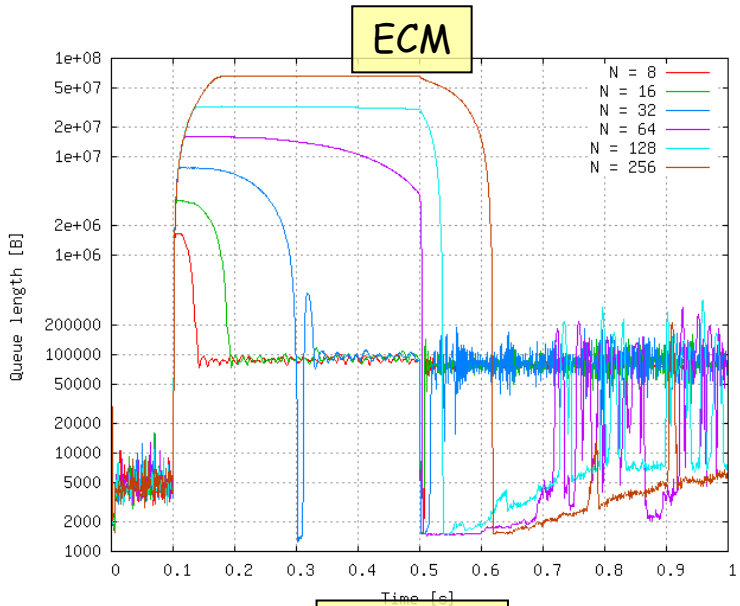


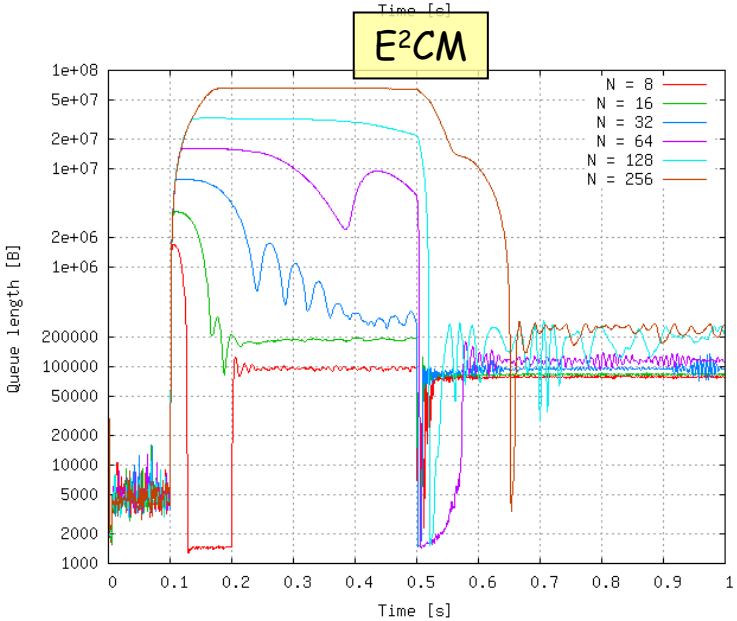
Hot queue length: no BCN(0,0), PAUSE disabled

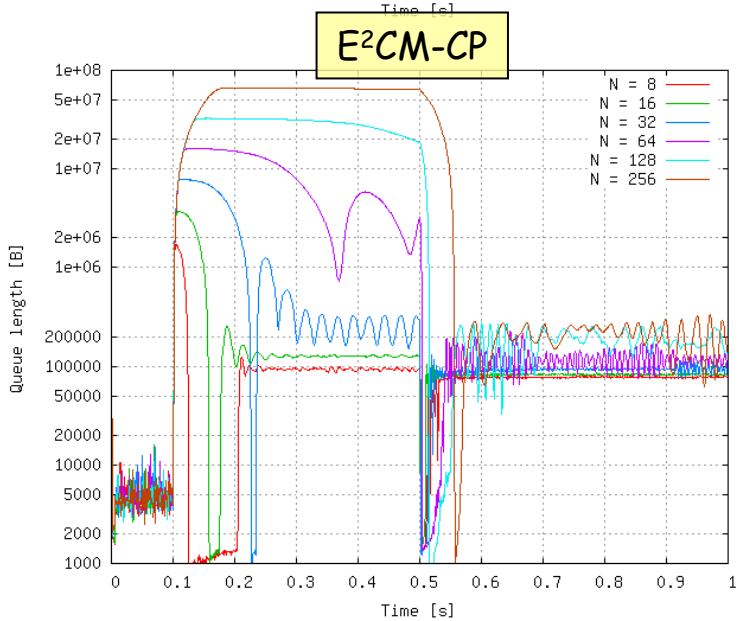


Hot queue length: no BCN(0,0), PAUSE enabled

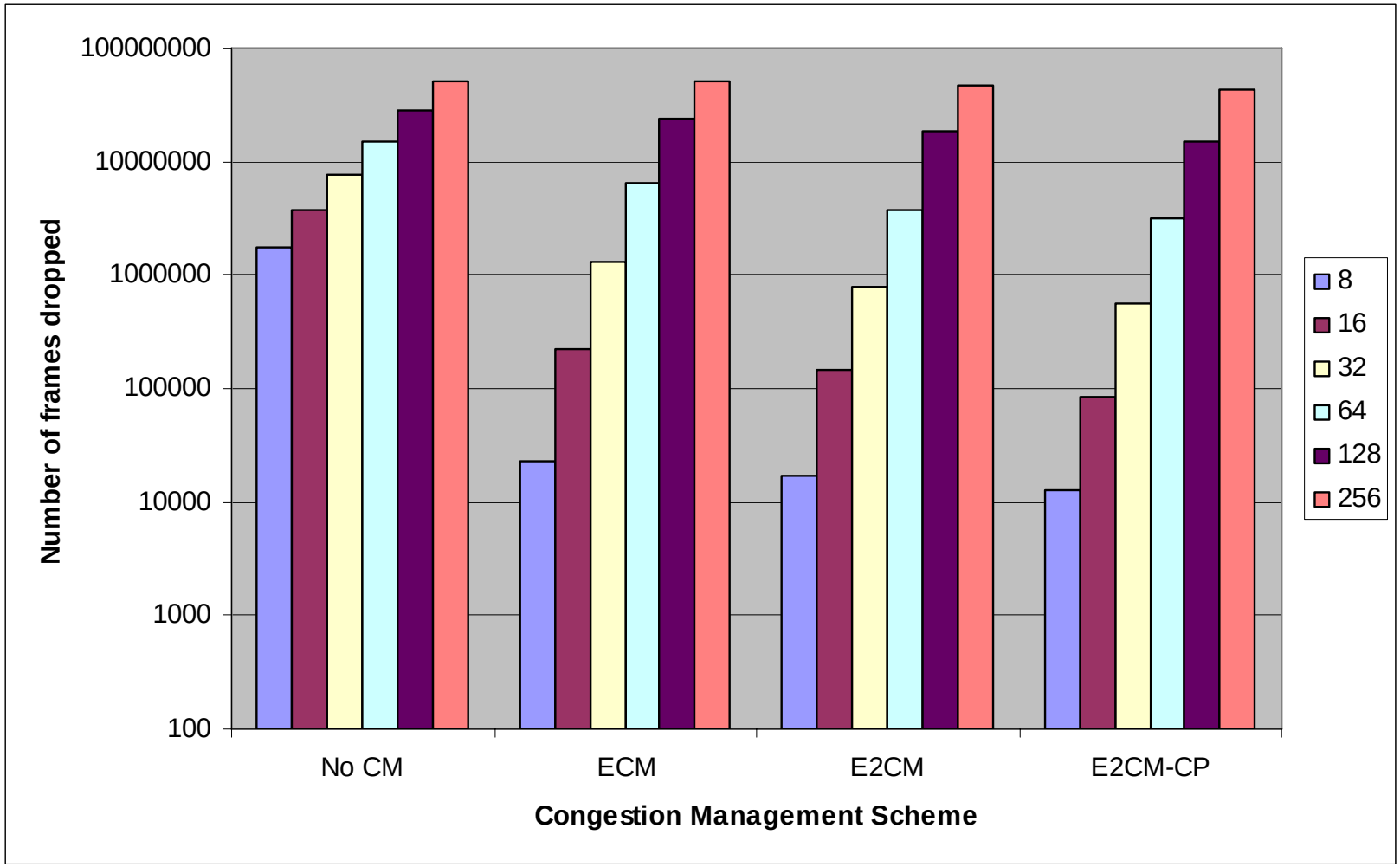






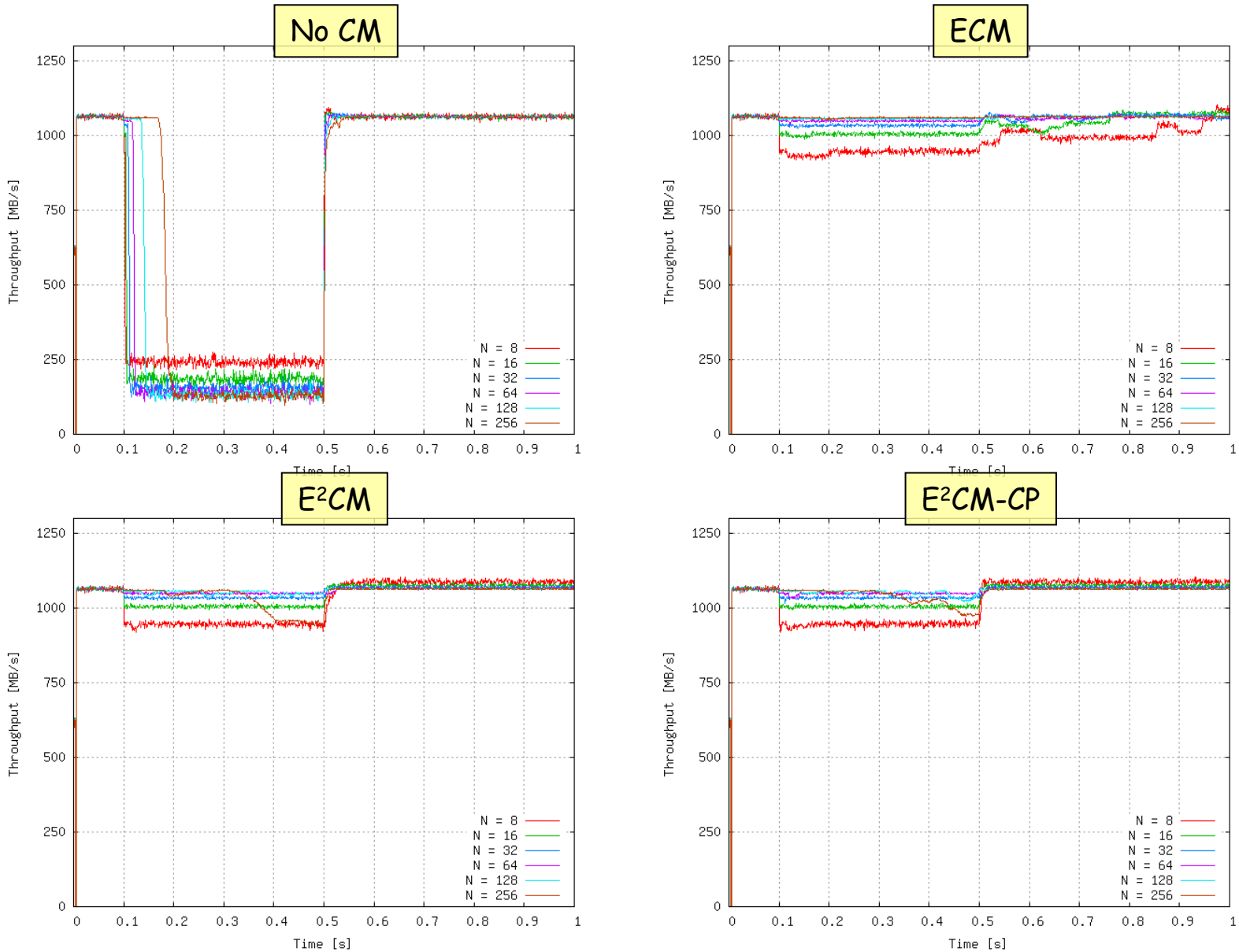


Frame drops: no BCN(0,0), PAUSE disabled

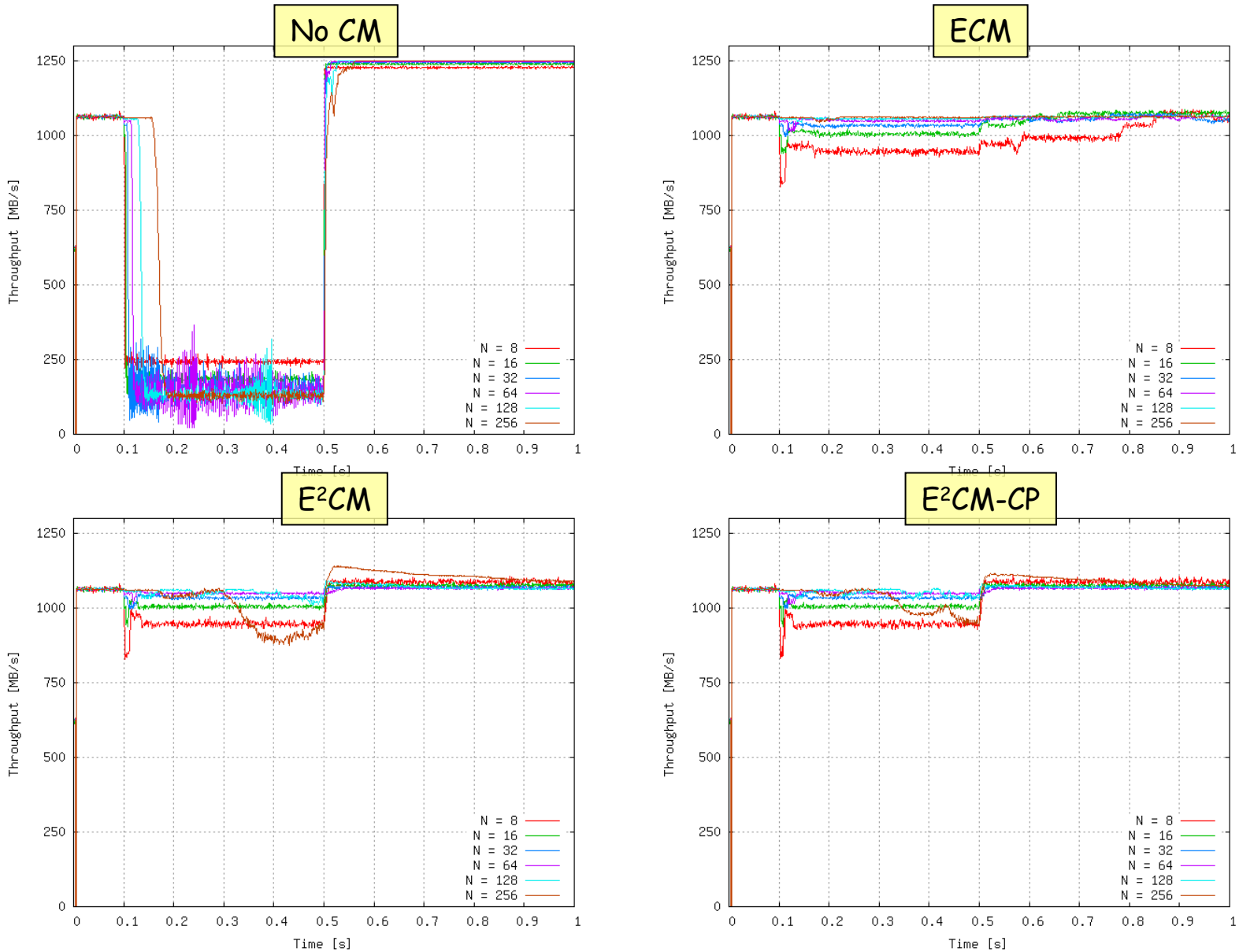


Simulation results w/ BCN(0,0)

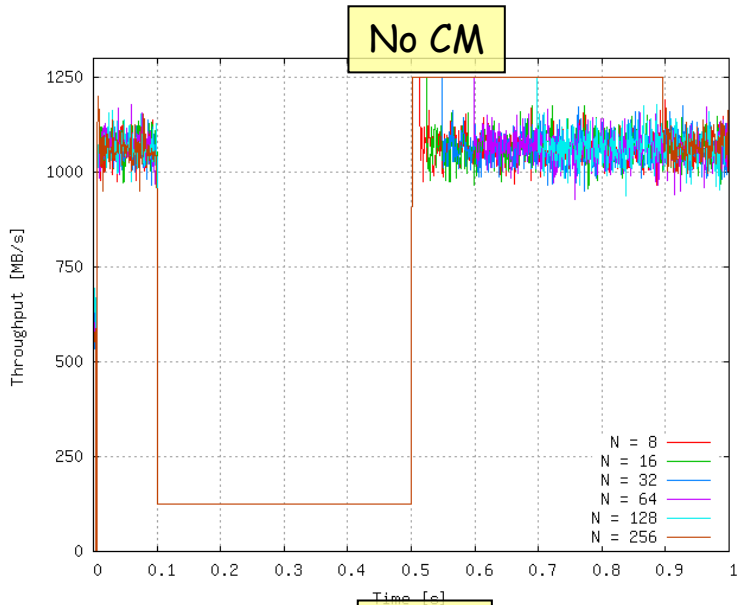
Aggregate throughput: w/ BCN(0,0), PAUSE disabled

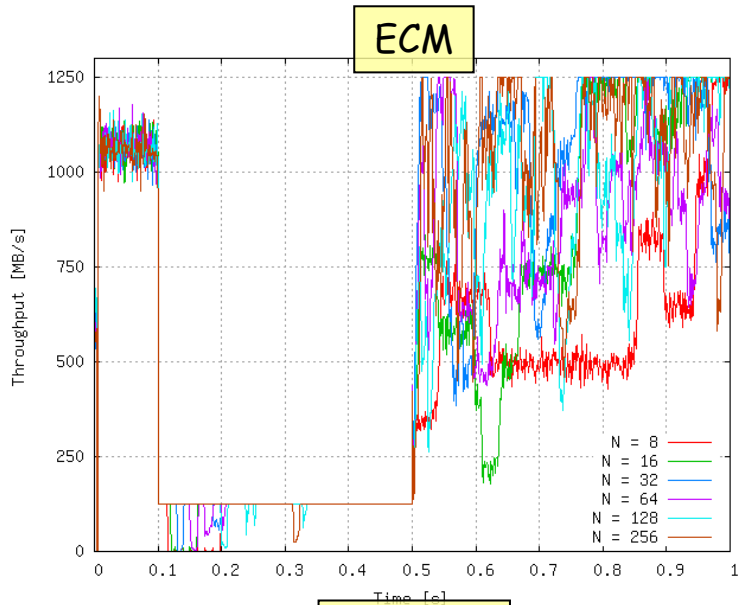


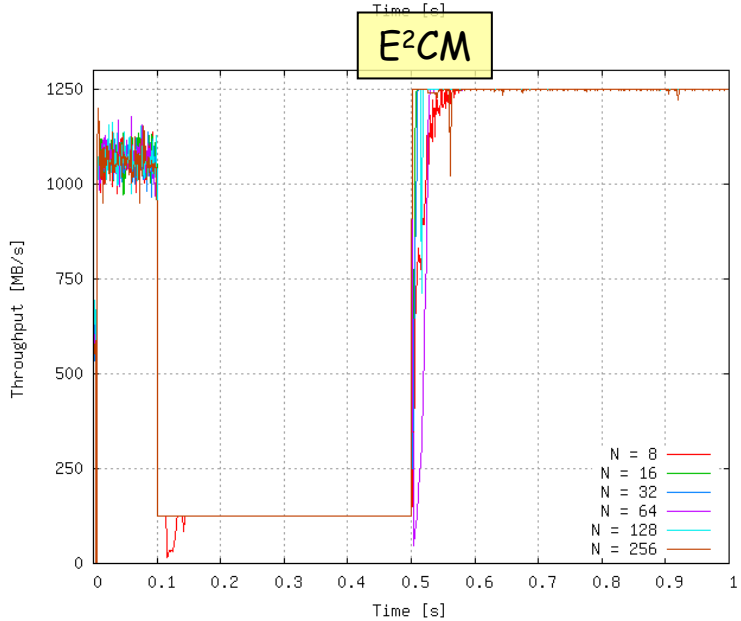
Aggregate throughput: w/ BCN(0,0), PAUSE enabled

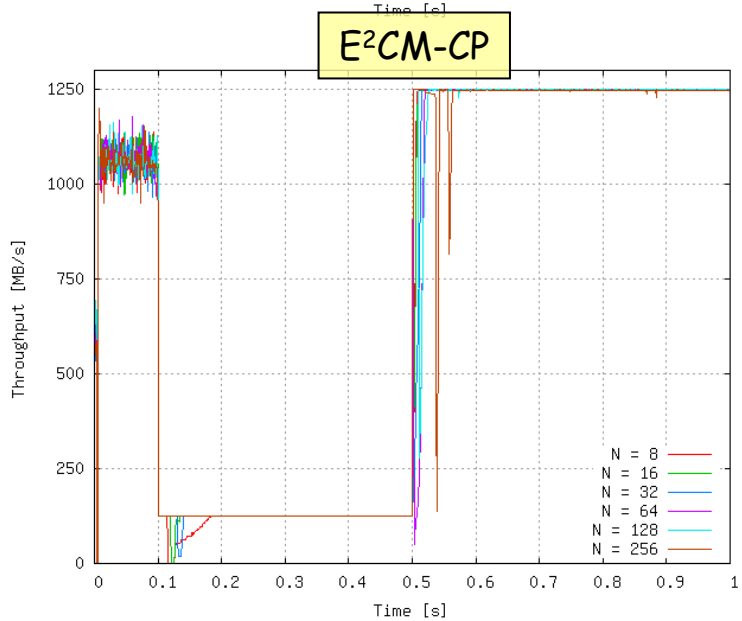


Hot port throughput: w/ BCN(0,0), PAUSE disabled

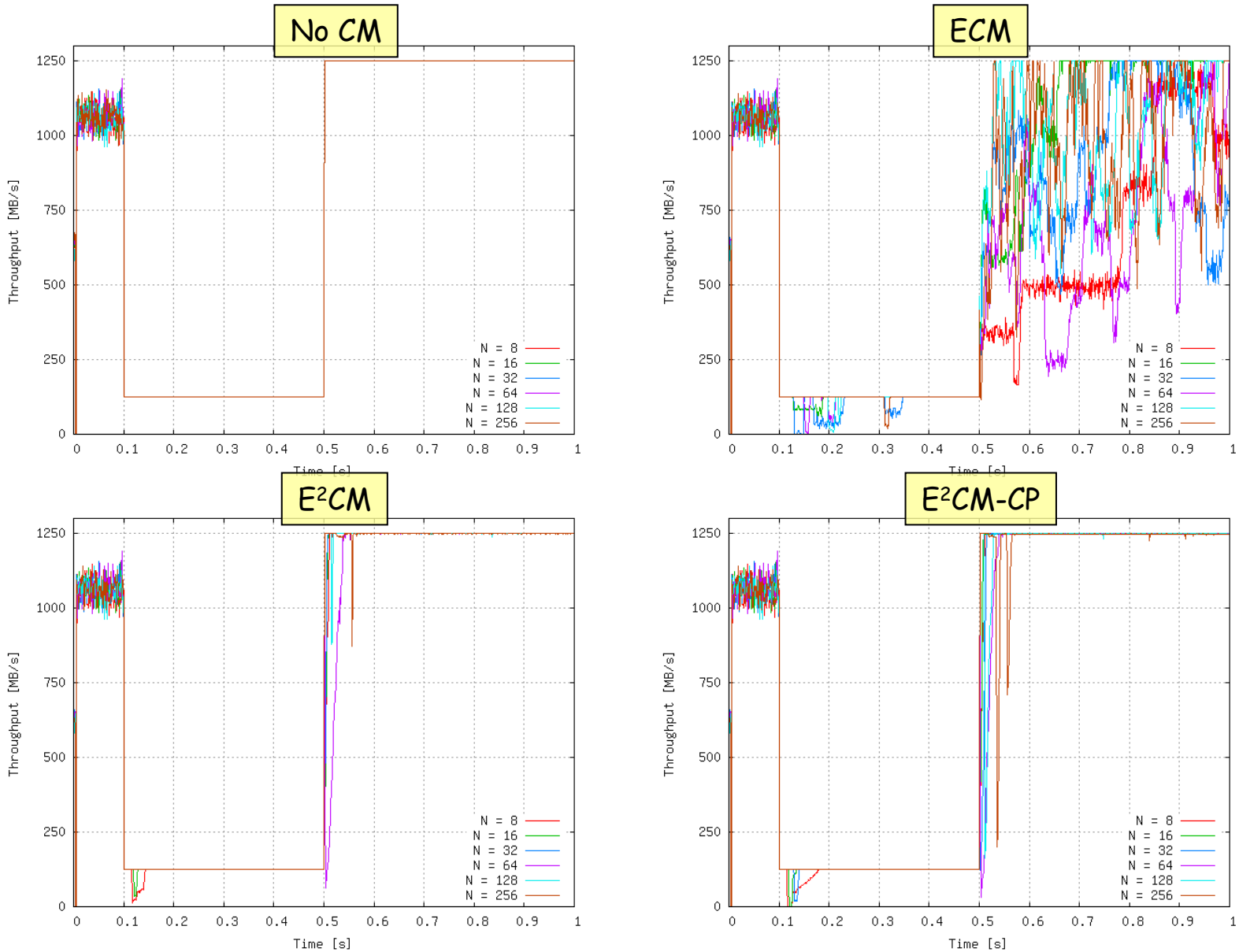




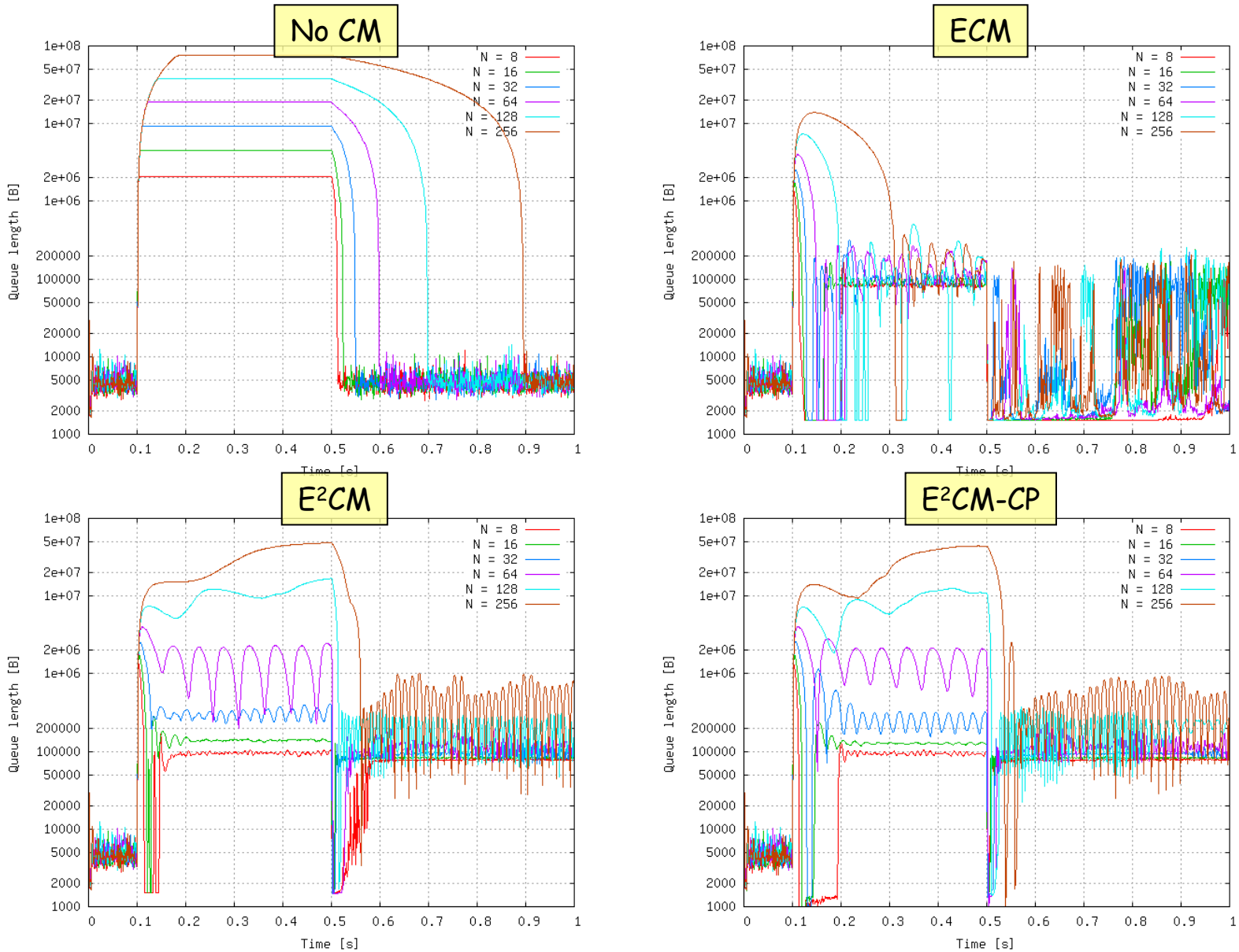




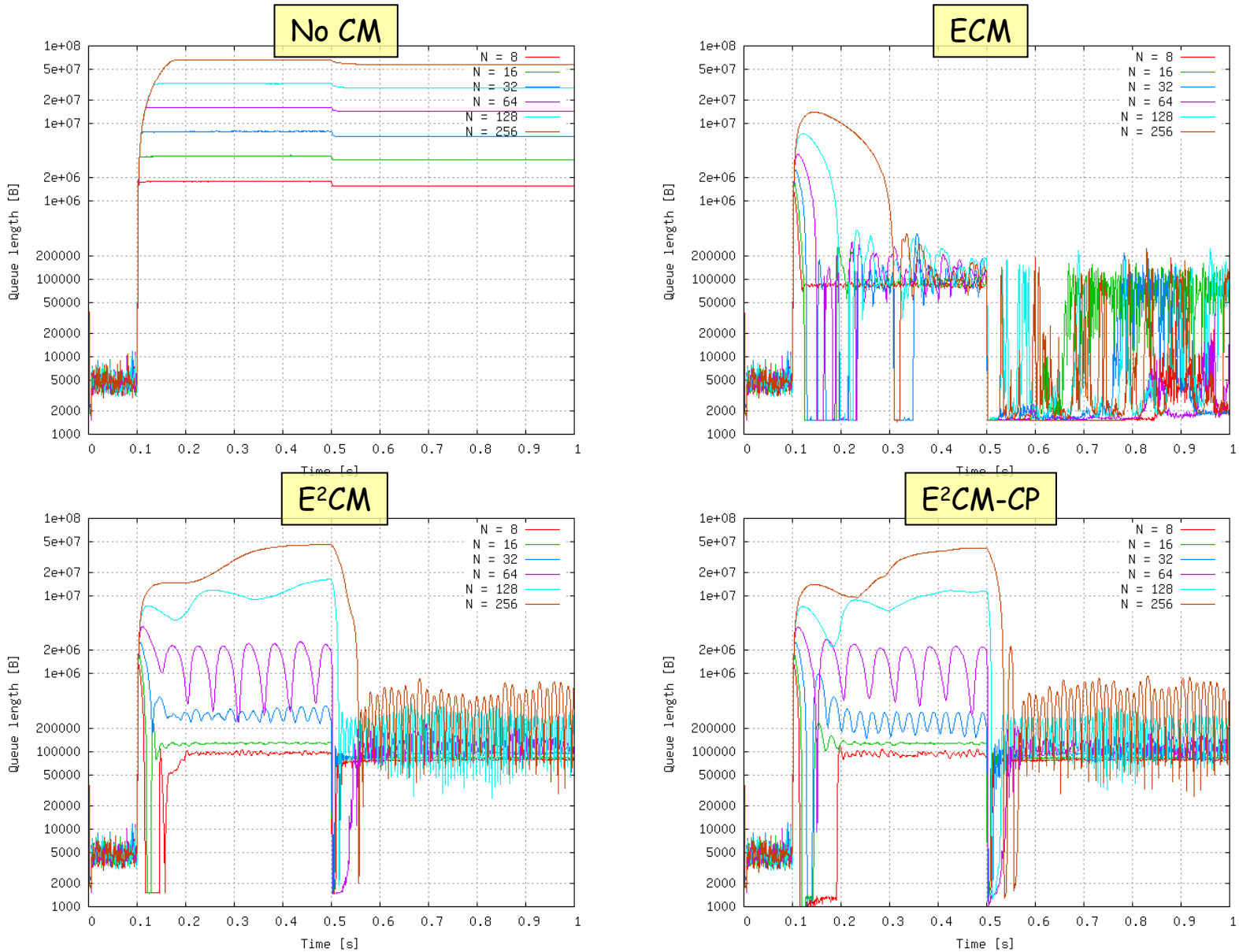
Hot port throughput: w/ BCN(0,0), PAUSE enabled



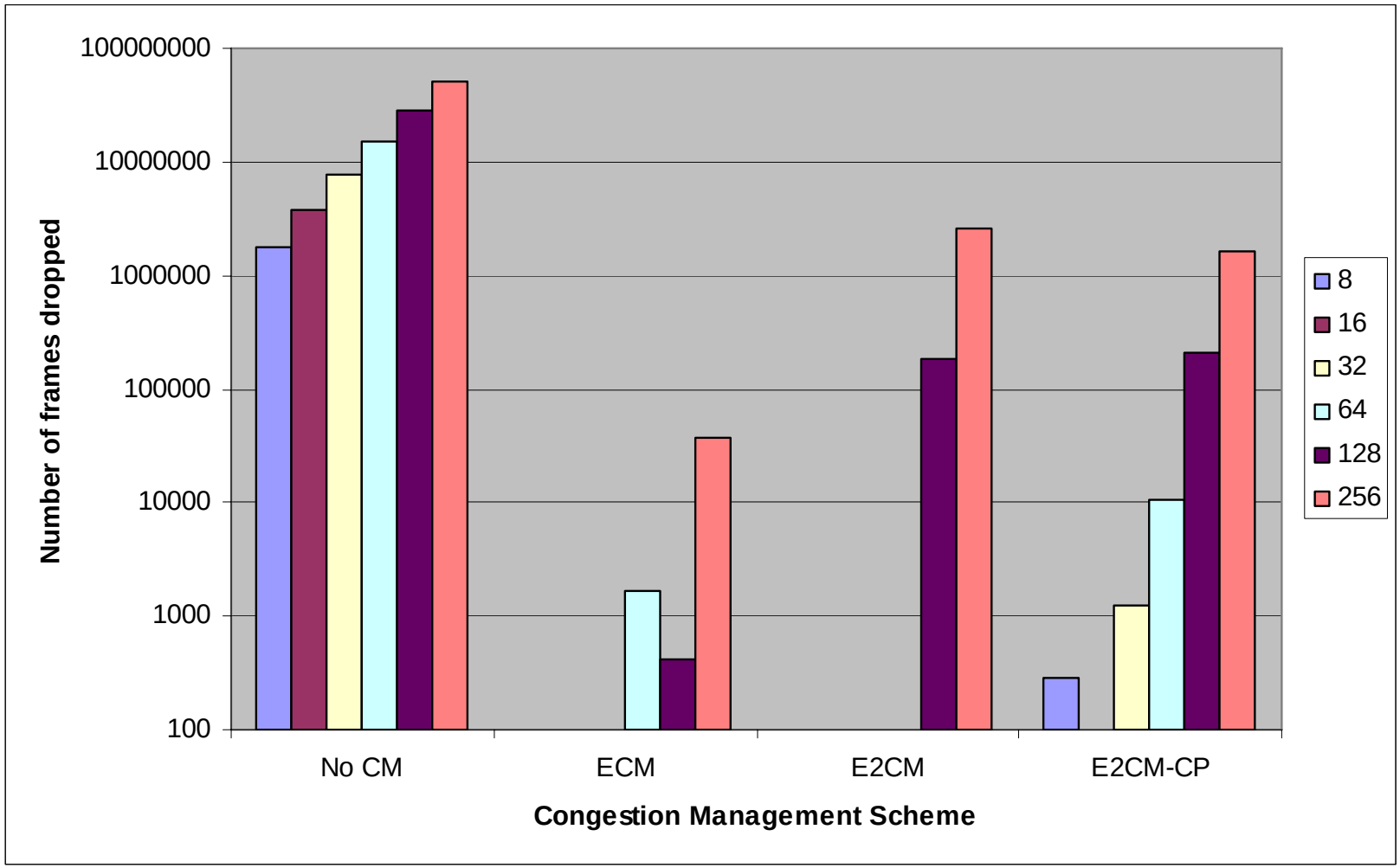
Hot queue length: w/ BCN(0,0), PAUSE disabled



Hot queue length: w/ BCN(0,0), PAUSE enabled



Frame drops: w/ BCN(0,0), PAUSE disabled



Conclusions on High-HSD OG w/ BCN(0,0)

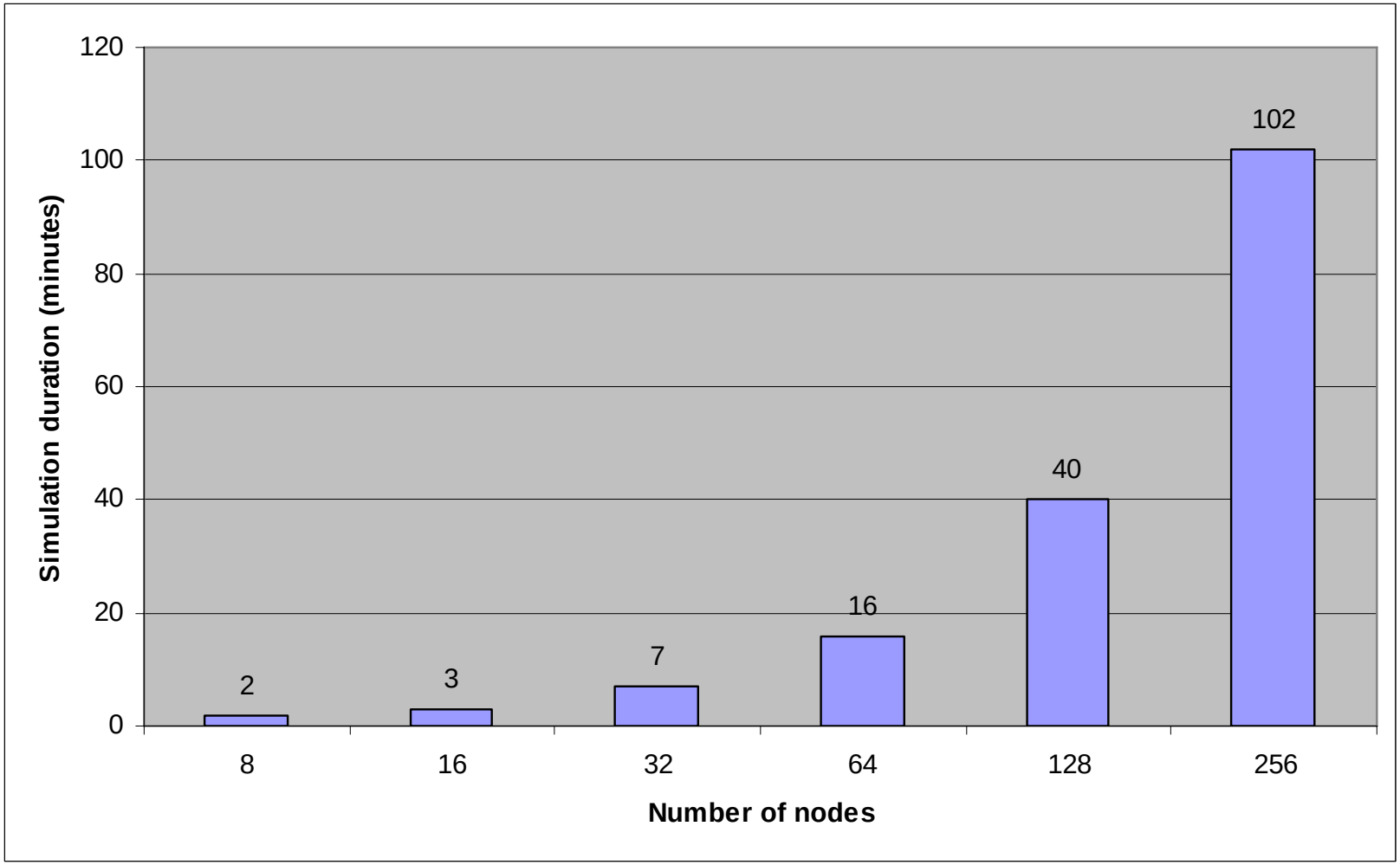
- Last week's conclusions from here still apply
  - Tough benchmark!
  - BCN_MAX is not sufficient to control this case

- BCN(0,0) provides major benefits in this case
  - No collapse of average throughput
  - Drastically reduced drop rates
  - Queue convergence even for large N
  - w/o having to retune the gains for such corner cases...

- Per-flow sampling (E^2CM)
  - Improves recovery speed and stability
    - However, ECM's recovery timer is not implemented
  - Has difficulty coping with high hotspot degree
    - E^2CM's $Q_{eq,flow}$ is not scaled down as N increases

Backup

Simulation duration per run



- Number of nodes x 2 ➔ simulation time x 2.5

Comparative Impact of BCN(0,0) on Loss w/ PAUSE Disabled