Extended Ethernet Congestion Management (E^2CM): Per Path ECM - A Hybrid Proposal

M. Gusat, C. Minkenberg and R. Luijten

IBM Research GmbH, Zurich

March 14th 2007

Outline

- Status at 802.1
- Critique Analysis of BCN
- IBM hybrid proposal: E^2CM
  - unite BCN/ECM + recent CM proposals
- E^2CM simulations
  - A) Orientative results
  - B) Reference results
- Conclusions

Status at 802.1au

1. 802.1 WG critique of BCN performance
   1. Stability: oscillations and 'slow' convergence
   2. Fairness: BCN's rate allocation is 'unfair'

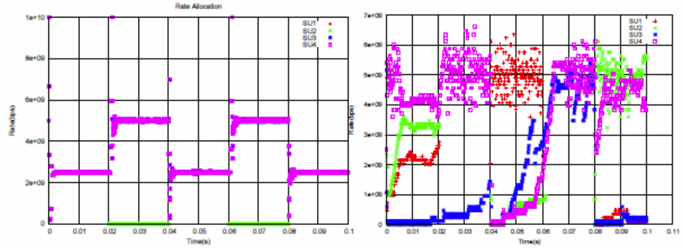
2. Recent rate-based (RB) CM proposals in 802.1
   1. BECN evolves into FECN (R. Jain)
   2. FECN-like destination-based probing / DBP (M. Seaman)

3. First reconciliation proposal: QCN (B. Prabhakar)
   - optimized BCN
     - same or better performance at *lower overhead*

- 802.1au agreement on CM method appears difficult...
  - ✓ adhoc simulation progress slowdown

Critique of BCN vs. ECN: Of Fairness, Stability and Speed

Symmetric Topology: Source Rates Rcvd



ECN BCN

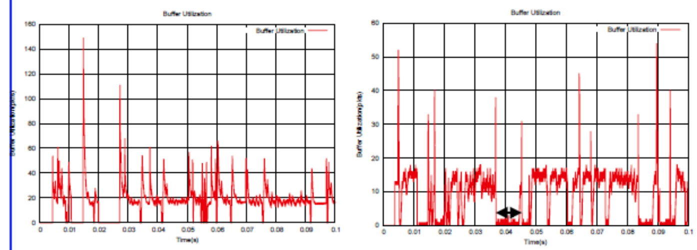
Conclusions:

1. ECN converges very fast and remains stable.
2. Perfect fairness results in only two visible curves.
   Note that ECN graphs have 4 curves.
3. Convergence time is a small multiple of measurement interval.

Washington University in St.Louis IEEE 802.1 Meeting Nov 14, 2006 Raj Jain

31

Bursty Traffic: Queue Lengths



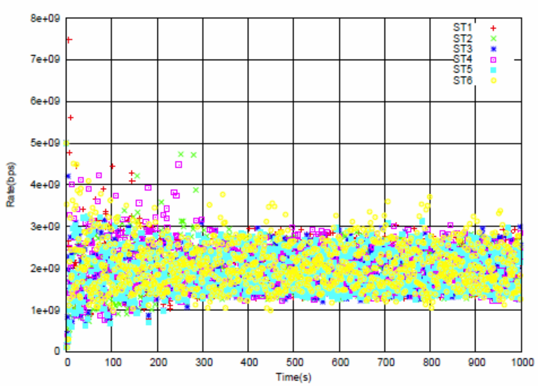
ECN BCN

- **Conclusion**: ECN has less chances of zero queue => Higher link utilization

Washington University in St.Louis IEEE 802.1 Meeting Nov 14, 2006 Raj Jain

44

Parking Lot: Rates for BCN

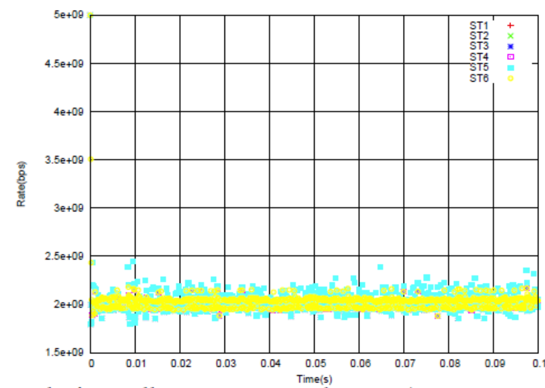


- Large Oscillations

Washington University in St.Louis IEEE 802.1 Meeting Nov 14, 2006 Raj Jain

36

Parking Lot: Source Rates for ECN



- Conclusion: All sources get 2 Gbps = C/5 => MAX-MIN Fairness

Washington University in St.Louis IEEE 802.1 Meeting Nov 14, 2006 Raj Jain

35

Current CM Proposals:
R. Jain's FECN + B. Prabhakar's QCN + M. Seaman's DBP

A) R. Jain presents in Monterey a new rate-based CM (RB-CM) proposal, i.e. FECN

- Performance claims (top 3):
  1. "Perfect Fairness" [Obs.: here "perfect" stands for max-min]
  2. "Fast Convergence" [Obs.: 10ms here]
  3. "No PAUSE required or issued" [Obs.: in certain cases]

Note: FECN results are to be validated by other simulation teams.

B) M. Seaman's DST-based probing (DBP) proposal

"...algorithm comprising three sets of algorithms for Sources, Bridges, and Destinations.

1. the Destination originates and transmits regular Rate Report (RR) frames to each active Source.
2. Each RR traces the reverse path from the Destination to the Source and carries an advertised rate for use by the Source in transmitting to that Destination. The RR originally carries a rate set by the destination to be its receiving link speed. "

C) B. Prabhakar's insight leading to the QCN proposal

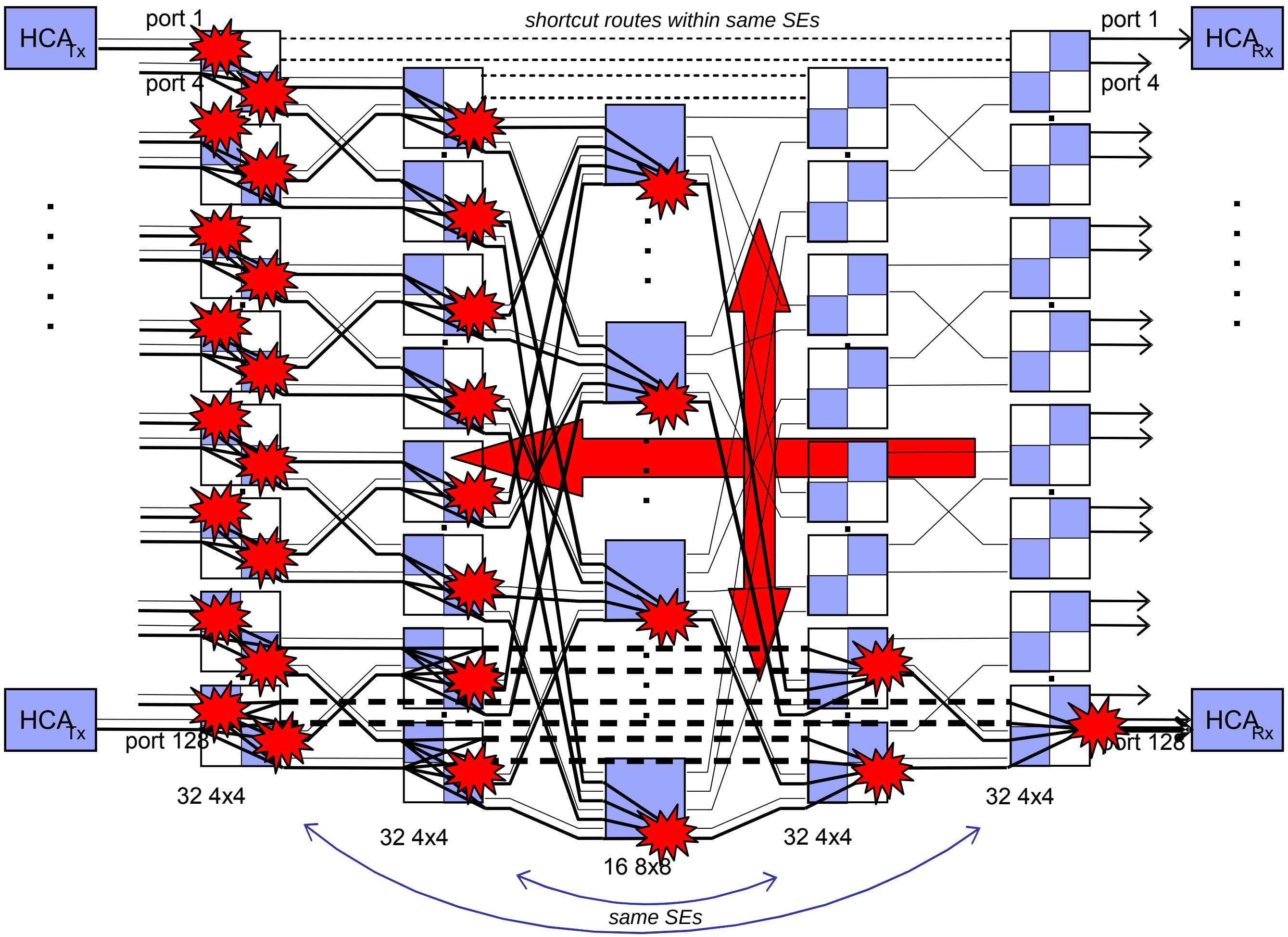
"BCN generates extra signaling traffic

- Hence sampling probability is kept at 1%; this can go up to 10% and improve responsiveness by a lot
- But, if [i] forward signaling is possible, or [ii] another means of signaling ***more frequently*** can be found, then we can send less information per signal"

IBM Proposal: BCN + FECN + DBP + QCN => E^2CM

The Hybrid Dilemma: How to Combine the Best Features ... Only?

Tree Saturation => Complex CM ... yet leads to our proposal (PPT animation)

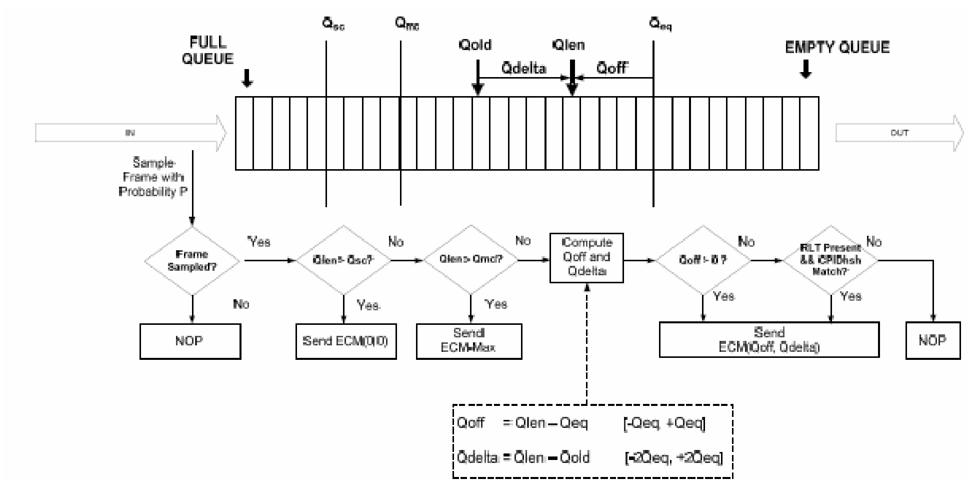


- Documented in papers from sim and h/w experiments [refs available]
  - Link-level flow control induces blocking chains in depth and breadth
  - Few hot flows hog resources (high-order HOL-B) => blocking many/all cold/victim flows
- Effect: nonlinear (saturation chain) and time-varying system (2 non-invariants)
  - Impact on piecewise linearization, particularly at linearization points
- Transport lag => transcendental characteristic equation
  - Root-locus and Routh-Hurwitz apply only to rational transfer functions...

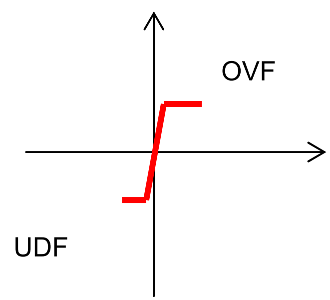
Critique – 1 Analysis : Stability of BCN

- A key observation: Baseline BCN has robust performance in the linear region!
  - However, its dynamic range (DR) is limited by the queue capacity
  - Furthermore, possibly fed by n simultaneous arrivals...

from ECM Spec



Saturated Integrator behavior...



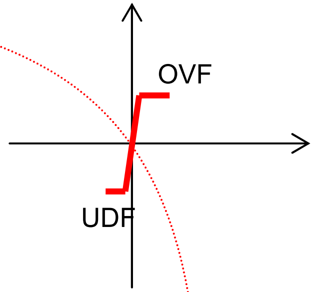
$\Rightarrow$ Feedback: $F_b(t) = -(q(t) - Q_{eq}) + w*(dq/dt) / (\mu_j * p_s)$ =>

$$0 \le q(t) \le q_{max}$$

- Fast transition between lower/upper saturation (n+1 stochastic procs)
  - requires frequent use of saturation signals: BCN_Max, BCN(0,0)
  - non-linear saturation patches reduce the efficiency of the baseline control alg.

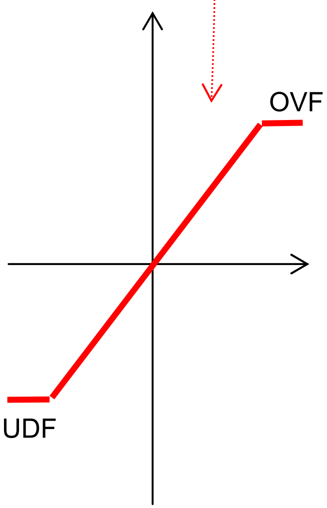
Extending the Linear Region of a Saturated Integrator

- How to scale BCN's stability properties w/ network size?
  1. Increase the dynamic range by chaining the j queues along the path i...
  2. Control the chain of queues instead of the individual queue



  ➢ **Adopt per path probing ...**

- Concatenate multiple queues along a path into a Path Queue -> Σq_{ij}



  ➢ State equations

  From local queue stability to per path stability:

LQueue) $dq/dt = HSD*\lambda(t) - \mu_j$, where

$\max(HSD) = N$, and $\max(\mu_j) = C_j$

PQueue) $Q'_{ij} = \Sigma_i dq_{ij}/dt = \Sigma_i \lambda_{ij}(t) - \mu_j$, $1 < i \le HSD$

Obs.: Slope steepness decreases: from n+1 to 2 stochastic procs

Critique - 2 Analysis : Fairness of BCN

- BCN's 'unfairness': from probabilistic sampling of aggregate occupancy
  - Queue contains frames from any flow => lack of per-flow state sensing
    - Per-flow state in the bridge is prohibited by PAR
- A) One approach is to calculate / iterate the fair share (FS) as in ABR methods such as FECN, UT, OSU, H2
  - TBD: scalability and h/w complexity
- B) Other methods were proposed by Stanford Univ. [Allerton paper]
- C) IBM proposal: A per-flow probing sensor in the edge node...
  - Probing: triggered by BCN, or autonomous (congestion avoidance)
  - Why per flow? -> fast max-min convergence even w/o FS in the bridge

E^2CM Principles: Dual Heritage

- E^2CM includes saturation tree mgnt: avoidance, isolation and recovery.
  - Historically this wasn't the case in TCP and ATM ABR

I) BCN and QCN heritage: E^2CM draws upon the baseline BCN

- PD control + feedback equation (extended, instead of piecewise linearized)
- AIMD-based SRF + parameters
  - ✓ Potential QCN optimizations:
    - Bridges do not send increase signals
    - Sparse quantization (6/5 bit)

II) **FECN and DBP** heritage: SRC probes for RTT + DST calculates rate

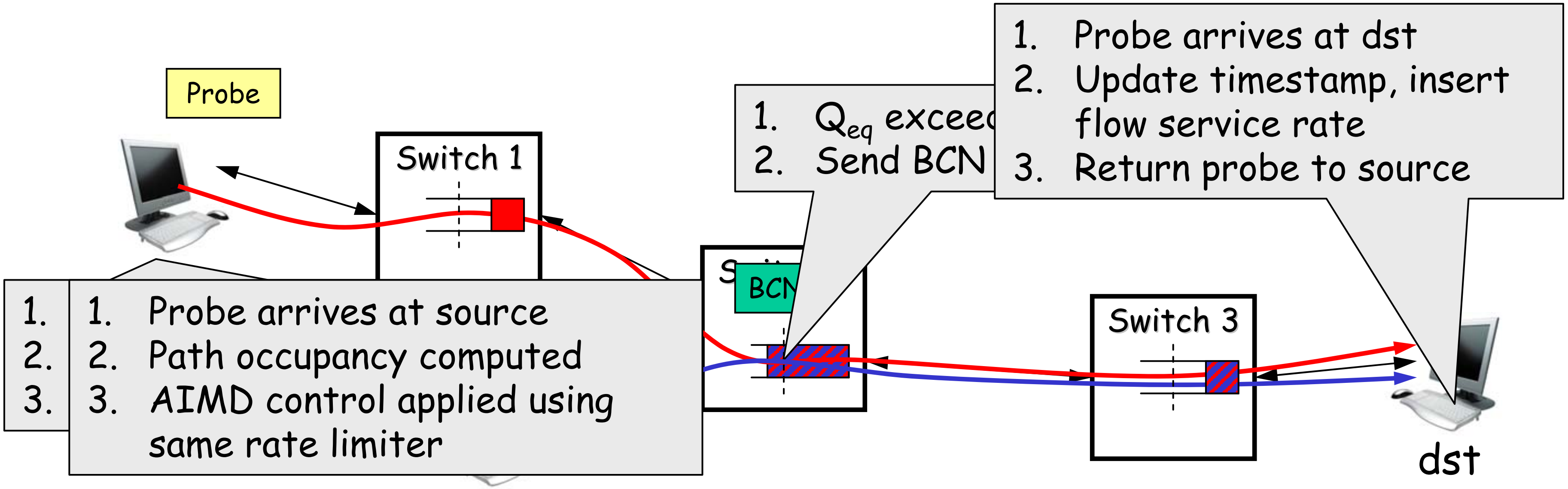
- from FECN and DBP we adopt per path probing and DST rate reporting

Effect: E^2CM extends BCN's dynamic range to a wider linear region proportional to the number of buffers traversed per path

- increased stability and phase margin (improved convergence)
- scalability w/ network size

- By adopting per-flow accounting in the end node, E^2CM converges to fair allocation rates w/o using rate-based calculations in the bridge

E^2CM Operation (PPT animation)



- Probing is triggered by BCN frames; only rate-limited flows are probed
  - Insert one probe every X KB of data sent per flow, e.g. X = 75 KB
  - Probes traverse network inband: Objective is to observe real current queuing delay
- Per flow, BCN and probes employ the same rate limiter
  - Control per-flow (probe) as well as per-queue (BCN) occupancy
  - CPID of probes = destination MAC
  - Rate limiter is never associated with probe CPID
  - Parameters re. probes may be set differently (in particular $Q_{eq,flow}$, $Q_{max,flow}$, $G_{d,flow}$, $G_{i,flow}$)

Proposed Compromise Scheme details

- **Extension: BCN reception triggers RTT and T_{put} probing in the end nodes**

  - A) While activated **SRC** periodically (t or n-pkts) insert probe frame for every RLT flow
    - ✓ Probe contains timestamp
    - ✓ Probes traverse network in-band with regular data frames

  - B) Upon reception of forward probe, the **DST** will
    1. Update timestamp to reflect forward latency *L*
    2. Calculate and report the flow service rate *R* since last flow probe
    3. Return probe to sending SRC

  - C) Upon reception of reverse probe back at the originating **SRC**
    1. Adjust latency *L* for flight time L_0
    2. Apply Little's Formula: **$Q = (L - L_0)*R$**
        1. Yields the mean number of bytes of probed flow stored on entire forward path
    3. Apply the extended BCN source response function
        1. E.g., set $Q_{equilibrium}$ and apply AIMD rate adjustment
    - ✓ One rate limiter per flow
    - ✓ Associated with last negative feedback

  - Net: E²CM extends buffer occupancy per path and flow (however, CM triggering may be done on rate –instead of size– thresholds)
    - ✓ per flow: Separates hot from cold flows
    - ✓ per path: extends the region of linear BCN operation

Reaction Point

```
if (bcn.type() == BCN_BCN) {
  // Compute BCN reaction as usual
  ...
} else if (bcn.type() == BCN_PROBE) {
  // Store minimum latency as time of flight
  if (flightTime > bcn.getLatency() || flightTime == 0.0)
    flightTime = bcn.getLatency();

  // Compute amount of data queued on forward path, adjusting for flight time
  flowQ = bcn.getThroughput()*(bcn.getLatency() - flightTime);
  flowdQ = max( min( flowQ - flowLastQ, 2*flowQeq ), -2*flowQeq );
  flowQoff = max( min( flowQeq - flowQ, flowQeq ), -flowQeq );

  if (flowQ > flowQmax) // Qmax threshold exceeded?
    feedback = -(1+2*W)*flowQeq; // Apply maximum negative feedback
  else
    feedback = (flowQoff - W*flowdQ); // Compute feedback

  flowLastQ = flowQ; // Store last queue estimate

  // Apply AIMD rate adjustment
  if (feedback > 0) // Additive increase
    rate = rate + flowGi*feedback*rateUnit;
  else if (feedback < 0) // Multiplicative decrease
    rate = rate * (1.0 + flowGd*feedback);
}
// If needed, instantiate new rate limiter or update rate
// Associate rate limiter with CP if feedback < 0 and not probe
...
```

E^2CM Frame Format

- General fields
  - Congestion point MAC
    - Inserted by congestion point
    - Probe congestion point MAC = flow destination MAC
  - Flow identifier
    - Hash based on src MAC, dst MAC, priority
  - BCN type
    - BCN, BCN_MAX, BCN_ZERO, BCN_PROBE_FWD, BCN_PROBE_REV

- BCN-specific fields
  - Queue offset
    - Inserted by congestion point
  - Queue delta
    - Inserted by congestion point

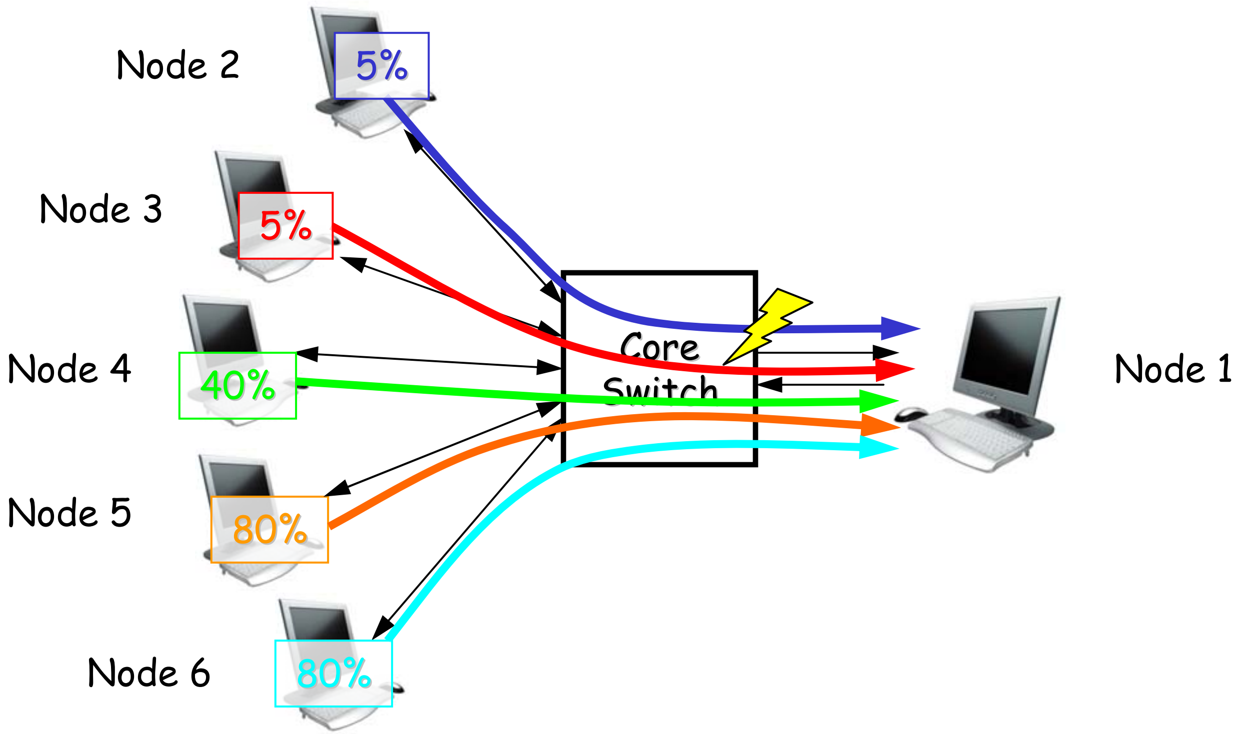
- Probe-specific fields
  - Forward latency
    - Already provided in original BCN format (but different usage)
    - Timestamp inserted by source node
    - Updated by destination node (latency = now – timestamp)
  - Flow throughput field
    - Inserted by destination node
    - Measured between two subsequent probes for same flow

* Red: new fields

E^2CM: Selected Initial Simulation Results

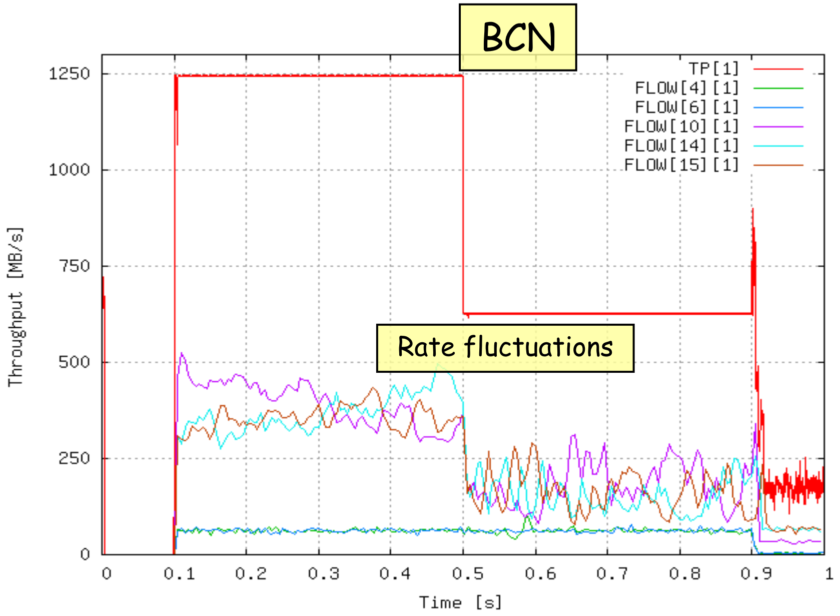
Baseline IG and OG simulations: Orientative Preview
(not for reference)

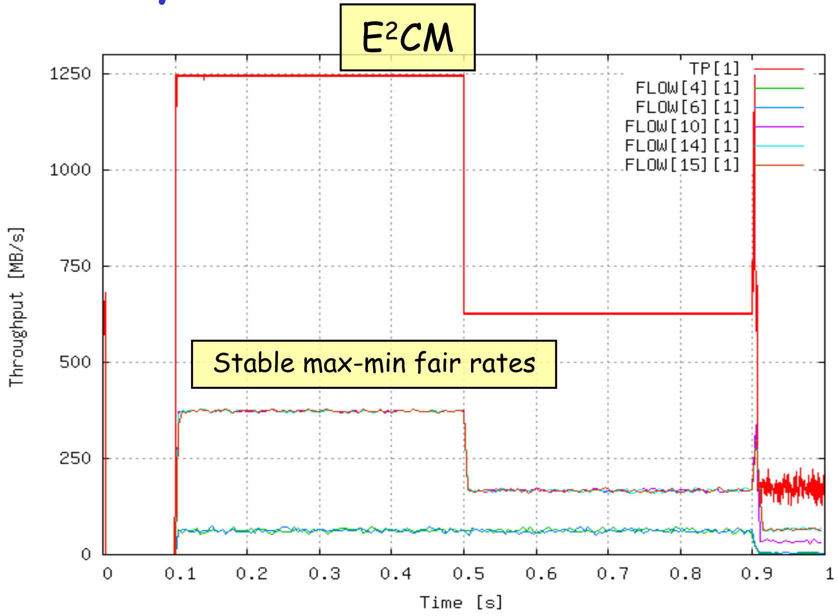
Input-generated Hotspot



- 5 flows sending to hotspot: aggregate load = 21 Gb/s
- Max-min fair rates = (0.5; 0.5; 3; 3; 3) Gb/s = (62.5; 62.5; 375; 375; 375) MB/s
- Hotspot starts at t = 0.1s; at t=0.5s, service rate of node 1 is reduced by half; fair rates = (62.5; 62.5; 167; 167; 167) MB/s

BCN vs. E²CM : Fair and Steady Rate Allocation



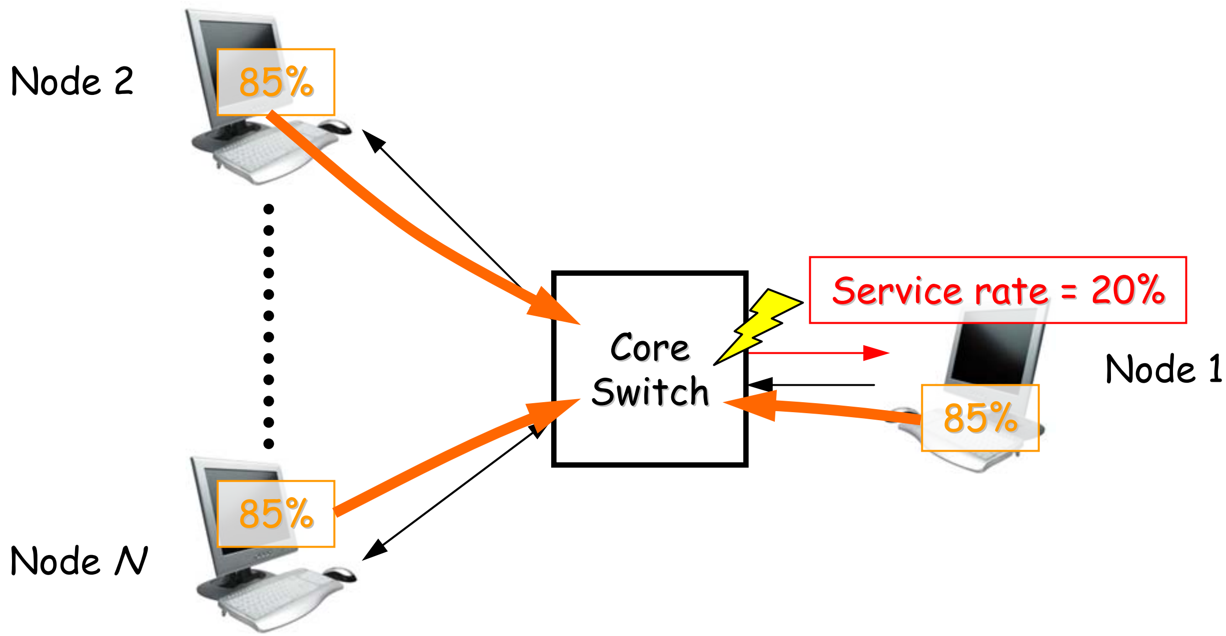


- Graphs show aggregate (red) and per-flow throughput

- Params

1. Qeq_BCN = 75 kB, Qeq_E2CM = 15 kB, Gd = 1.333*10^-6, Gi = 6.6667*10^-4
2. (Gdf = 0.5, Gif = 0.1; E2CM gains are 5 times as high, because Qeq is 5 times as low)
3. Ru = Rmin = 250 kB/s = 2 Mb/s
4. M = 300 kB/port, Thr_hi = 295500, Thr_lo = 147750
5. sample interval = 75 kB (for BCN as well as E2CM), 15 kB for rate-limited flows
6. BCN_MAX disabled

Output-Generated Single-Hop Hotspot

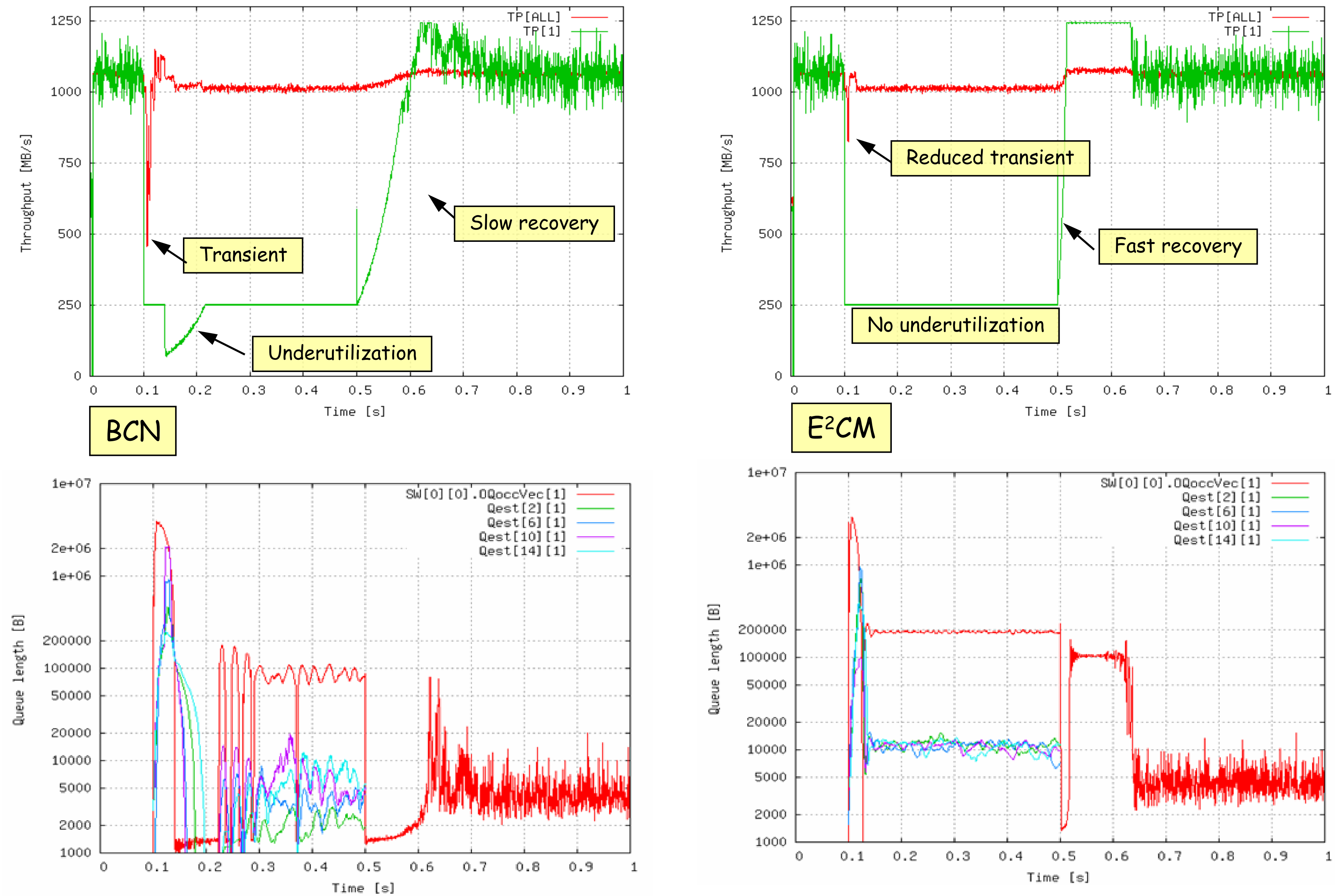


- All nodes: Uniform destination distribution, load = 85% (8.5 Gb/s)
- Node 1 service rate = 20%
- One congestion point
  - Hotspot degree = N-1
  - All flows affected

- Params

1. Qeq_BCN = 75 kB, Qeq_E2CM = 15 kB, Gd = 1.333*10^-6, Gi = 6.6667*10^-5
2. (Gdf = 0.5, **Gif = 0.01**; E2CM gains are 5 times as high, because Qeq is 5 times as low)
3. Ru = Rmin = 250 kB/s = 2 Mb/s
4. M = 300 kB/port, Thr_hi = 295500, Thr_lo = 147750
5. sample interval = 75 kB (for BCN as well as E2CM), 15 kB for rate-limited flows
6. BCN_MAX enabled (Qsc = 280500)

BCN vs. E²CM : Output-generated (Tp and Q)



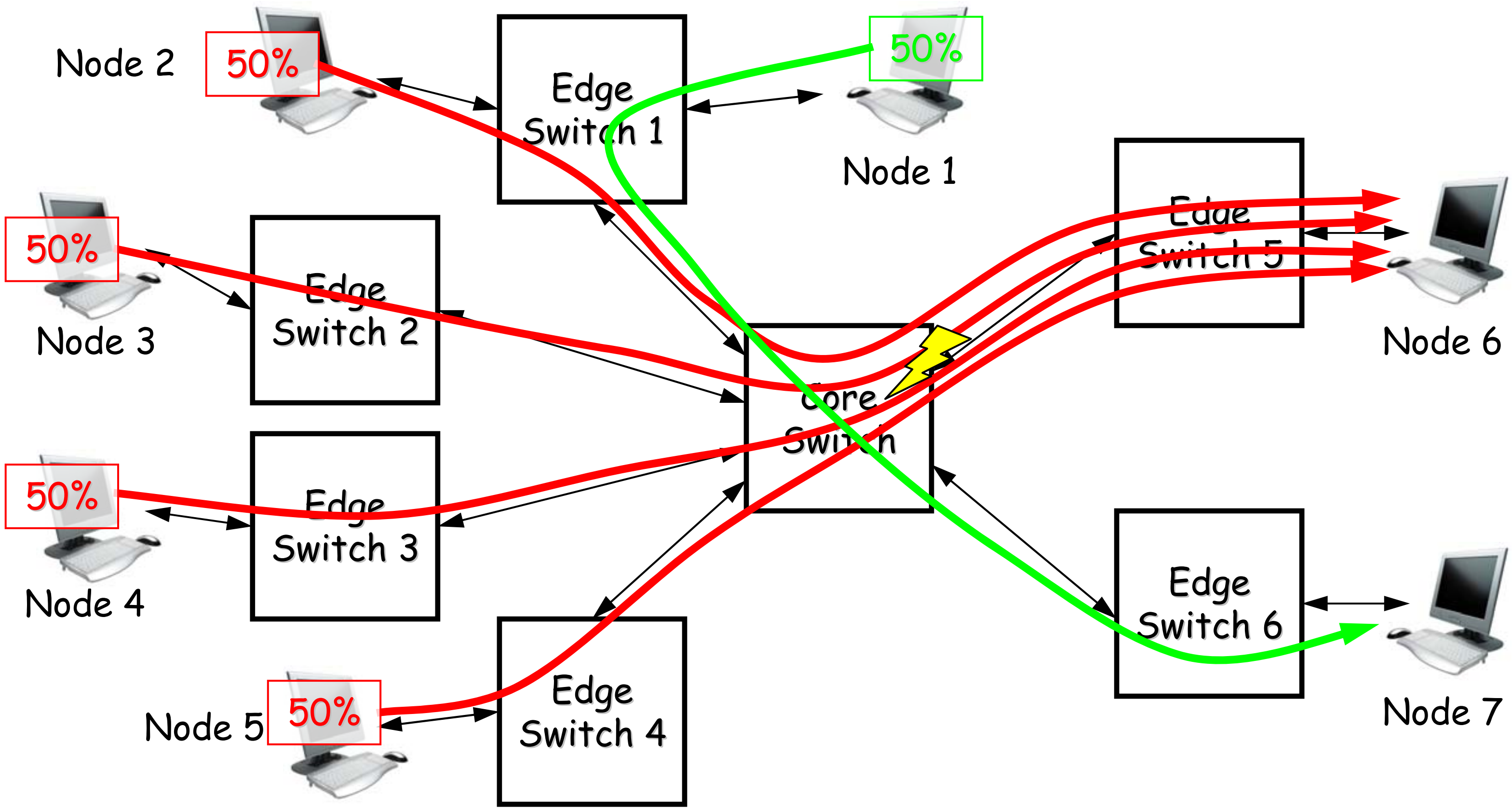
Reference Simulation Results

E2CM r1.0

Simulation Setup & Parameters

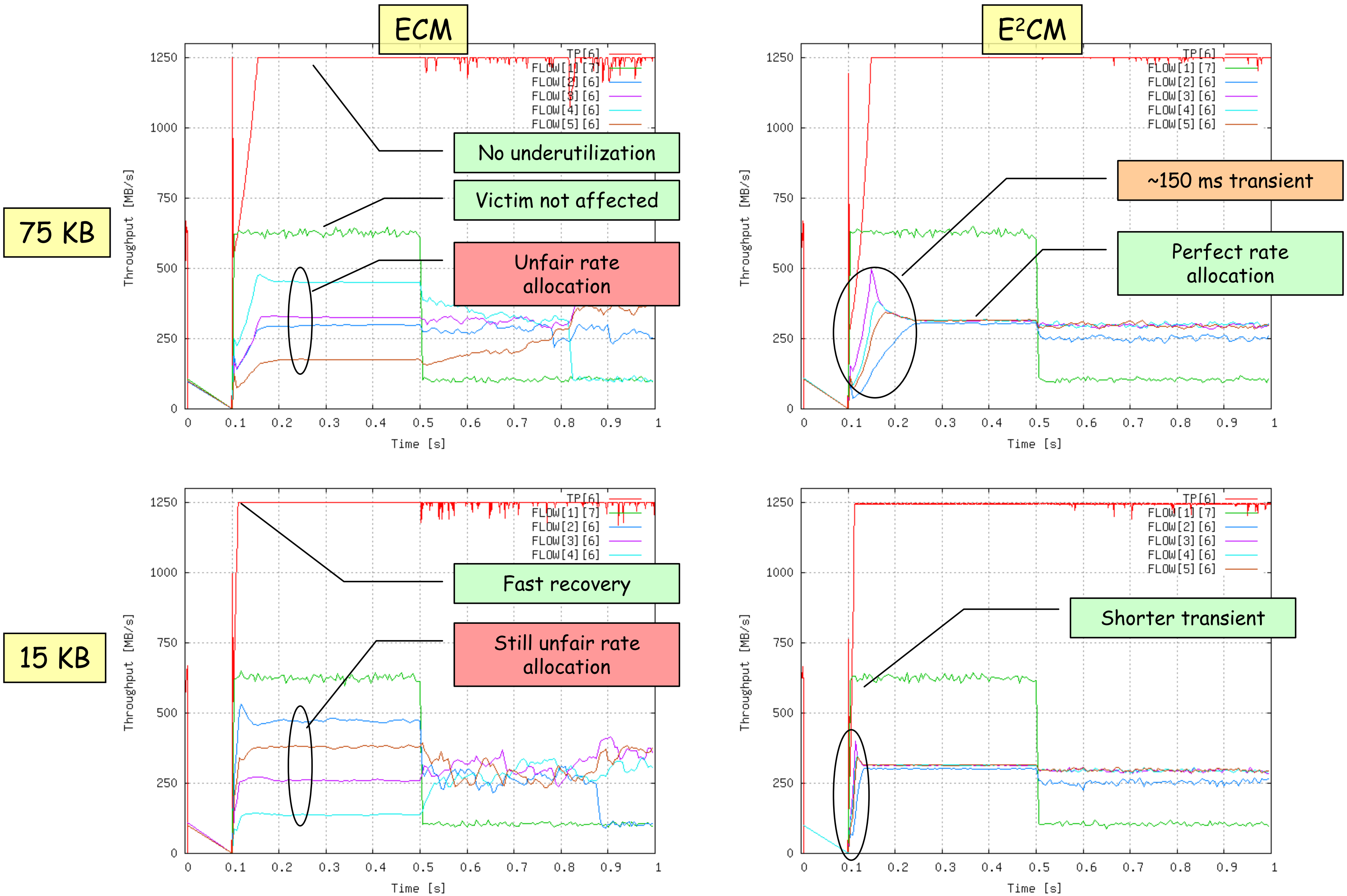
- Traffic
  - I.i.d. Bernoulli arrivals
  - Uniform destination distribution (to all nodes except self)
  - Fixed frame size = 1500 B
- Scenarios
  1. Baseline input-generated (IG)
  2. Max-min (mice-elephant) IG
  3. Single-hop output-generated (OG)
  4. Multi-Hop OG background HS
  5. Bursty On-Off
  6. Parking lot
- Switch
  - M = 300 KB/port
  - Partitioned memory per input, shared among all outputs
  - No limit on per-output memory usage
  - PAUSE enabled
    - Applied on a per input basis based on local high/low watermarks
    - $watermark_{high}$ = 280 KB
    - $watermark_{low}$ = 260 KB
- Adapter
  - Per-node virtual output queuing
  - No limit on number of rate limiters
  - Unlimited ingress buffer size
  - Egress buffer size = 150 KB
  - PAUSE enabled
    - $watermark_{high}$ = 140 KB
    - $watermark_{low}$ = 130 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) or 10% (15 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) or 10% (15 KB)
  - $R_{unit} = R_{min}$ = 1 Mb/s
  - BCN_MAX enabled, threshold = 56 KB

1. Baseline Input-Generated Hotspot



- Four culprit flows of 5 Gb/s each from nodes 2, 3, 4, 5 to node 6 (hotspot)
- One victim flows of 5 Gb/s from node 1 to node 7
- Fair allocation provides 2.5 Gb/s to all culprits and 5 Gb/s to the victim

Results Baseline scenario (Tp)

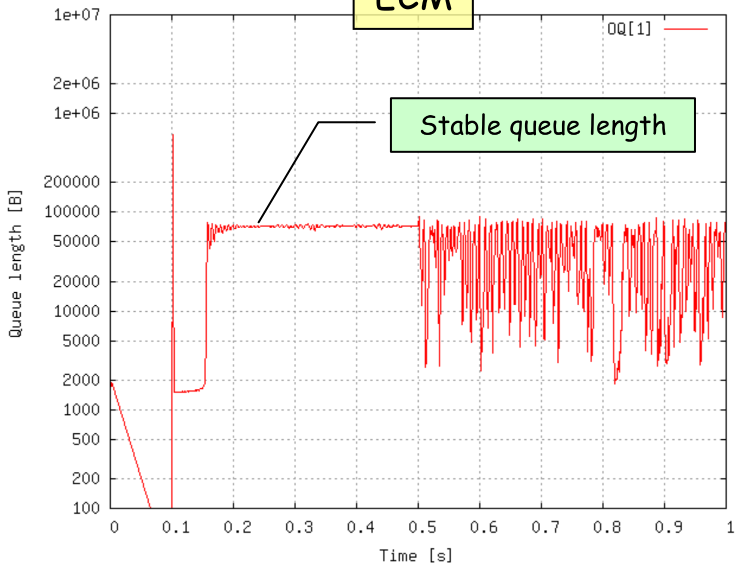


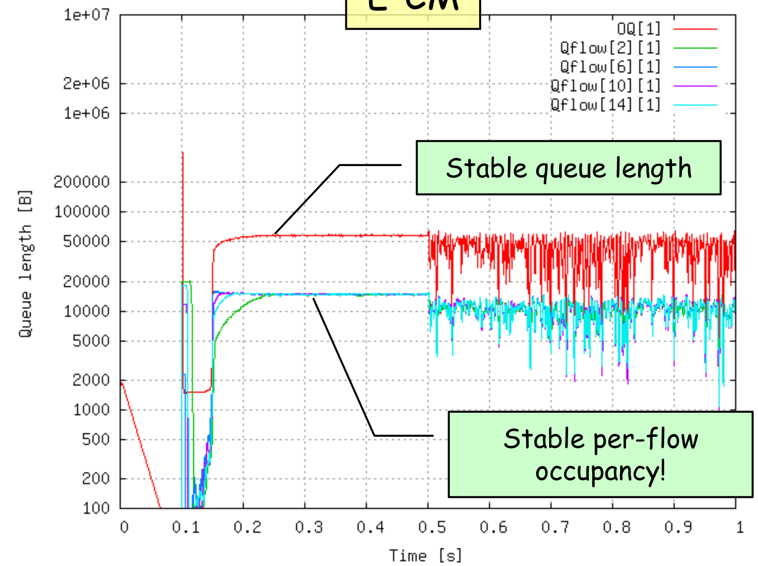
Results Baseline scenario (Q)

ECM

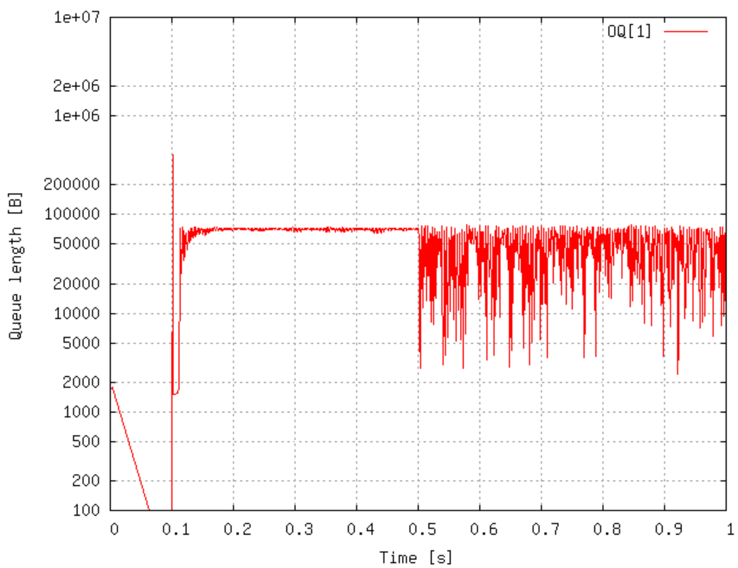
E²CM

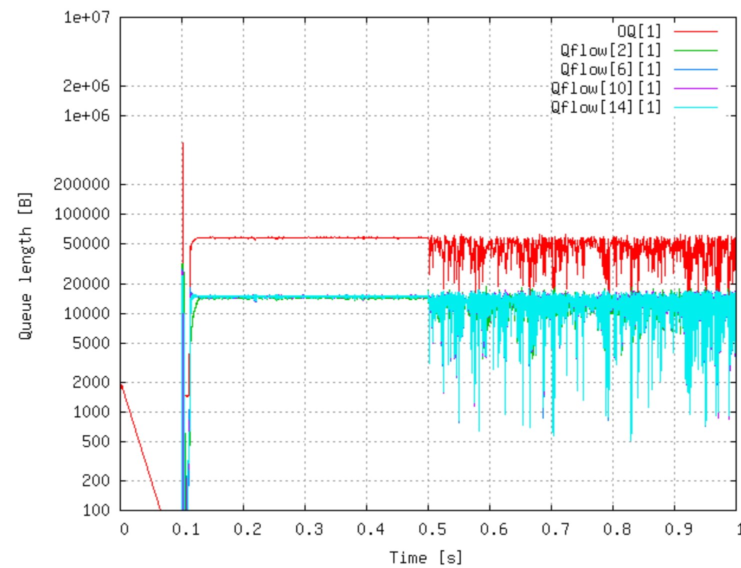
75 KB



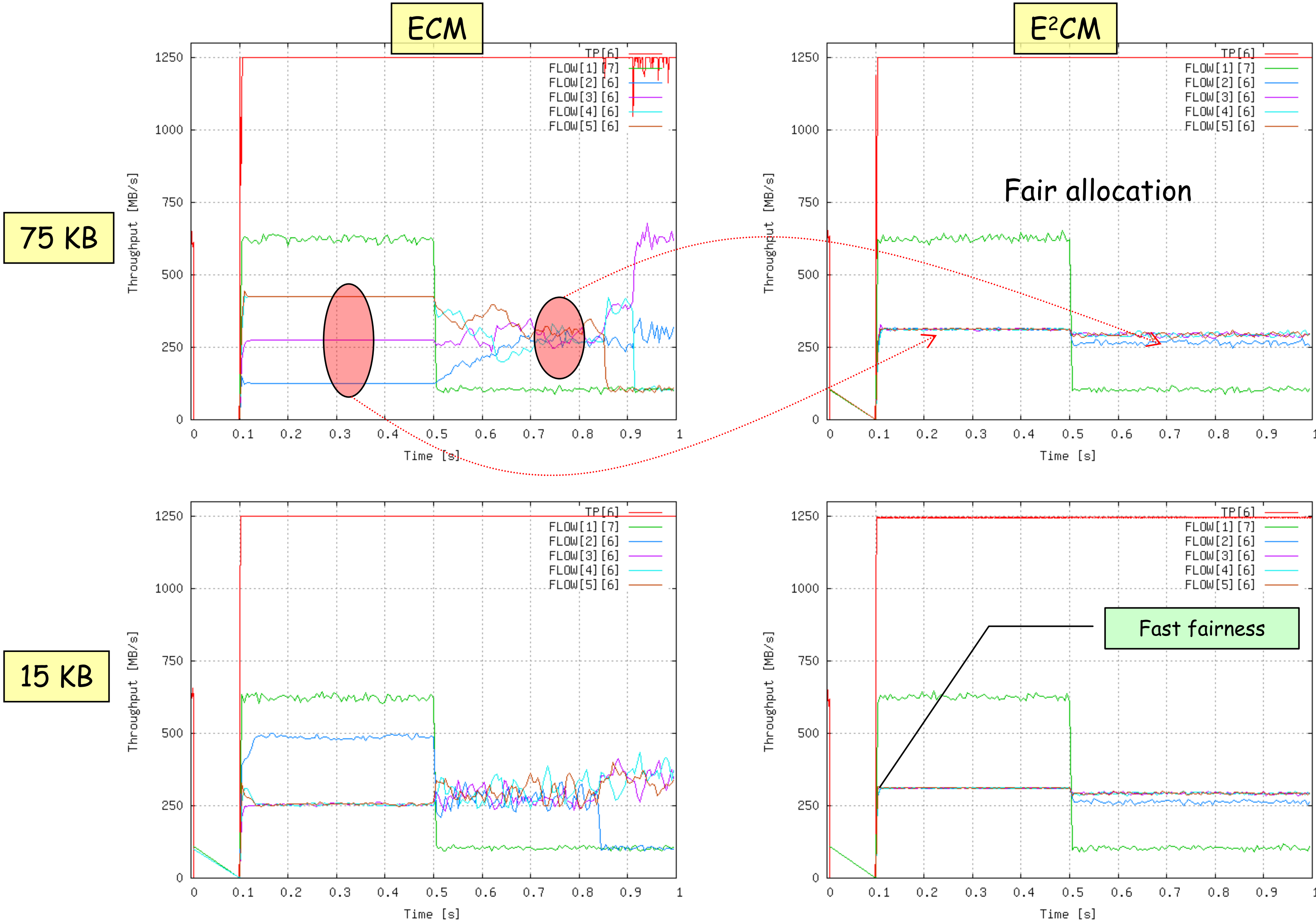


15 KB



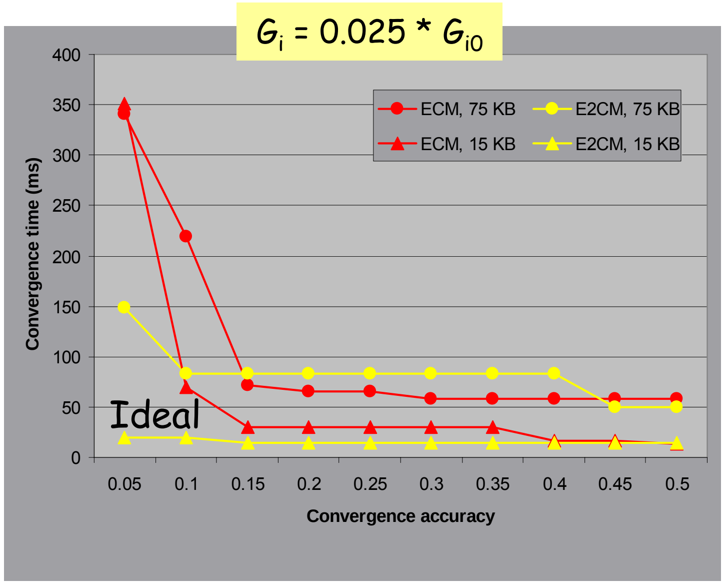


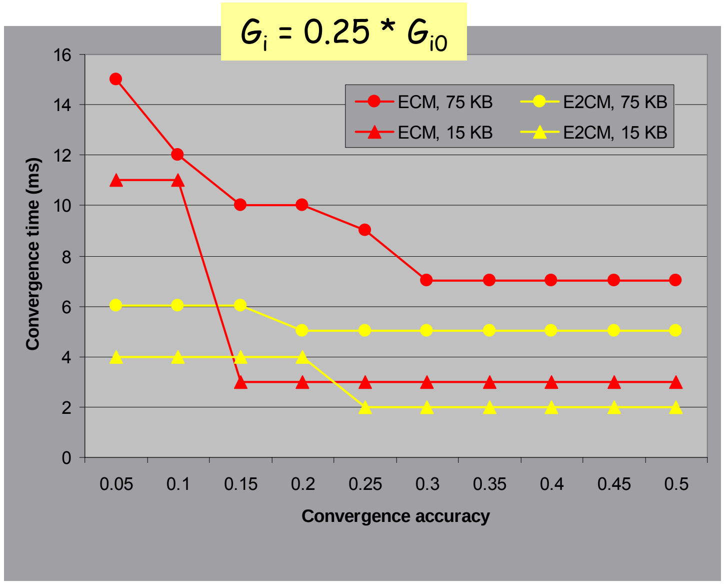
Results Baseline scenario (Tp, $G_i = 0.25*G_{i0}$)



Convergence times IG single-hop

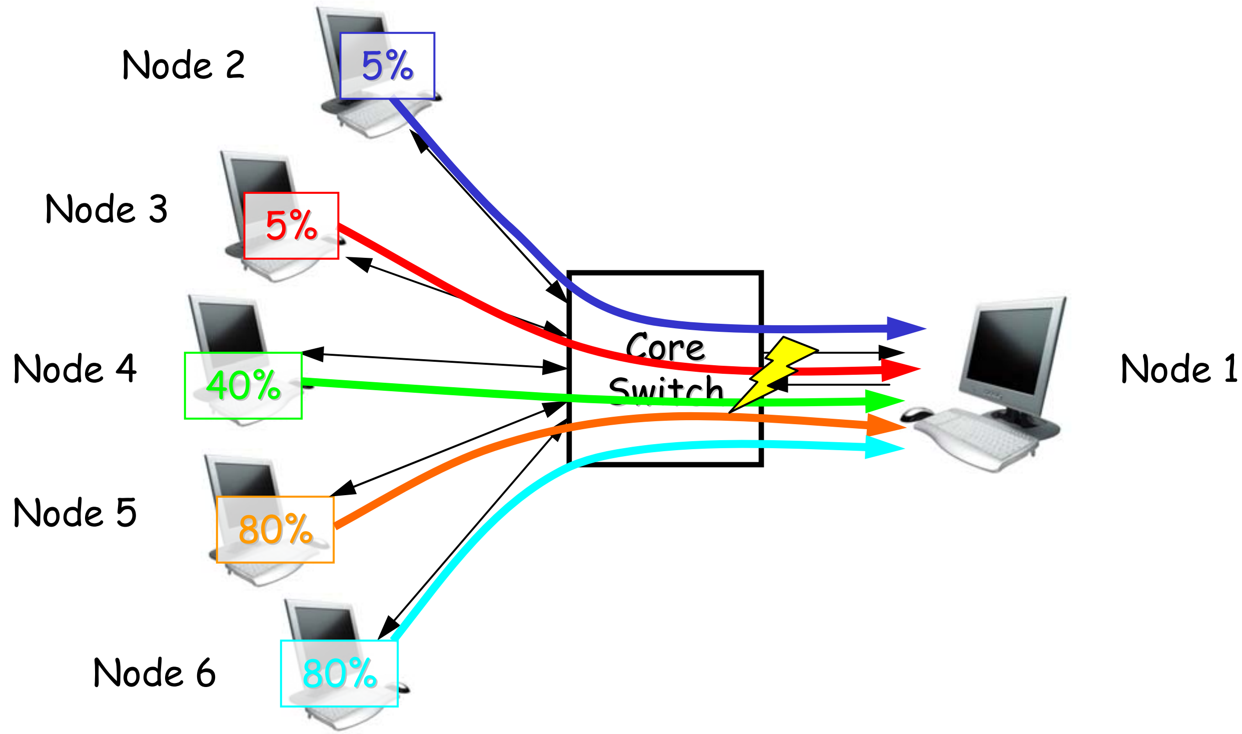
relative





- Convergence times determined over 1-ms averages of hot OQ length
- Relative accuracy a means that measured values stay within band [v*(1-a), v*(1+a)], where v is the steady-state value, so band width = 2*a

2. Input-Generated Mice-Elephant Hotspot

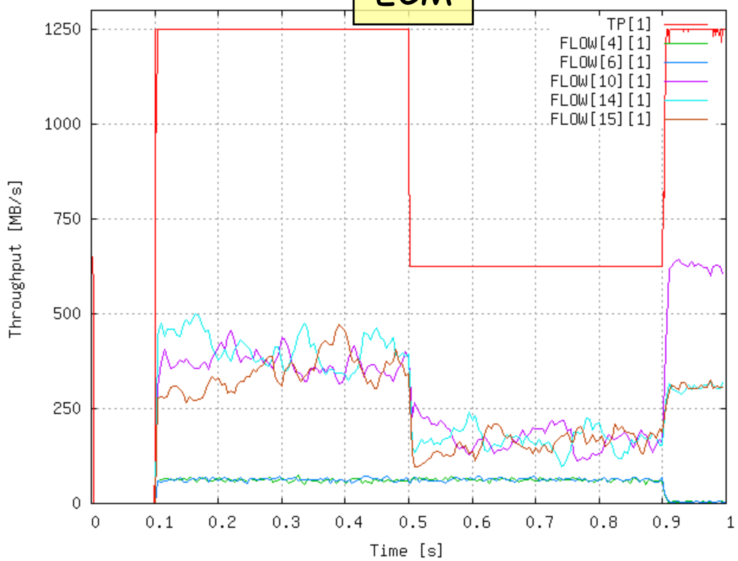


- 5 flows sending to hotspot: aggregate load = 21 Gb/s
- Max-min fair rates = (0.5; 0.5; 3; 3; 3) Gb/s =
  = (62.5; 62.5; 375; 375; 375) MB/s
- Hotspot max-min fair rates = (62.5; 62.5; 167; 167; 167) MB/s
- Achieved...

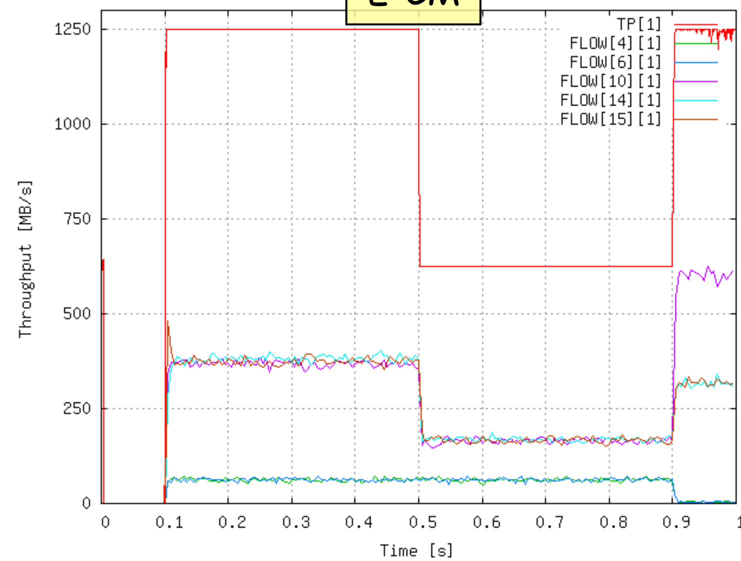
Results “Mice-Elephant” scenario (Tp and Q)

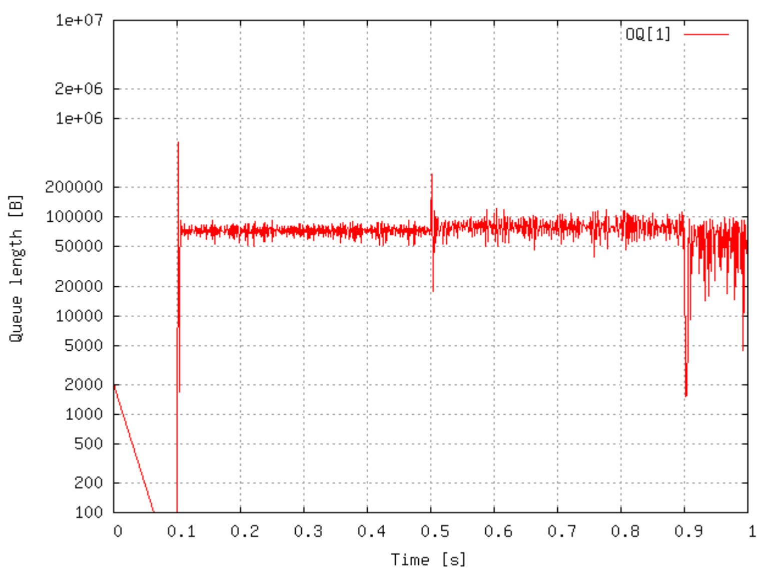
75 KB

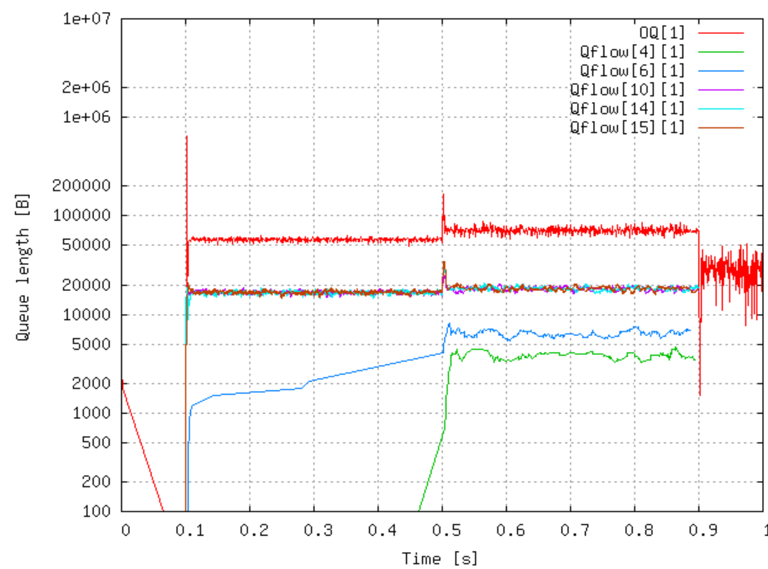
ECM



E²CM

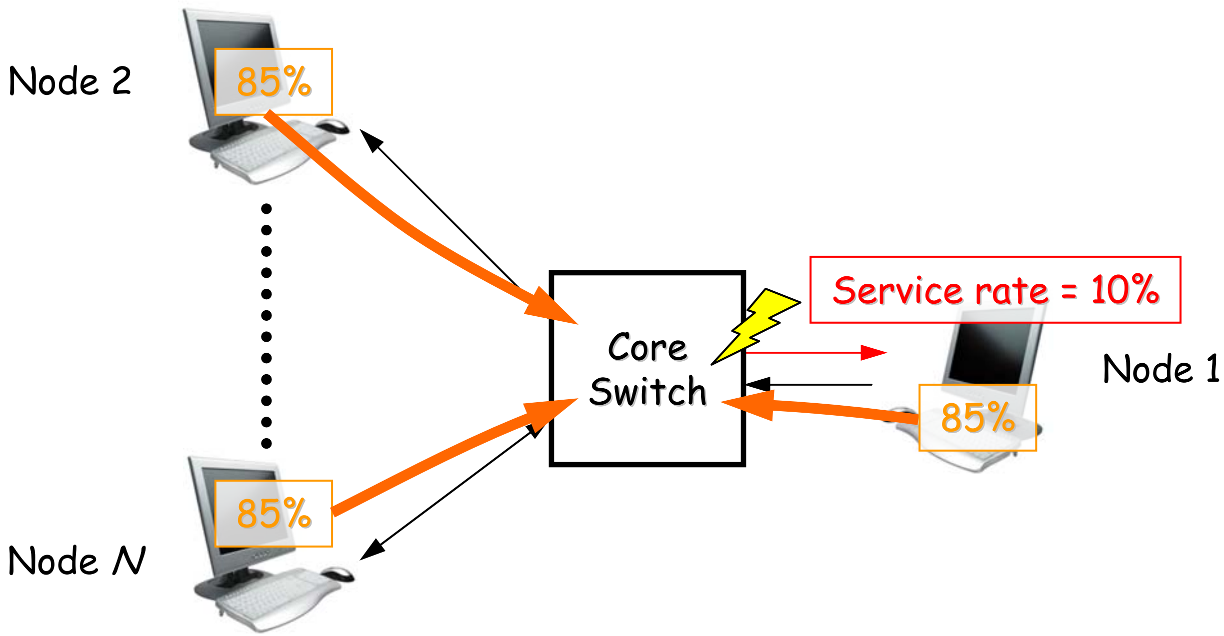






$G_i = 0.1 * (R_{link} / R_{unit}) * ((2*W+1)*Q_{eq})$

3. Output-Generated Single-Hop Hotspot

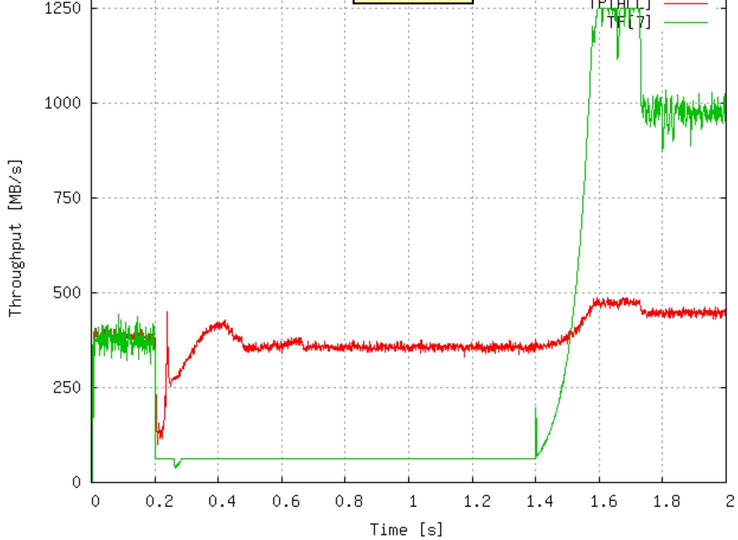


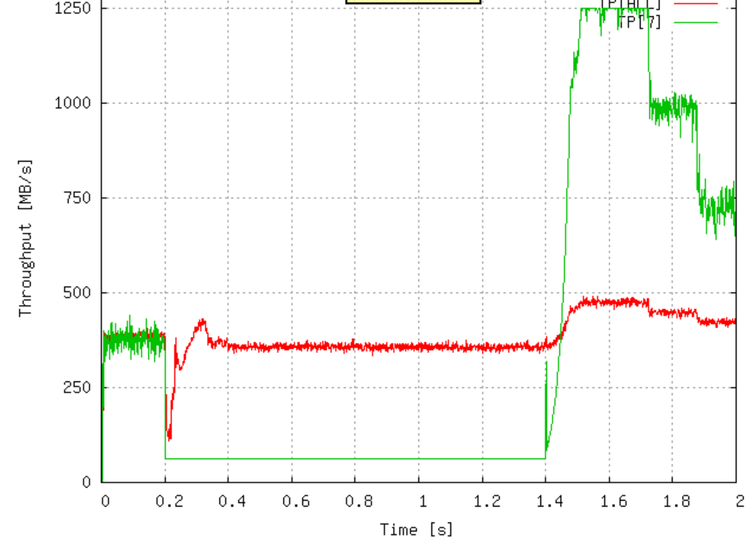
- All nodes: Uniform destination distribution, load = 85% (8.5 Gb/s)
- Node 1 service rate = 10%
- One congestion point
  - Hotspot degree = N-1
  - All flows affected => step response (test stability)

Results OG multi-hop scenario (Tp)

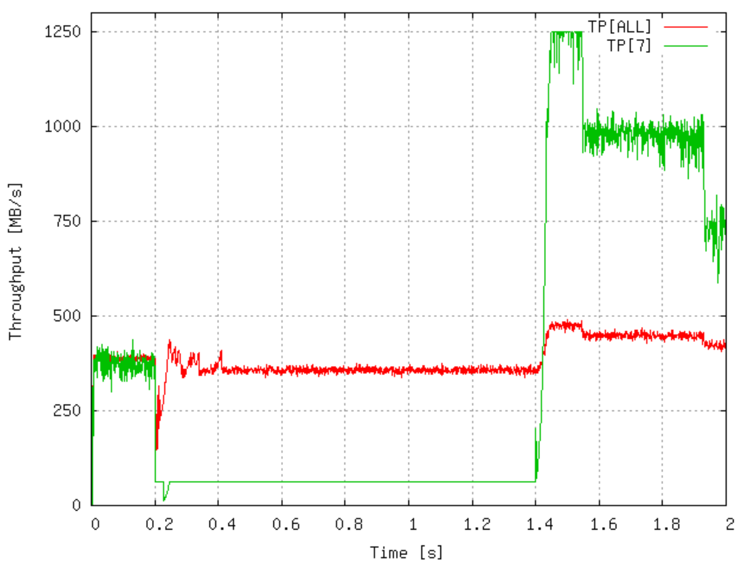
ECM | E^2CM

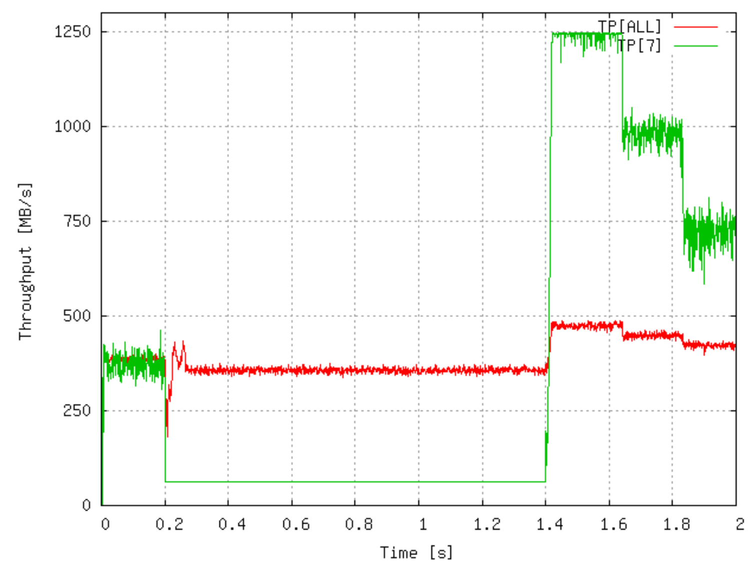
75 KB



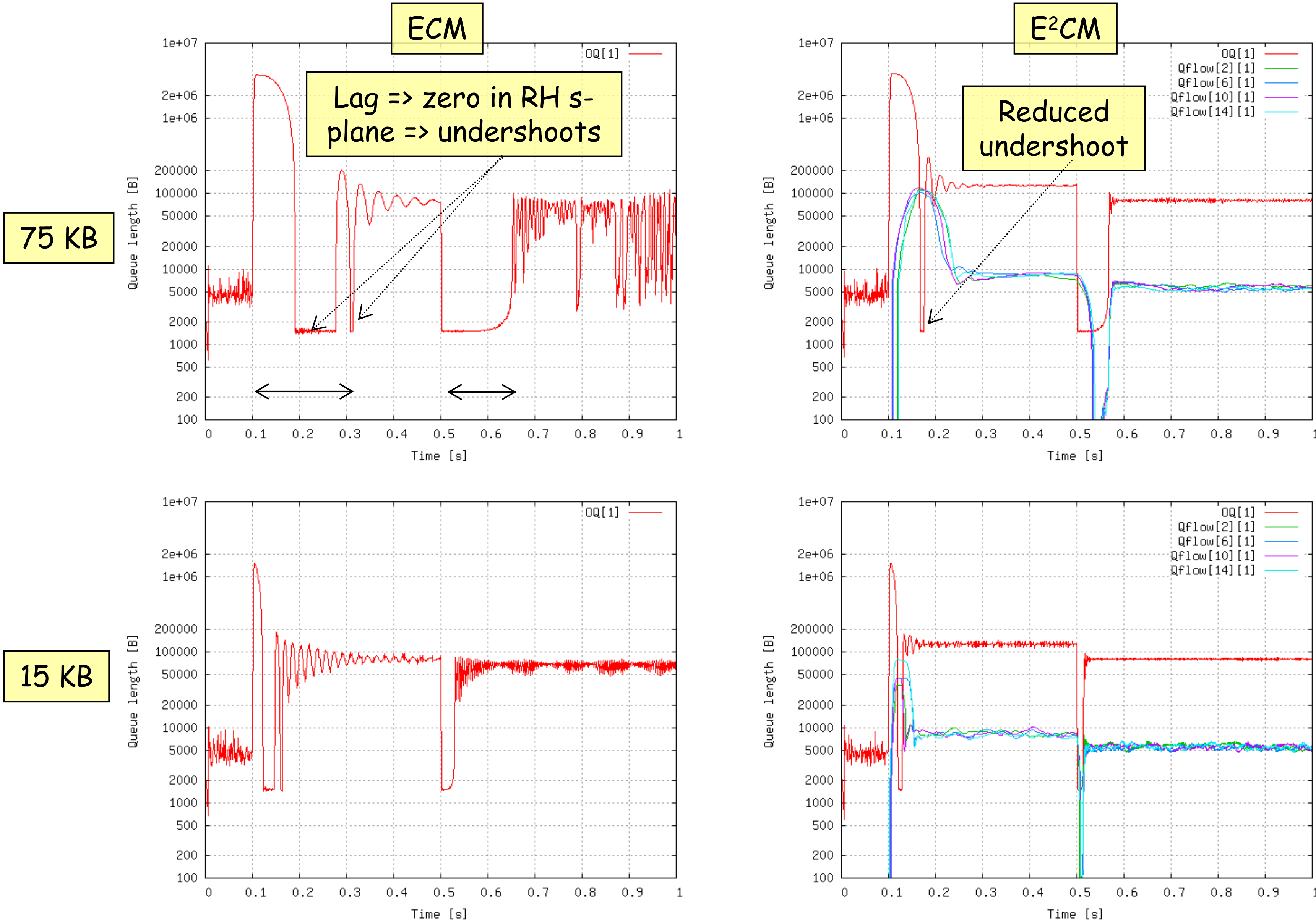


15 KB

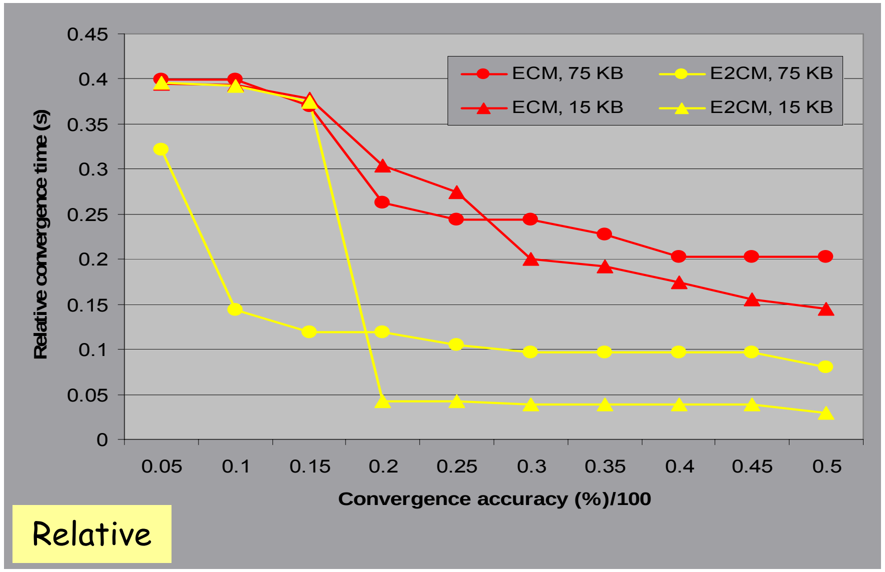


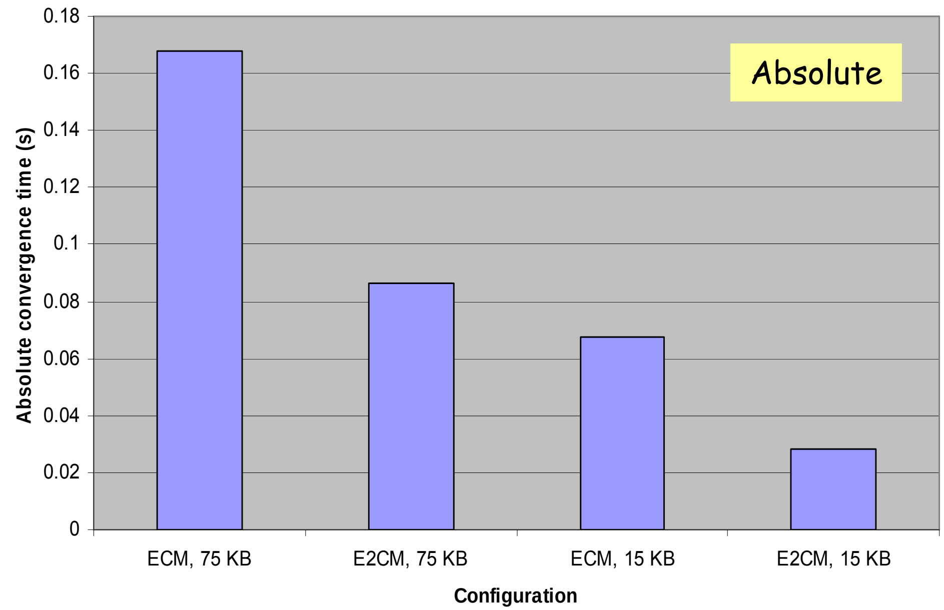


Results OG (Q): >2x Faster Convergence..!



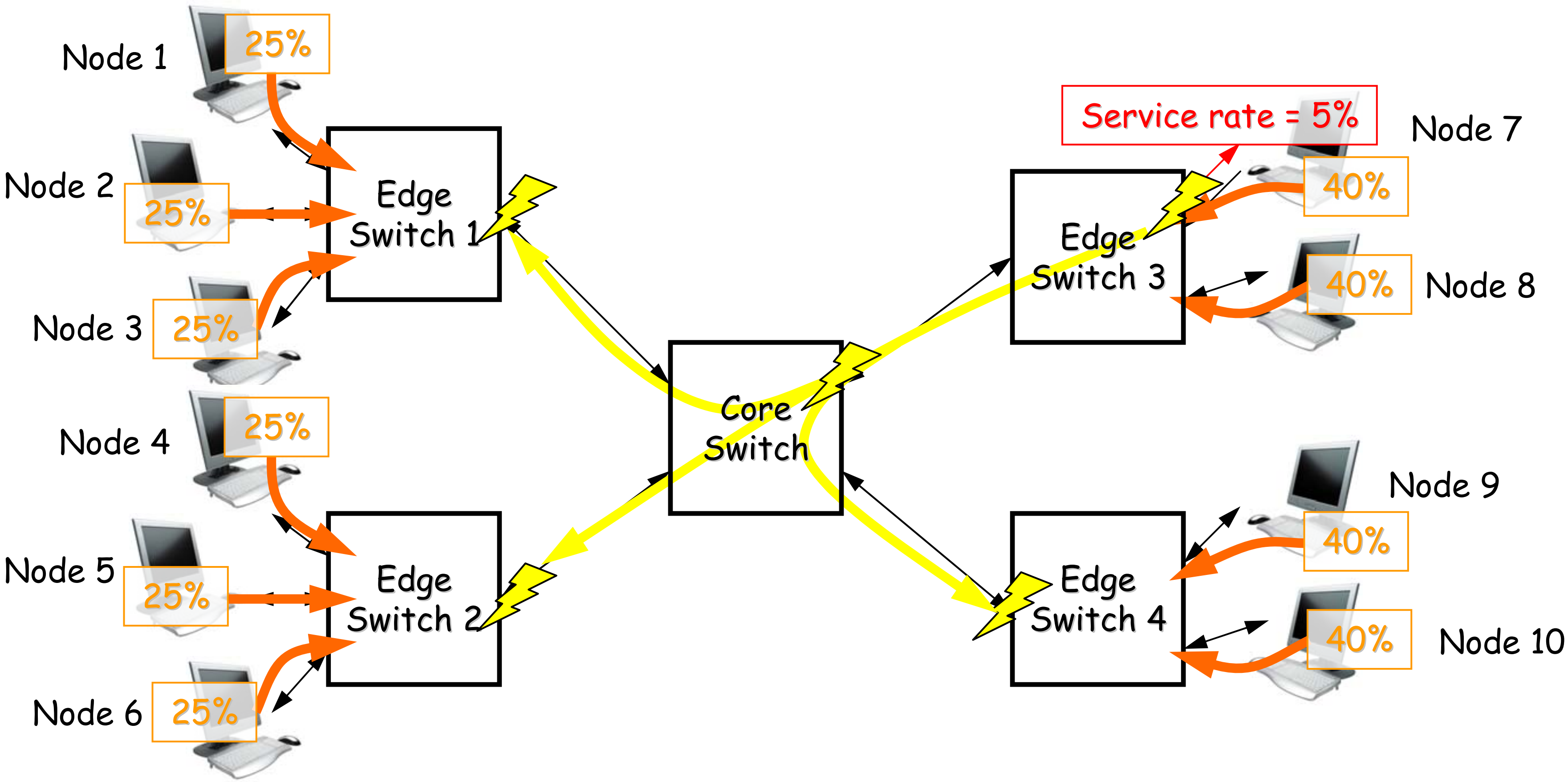
Convergence times single-hop OG scenario





- Convergence times determined over 1-ms averages of hot OQ length
- Relative accuracy *a* means that measured values stay within band [*v* *(1-*a*), *v* *(1+*a*)], where *v* is the steady-state value, so band width = 2**a*
- Absolute accuracy means that measured values stay within band [1.5, 280] KB

4. OG Multi-Hop Background Traffic Hotspot



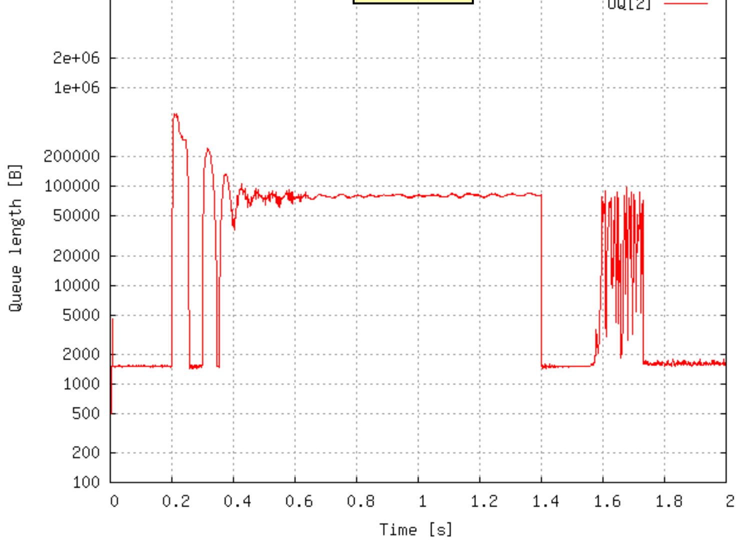
- All nodes: Uniform destination distribution
- Nodes 1-6 load = 25% (2.5 Gb/s), nodes 7-10 load = 40% (4 Gb/s)
  - Mean aggregate load = (6*.25+4*.4)/10 = 31% (3.1 Gb/s)
- Node 7 service rate = 5%
- Five congestion points
  - All switches and all flows affected

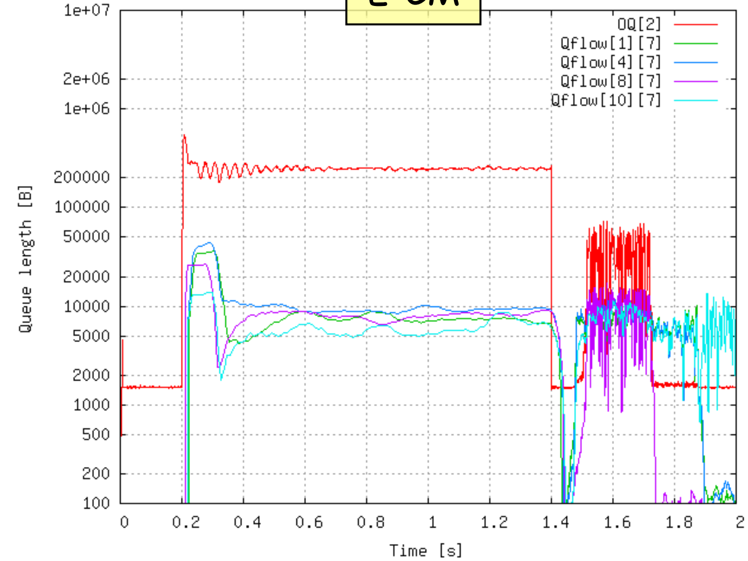
Results OG multi-hop BGND (Q)

ECM

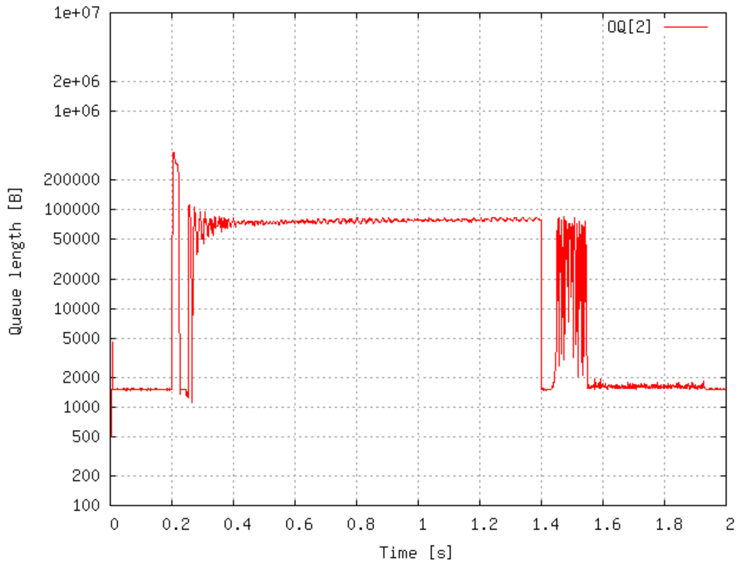
E^{2}CM

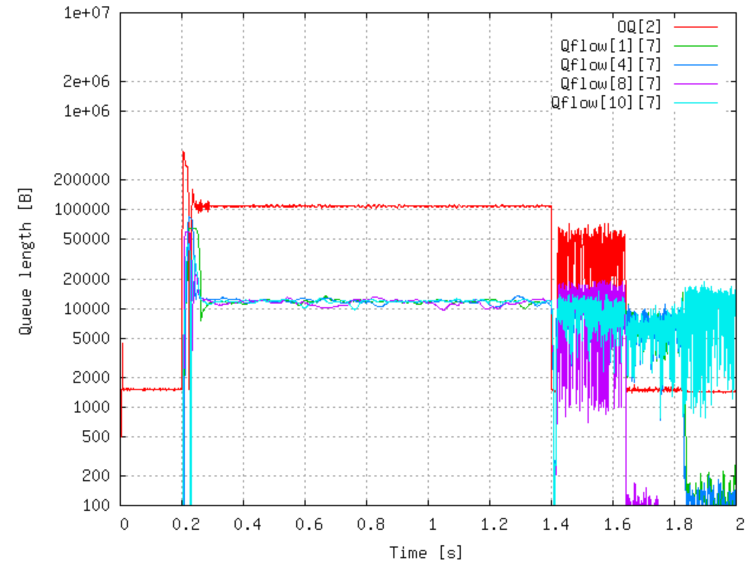
75 KB



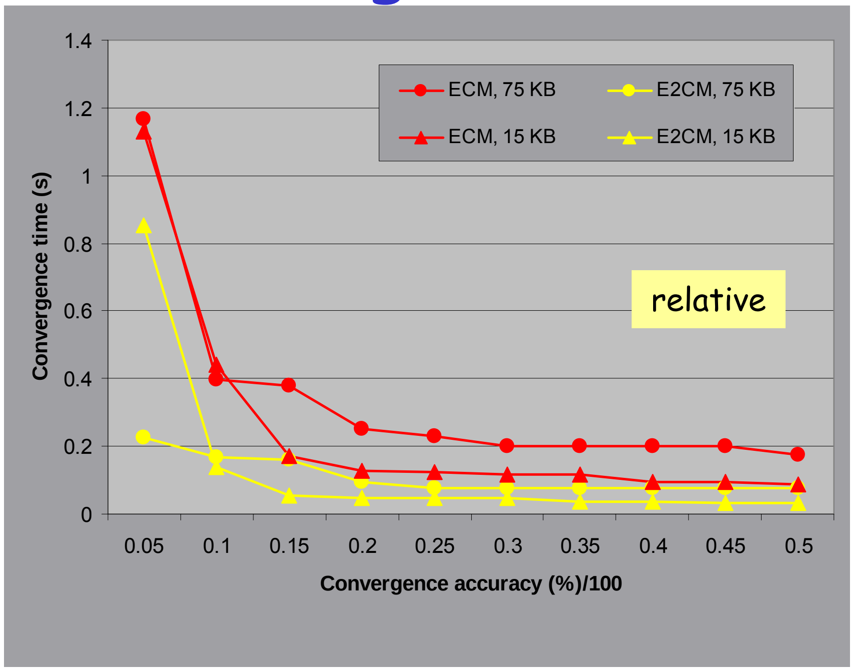


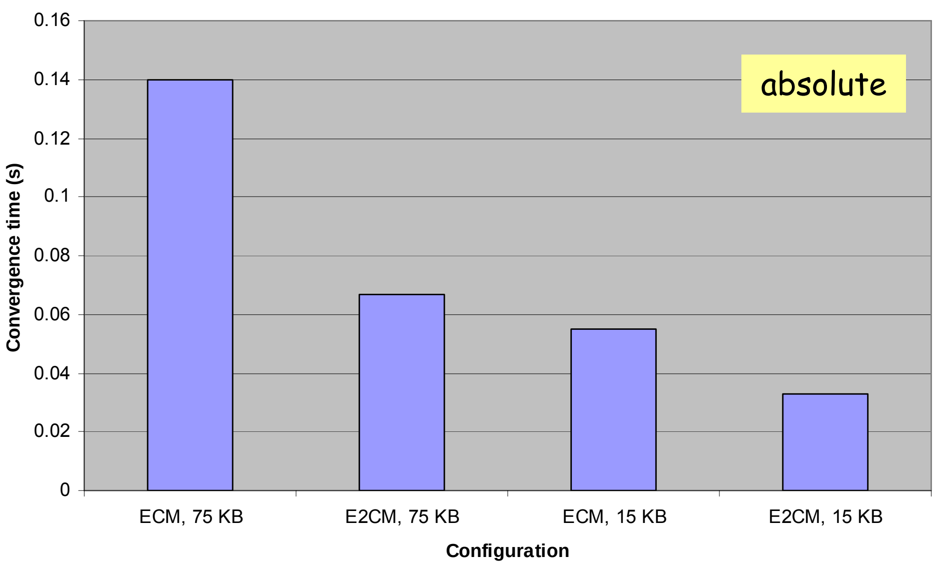
15 KB





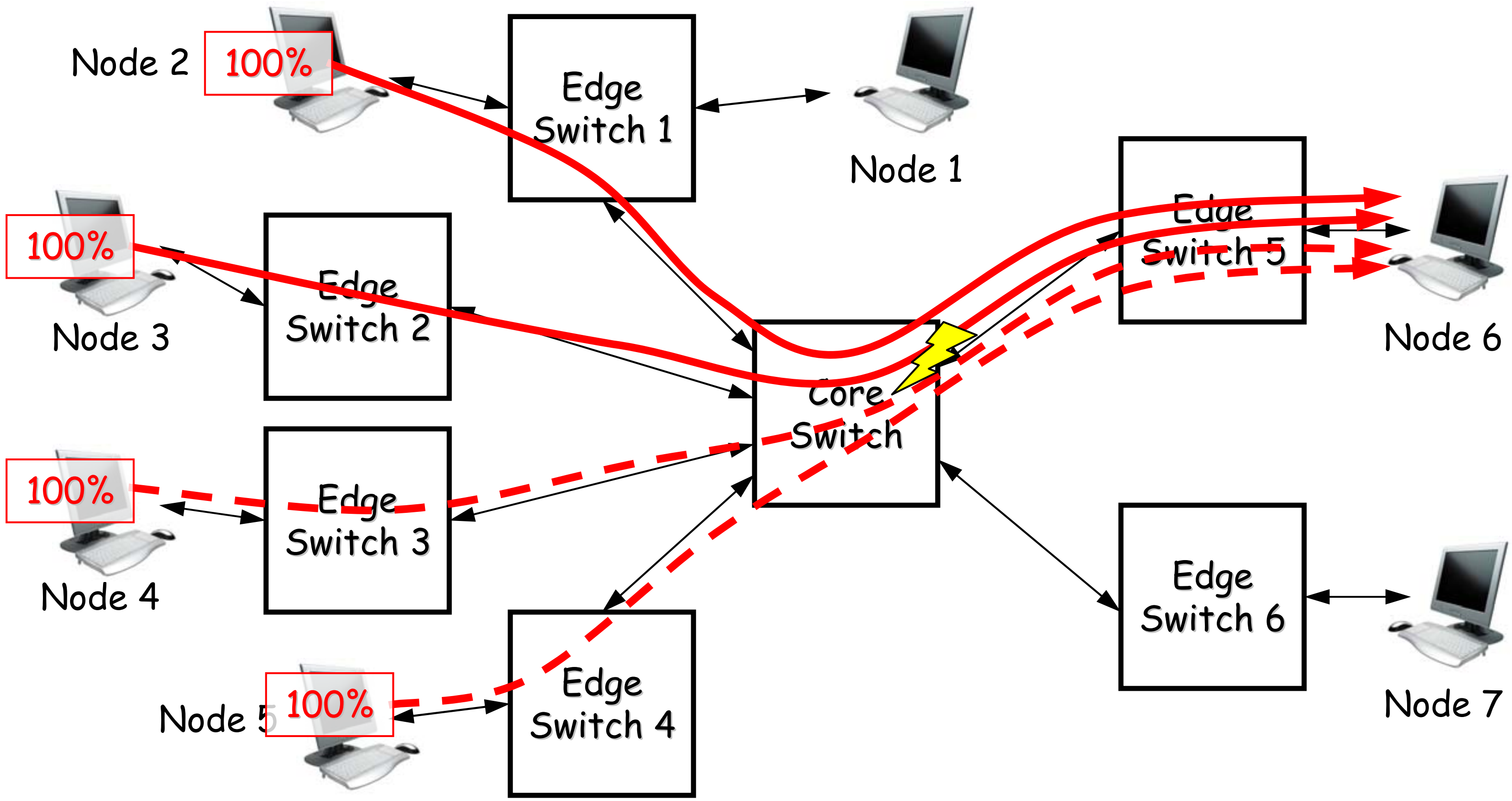
Convergence times multi-hop OG scenario





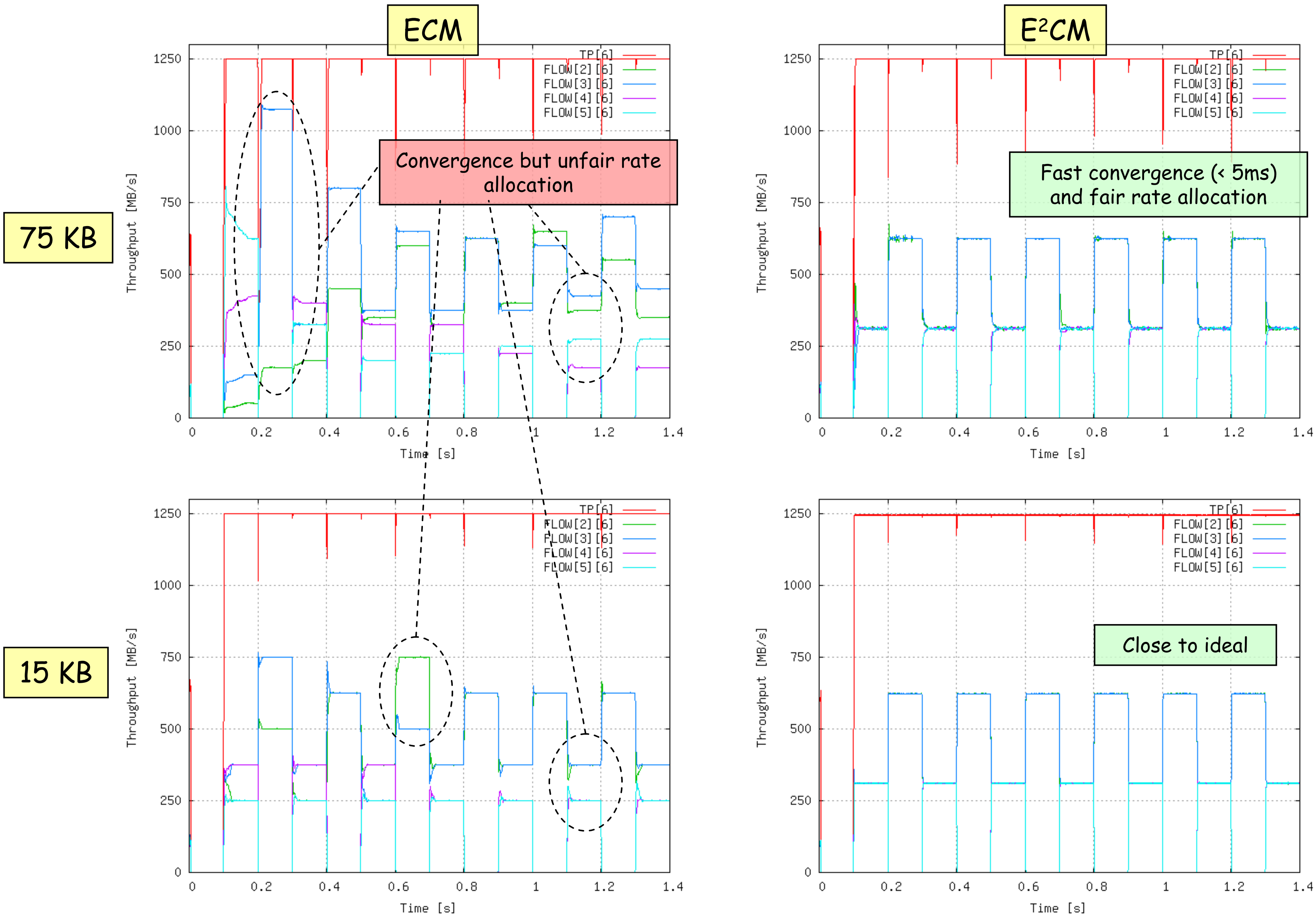
- Convergence times determined over 1-ms averages of hot OQ length
- Relative accuracy *a* means that measured values stay within band [*v**(1-*a*), *v**(1+*a*)], where *v* is the steady-state value, so band width = 2**a*
- Absolute accuracy means that measured values stay within band [1.5, 280] KB

5. Bursty Baseline Input-Generated Hotspot



- Four hot flows of 10 Gb/s each from nodes 2, 3, 4, 5 to node 6 (hotspot)
- Every 100 ms, flows from nodes 4 and 5 are switched from off to on and vice versa (duty cycle = 200 ms)
- Fair allocation provides 2.5 Gb/s per flow when 4 are active, 5 Gb/s when 2 are active
- **Pause disabled, very small adapter buffers (10 frames)**

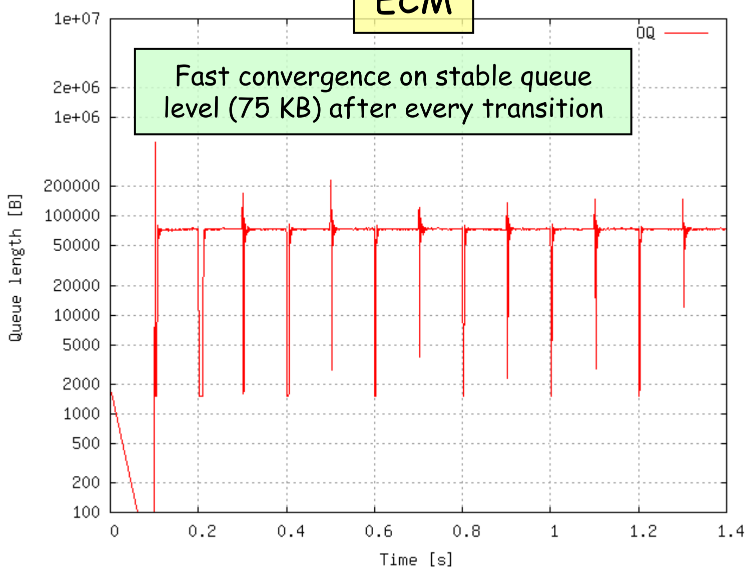
Bursty Baseline IG scenario (Tp): Convergence < 5ms



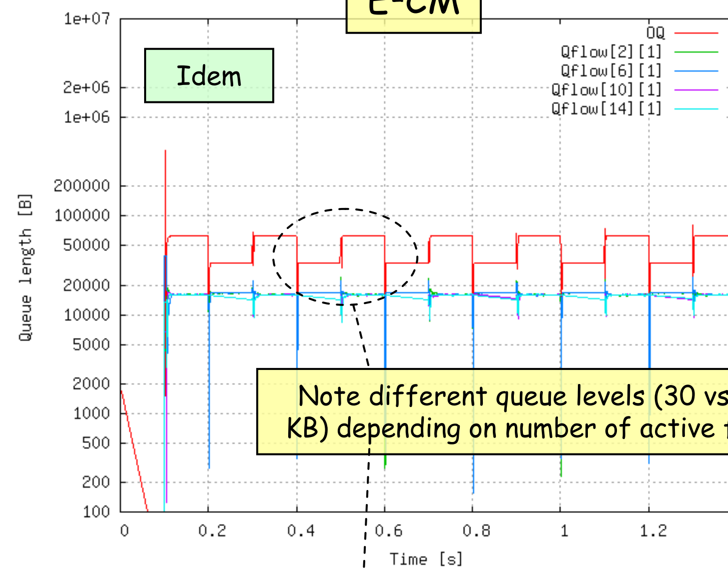
Bursty Baseline IG scenario (Q)

	ECM	E²CM
75 KB	Fast convergence on stable queue level (75 KB) after every transition	Idem
15 KB	Idem	Idem

75 KB, ECM:

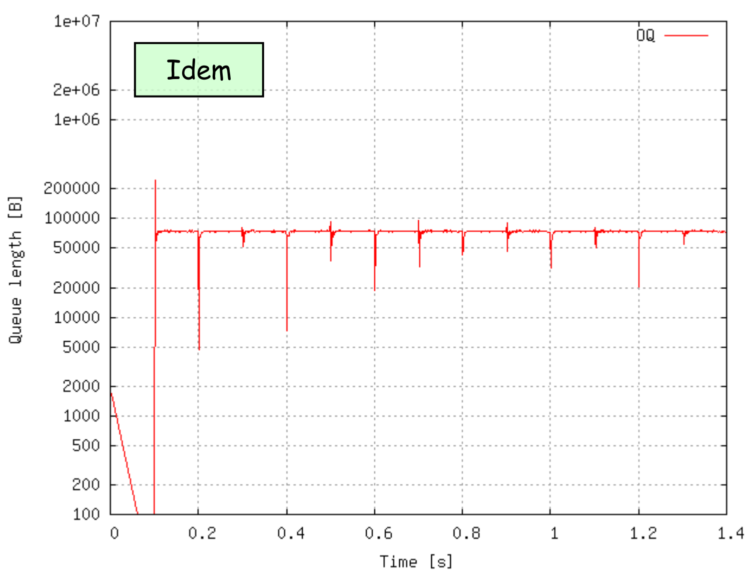


75 KB, E²CM:

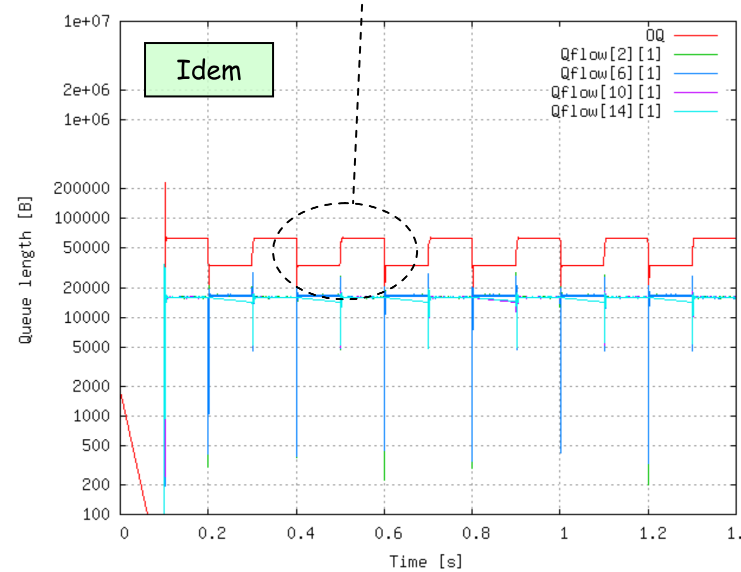


Note different queue levels (30 vs. 60 KB) depending on number of active flows

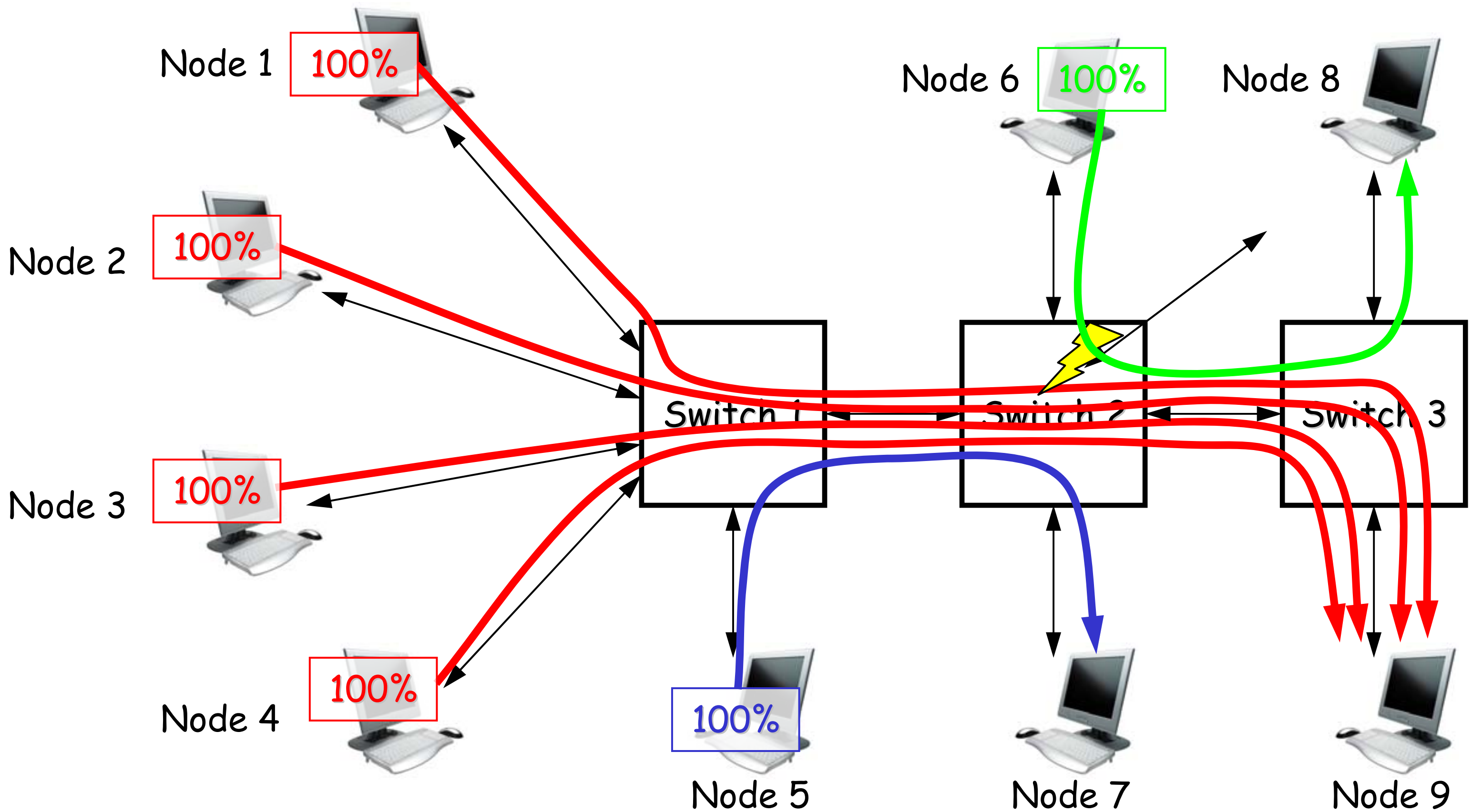
15 KB, ECM:



15 KB, E²CM:

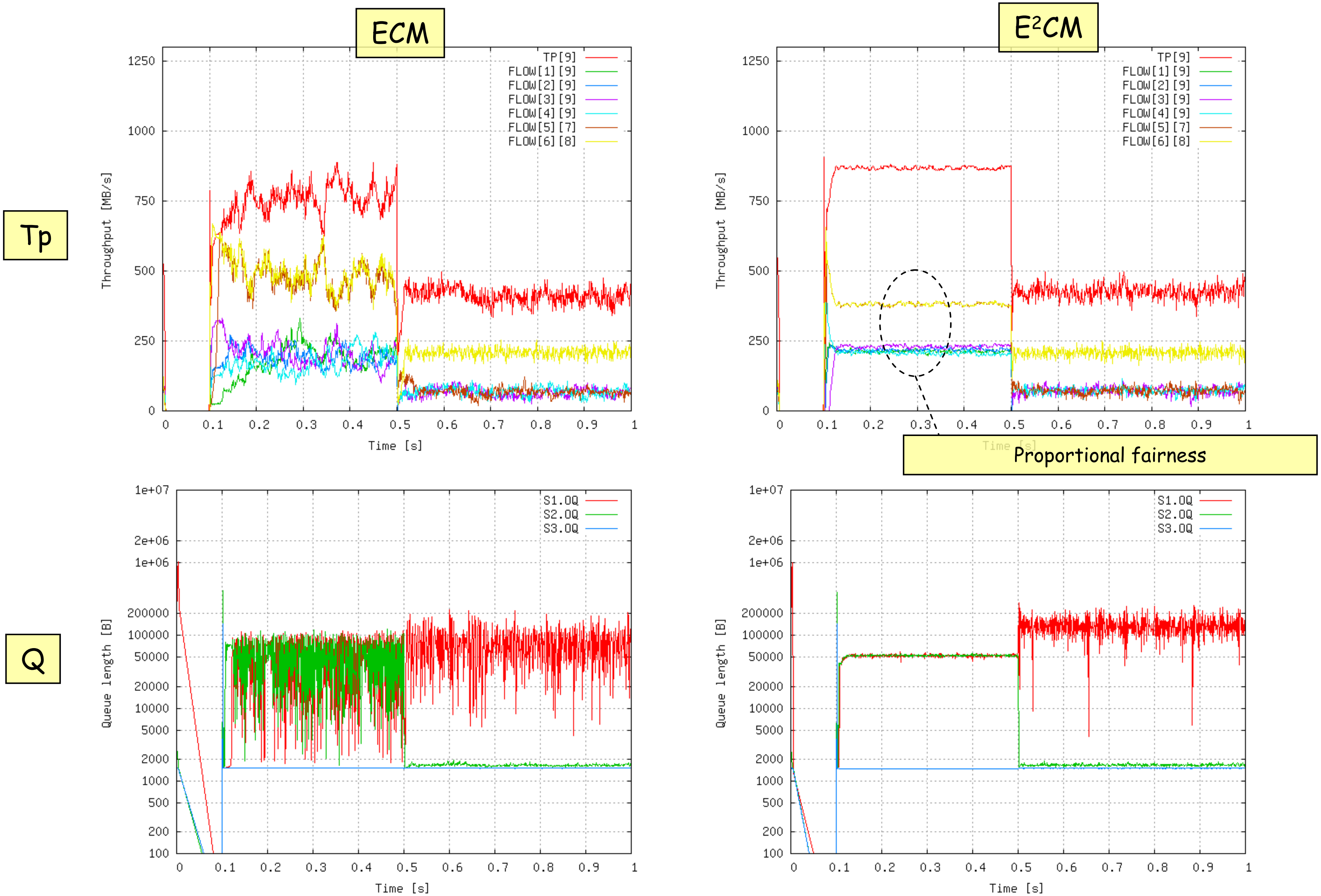


6. Parking Lot Scenario



- Four hot flows of 10 Gb/s each from nodes 1, 2, 3, 4 to node 9 (hotspot)
- Two cold flows of 10 Gb/s from node 5 to 7 and 6 to 8
- Max-min fair allocation provides 2.0 Gb/s to **all** flows
- Proportionally fair allocation provides 1.67 Gb/s to all hot flows and 3.33 Gb/s to all cold flows
- **Pause disabled, very small adapter buffers (10 frames)**

Parking Lot Scenario (75KB) - $Q_{eq,\ all}$ = 15 KB

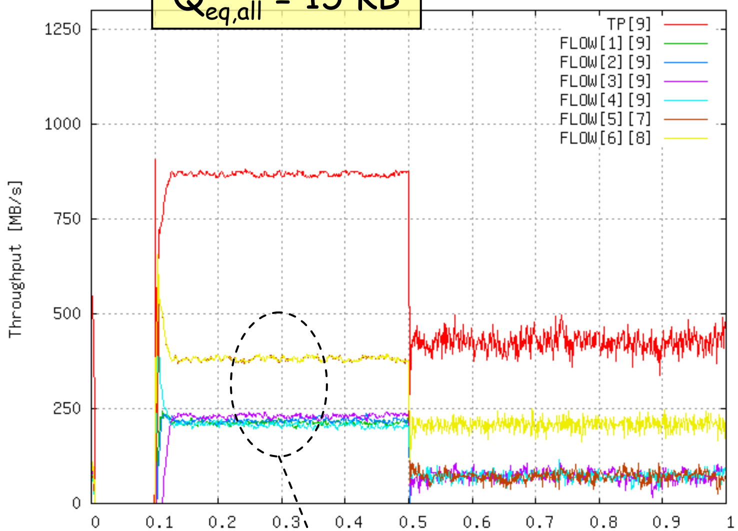


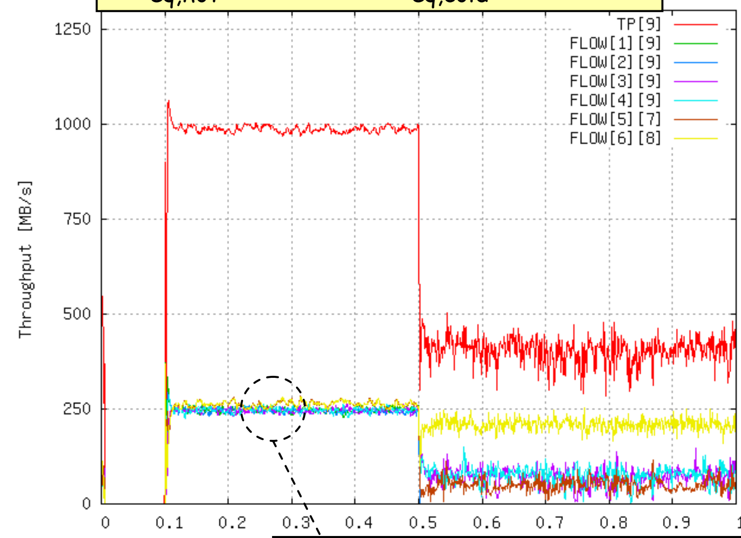
E^2CM - Parking Lot Scenario (75KB) – Per-flow Q_{eq}

$Q_{eq,all}$ = 15 KB

$Q_{eq,hot}$ = 15 KB, $Q_{eq,cold}$ = 7.5 KB

Tp

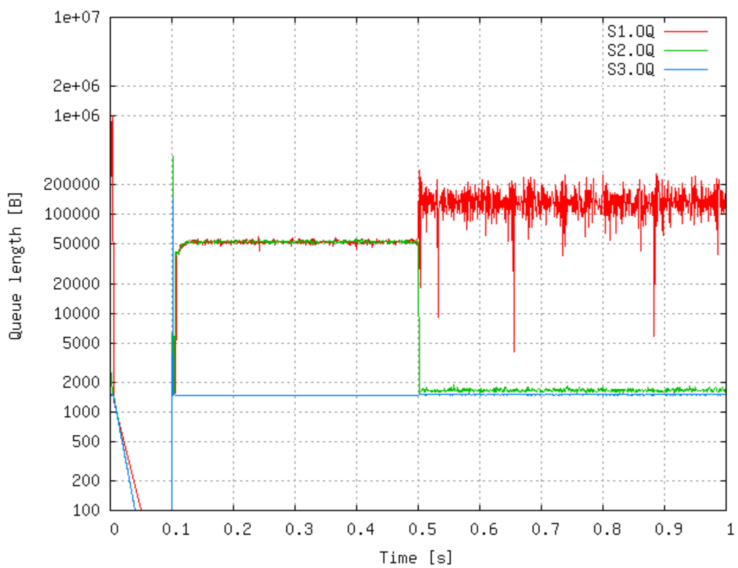


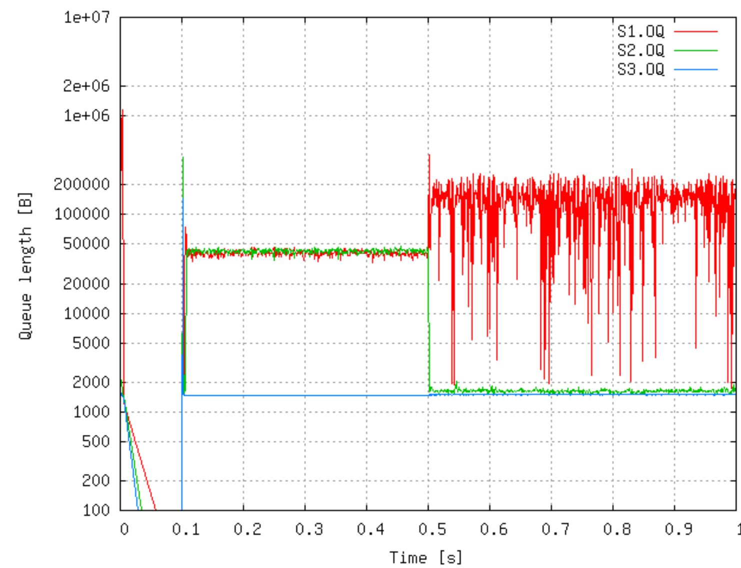


Proportional fairness: 4x1.67 Gb/s, 2x3.33 Gb/s

Max-min fairness: 6x2 Gb/s

Q





Conclusion

- E^2CM extends baseline BCN w/ per-path probing adopted from DBP and FECN

- E^2CM addresses the main critiques of BCN
  - improves performance in
    - ✓ stability and speed, based on an saturated integrator w/ extended dynamic range (distributed queue $Q_{ij} = \Sigma q_{ij}$)
    - ✓ linear range of F_b is now scalable w/ no. of buffers in the network
    - ✓ fairness (per flow accuracy possible in end-nodes, when needed)
      - scalability cost to 100+ Gbps Ethernet and 1M-node datacenter is TBD

- User-defined fairness:
  - 1. max-min (canonical, beneficial for 'mice')
  - 2. proportional (tempers 'remote' flows w/ long routing distance)
  - 3. max-T_{put} (maximize utilization at cost of unfairness, but no starvation)

- Synergy w/ FECN/DBP (probing) and w/ baseline ECM (param tuning)

- We propose the hybrid E^2CM as baseline CM approach

That's all, thanks!