Source-based E^2CM:

Validation of the Orlando Proposal

IBM Research GmbH, Zurich

March 22, 2007

Some Concerns re. E2CM Raised in Orlando

1. Destination (DST)-based per-flow RX rate calculation (throughput accounting)
   - Preferably, the source (SRC) should handle this job

2. Global clock synchronization required for forward latency measurement
   - Too costly

Modification addressing Concern #1

1. SRC measures throughput in between probes
   1. Generally this equals the configured mean probe interval (e.g. 75 KB)
      1. May vary due to imposed interval jitter and max-interval time limit (e.g. 10 ms)
   2. Byte count $B(P_n)$ is included in probe P_n
      1. Optionally, source may store byte count locally
   3. Upon reception, DST returns probe P_n including $B(P_n)$ and records probe arrival time $T_{dst}(P_n)$ in probe P_n
   4. Upon return, SRC stores $T_{dst}(P_n)$ for this particular flow
   5. SRC computes throughput as follows: $B(P_n) / (T_{dst}(P_n) - T_{dst}(P_{n-1}))$
      1. Clock synchro is not an issue: both time stamps are recorded at DST

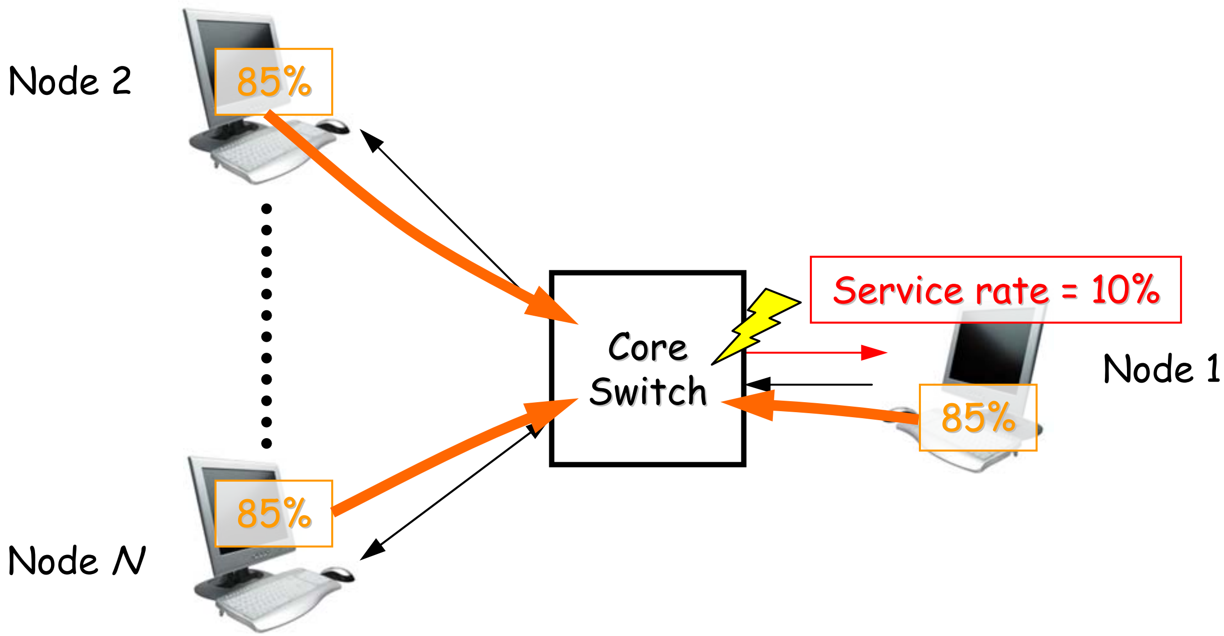
Potential demerits:

1. Does not account for dropped frames
2. Less robust to lost/corrupted probes

Modification addressing Concern #2

- Use SRC clock to determine forward latency
  - Expedite probes on reverse path
    - Use top priority traffic class
    - Switches automatically preempt other traffic for probes
  - SRC includes time stamp $T_{src}(P_n)$ in probe P_n
  - Upon return, SRC computes round-trip latency $L(P_n)$ = now - $T_{src}(P_n)$
  - SRC keeps track of minimum round-trip latency $L_0 = min_n(P_n)$
  - SRC computes effective forward latency as $L(P_n) - L_0$

Output-Generated Single-Hop Hotspot



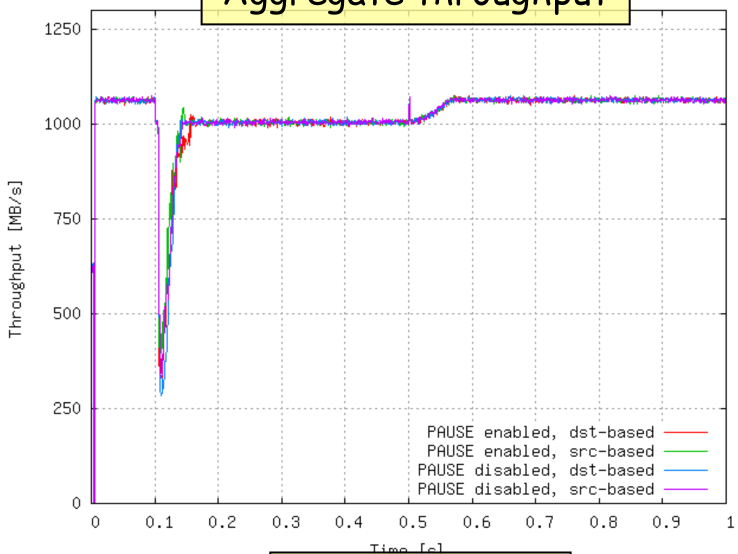
- All nodes: Uniform destination distribution, load = 85% (8.5 Gb/s)
- Node 1 service rate = 10%
- One congestion point
  - Hotspot degree = N-1
  - All flows affected

Simulation Setup & Parameters

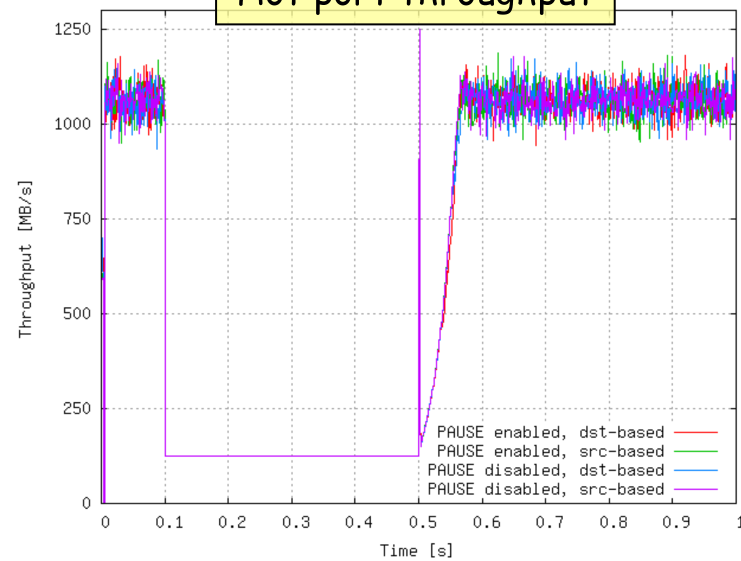
- Traffic
  - I.i.d. Bernoulli arrivals
  - Uniform destination distribution (to all nodes except self)
  - Fixed frame size = 1500 B
- Scenarios
  1. Single-hop output-generated hotspot
- Switch
  - M = 300 KB/port
  - Partitioned memory per input, shared among all outputs
  - No limit on per-output memory usage
  - PAUSE enabled or disabled
    - Applied on a per input basis based on local high/low watermarks
    - $watermark_{high}$ = 280 KB
    - $watermark_{low}$ = 260 KB
    - If disabled, frames dropped when input partition full
- Adapter
  - Per-node virtual output queuing
  - No limit on number of rate limiters
  - Unlimited ingress buffer size
  - Egress buffer size = 1500 KB
  - PAUSE enabled
    - $watermark_{high}$ = 1500 – rtt*bw KB
    - $watermark_{low} = watermark_{high}$ - 10 KB
- ECM
  - W = 2.0
  - Q_{eq} = 75 KB (= M/4)
  - $G_d = 0.5 / ((2*W+1)*Q_{eq})$
  - $G_{i0} = (R_{link} / R_{unit}) * ((2*W+1)*Q_{eq})$
  - $G_i = 0.005 * G_{i0}$
  - P_{sample} = 2% (on average 1 sample every 75 KB
  - $R_{unit} = R_{min}$ = 1 Mb/s
  - BCN_MAX enabled, threshold = 280 KB
  - No BCN(0,0), no self-increase
- E^2CM (per-flow)
  - W = 2.0
  - Q_{eq} = 15 KB
  - $G_d = 2.5 / ((2*W+1)*Q_{eq})$
  - $G_i = 0.025 * G_{i0}$
  - P_{sample} = 2% (on average 1 sample every 75 KB)
  - $R_{unit} = R_{min}$ = 1 Mb/s
  - BCN_MAX enabled, threshold = 56 KB

Results single-hop OG scenario (N = 16)

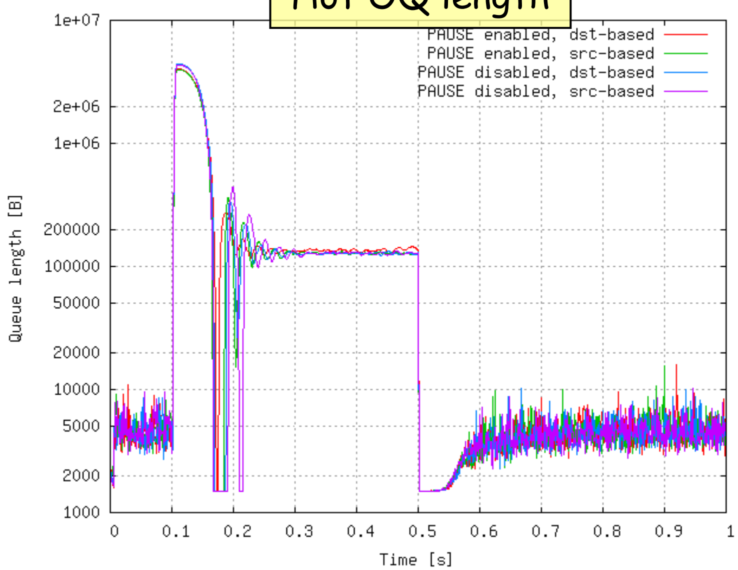
Aggregate throughput



Hot port throughput



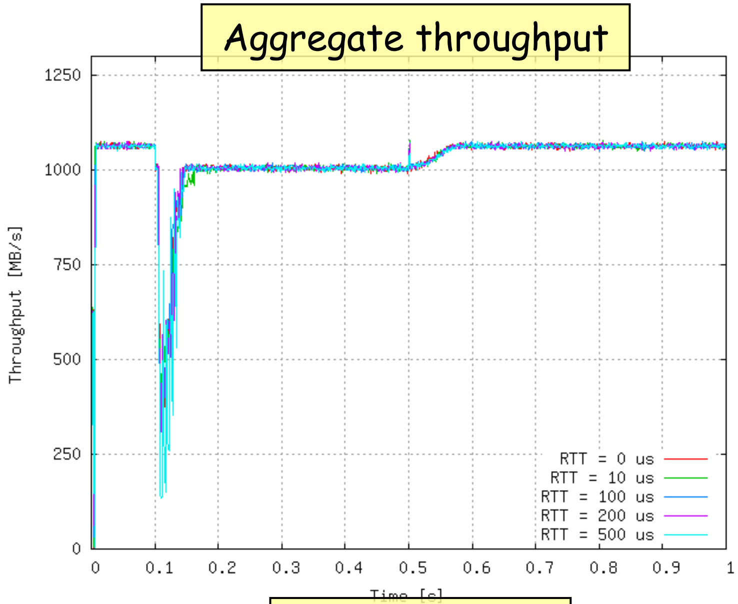
Hot OQ length

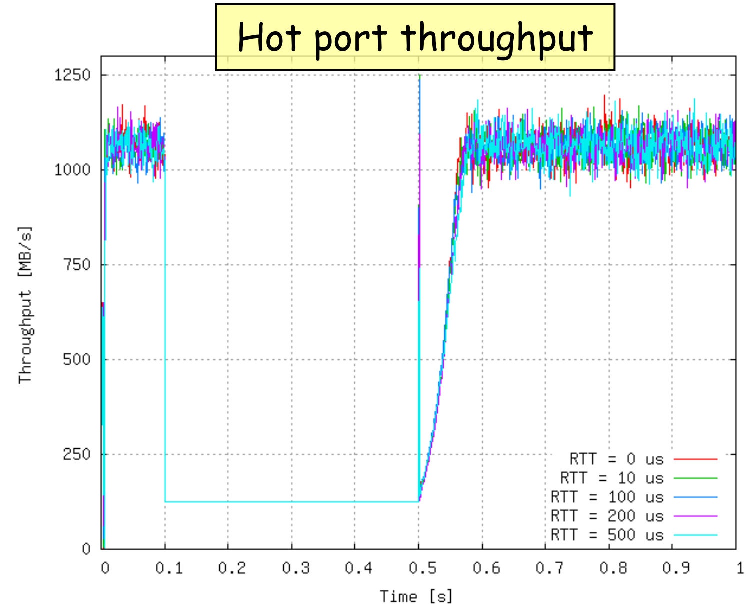


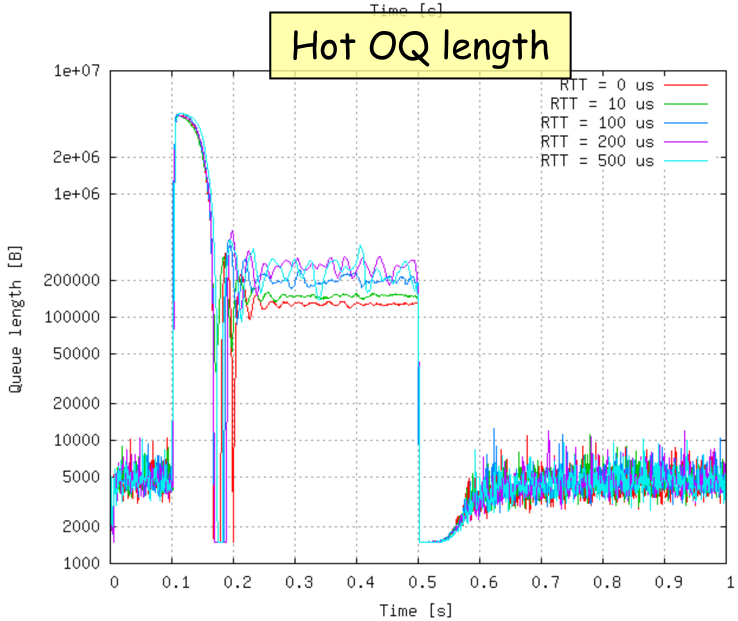
- Source- vs. destination-based (both mods 1 and 2)
- Switch PAUSE enabled/disabled
- No thresholding of OQ (unlimited within h/w boundaries)

Switch frame drops	Dst-based	Src-based
PAUSE on	0	0
PAUSE off	146,595	130,268

E2CM single-hop OG - Impact of RTT



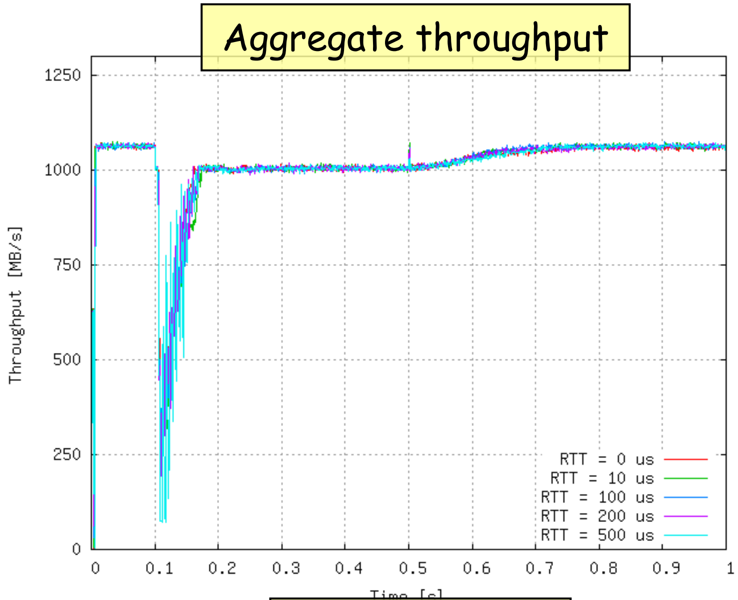


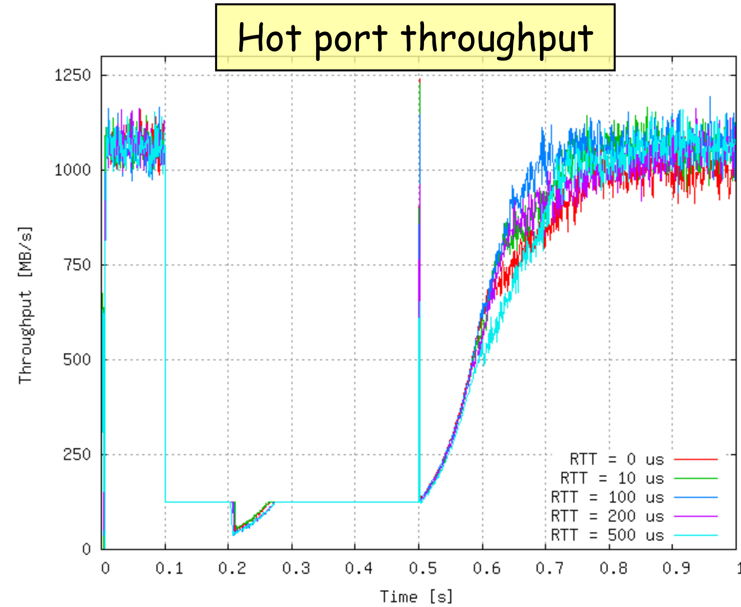


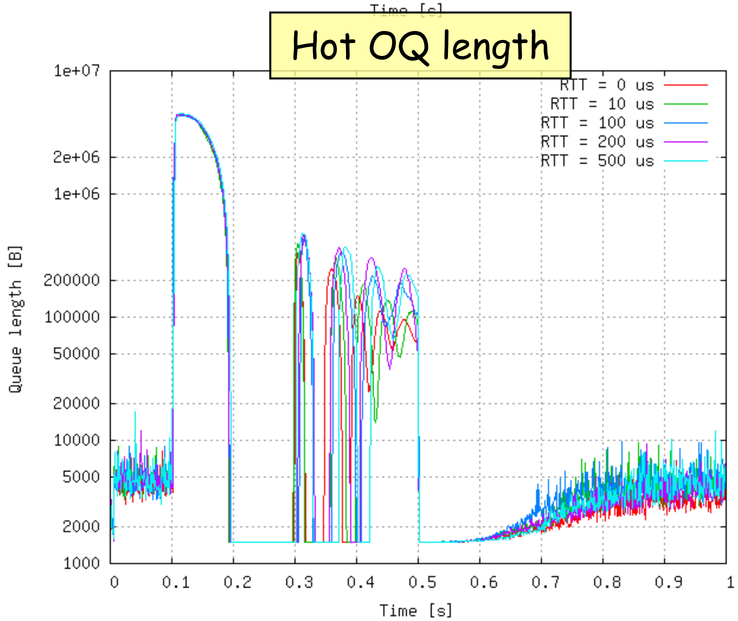
- Source-based (both mods 1 and 2)
- Switch PAUSE disabled
- Unlimited output queue length (hoggable)
- RTT = [0, 10, 100, 200, 500] μs

RTT (μs)	Switch frame drops
0	134,879
10	148,816
100	135,874
200	144,239
500	189,371

ECM results single-hop OG scenario – Impact of RTT



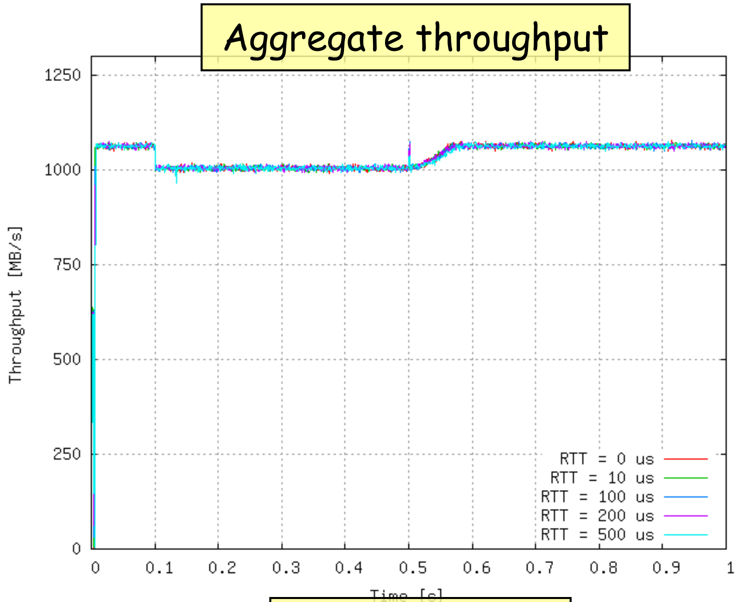


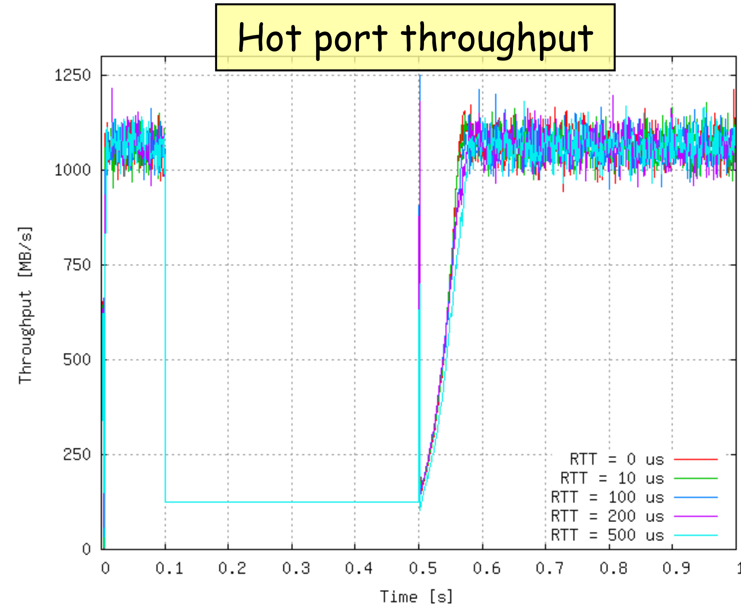


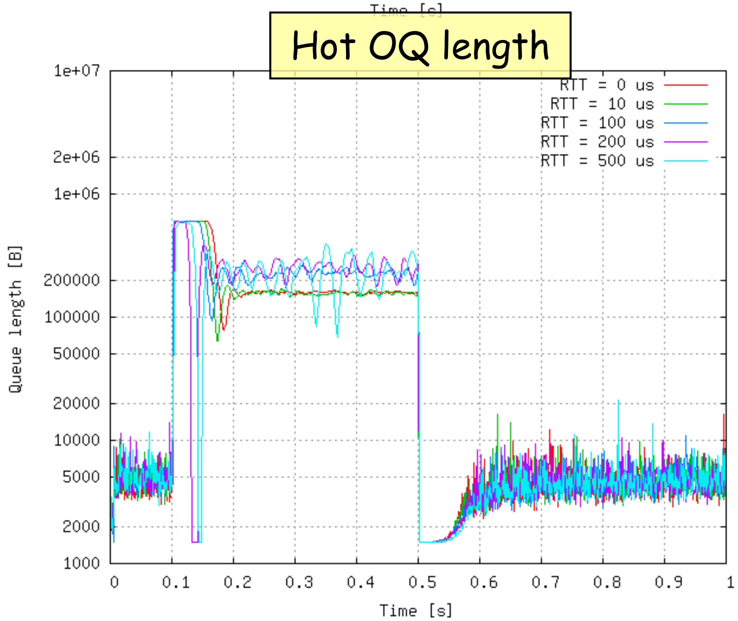
- Source-based (both modifications)
- Switch PAUSE disabled
- No limit on output queue length
- RTT = [0, 10, 100, 200, 500] μs

RTT (μs)	Switch frame drops
0	211,246
10	250,801
100	219,003
200	212,431
500	243,122

E2CM single-hop OG - OQ limit



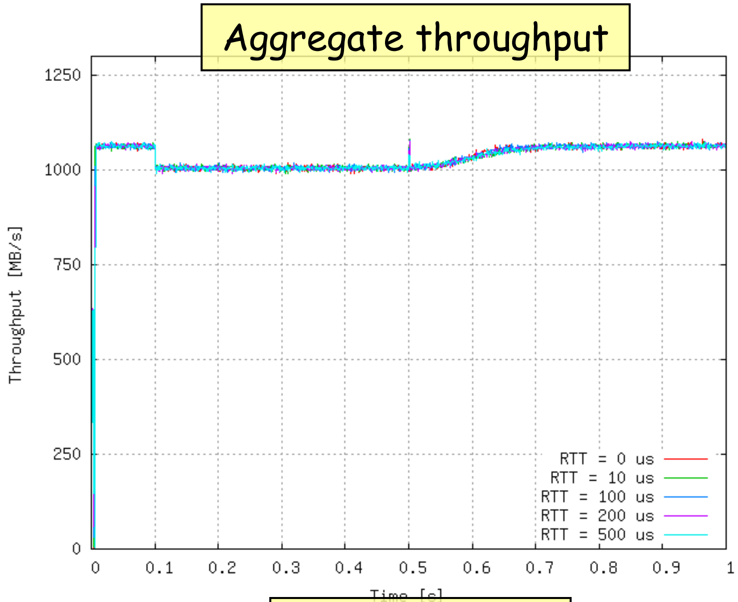


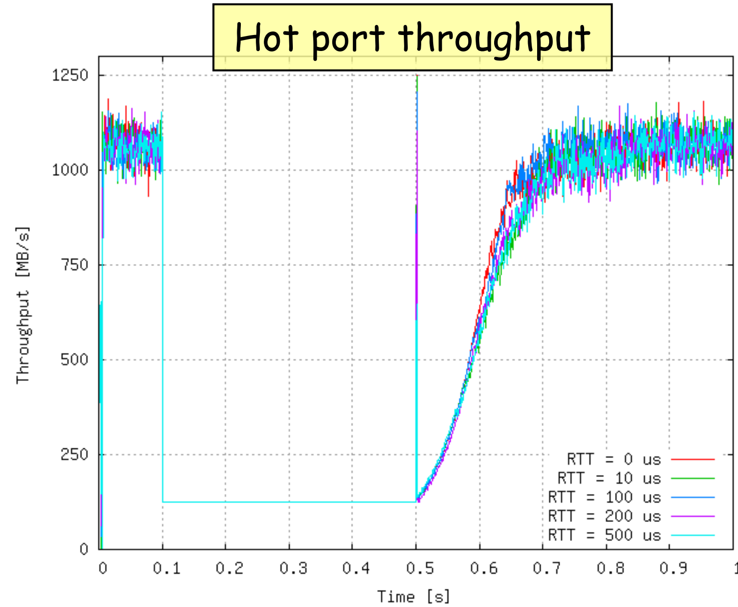


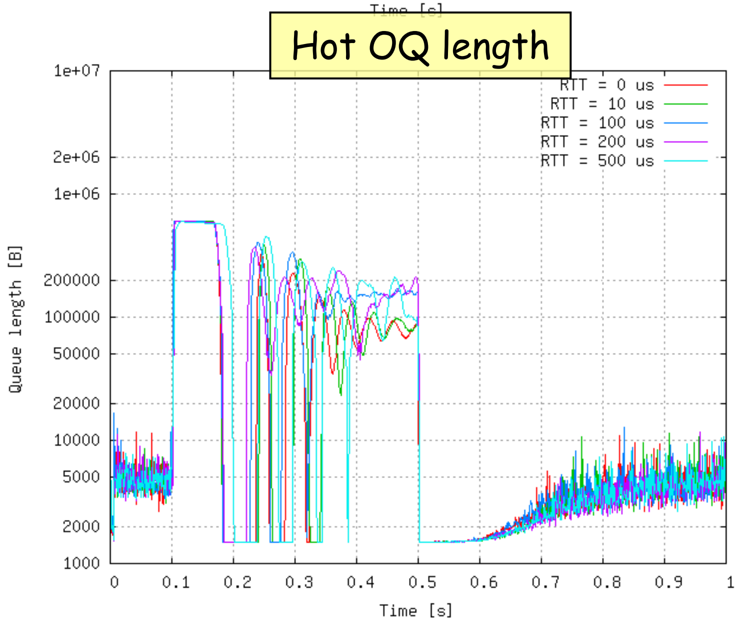
- Source-based (both modifications)
- Switch PAUSE disabled
- 600 KB limit on output queue length
- RTT = [0, 10, 100, 200, 500] μs

RTT (μs)	Switch frame drops
0	16,083
10	14,230
100	11,116
200	7,171
500	11,300

ECM results single-hop OG scenario – OQ limit







- Source-based (both modifications)
- Switch PAUSE disabled
- 600 KB limit on output queue length
- RTT = [0, 10, 100, 200, 500] μs

RTT (μs)	Switch frame drops
0	25,335
10	25,162
100	25,261
200	25,338
500	27,511

Conclusions: Pat's Orlando Proposal Works...

- Source-based and destination-based E^2CM are practically indistinguishable in terms of SH-OG performance
  - consequential for h/w implementation...
- Stability is achieved even with RTTs up to 500 μs
  - However, mean queue level increases with RTT as consequence of additional transport lag
- In PAUSE-less mode frame drops* can be significantly (~10x) reduced by using per-OQ drop threshold
  - such, or more sophisticated, partitioning is recommended

* An arguable pursuit (reducing loss rate w/o LL-FC) ...