Enhanced Forward Explicit Congestion Notification for Data Center Ethernet Networks

Chakchai So-In, Jinjing Jiang and Raj Jain
Washington University in Saint Louis
Saint Louis, MO 63130
Jain@cse.wustl.edu

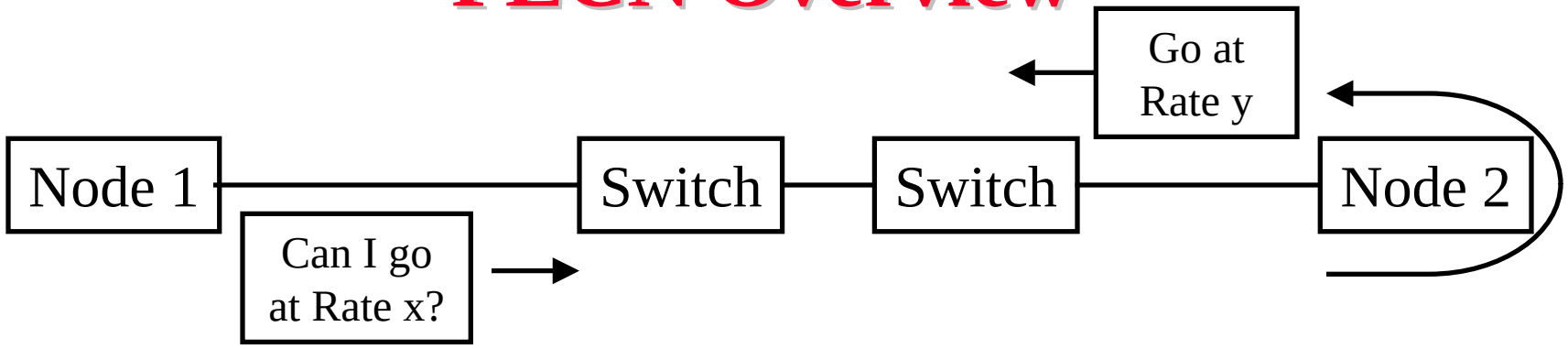
IEEE 802.1au Interim Meeting, Geneva, May 29, 2007

These slides are also available at:

http://www.cse.wustl.edu/~jain/ieee/fecn705.htm

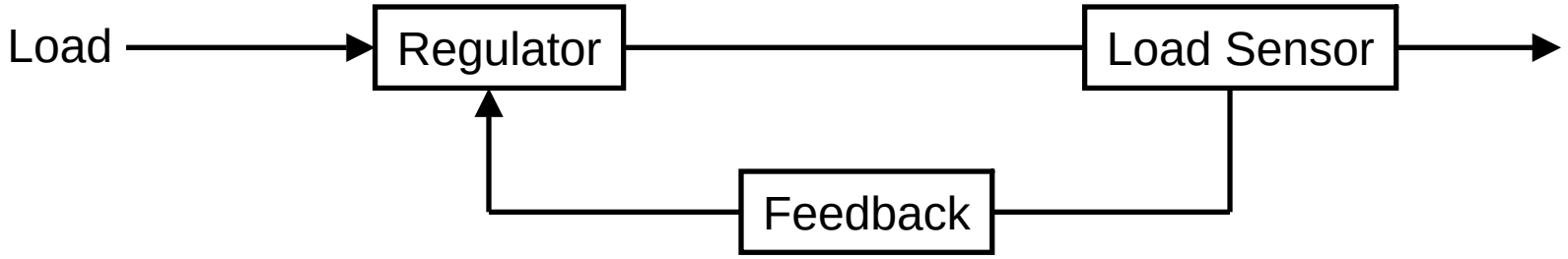
- Enhanced FECN
- Congestion Control and Avoidance
- Rate Probe Reflection and Generation
- Preliminary simulation results

FECN Overview



- Periodically, the sources probe the network for best available rate using "Rate Discovery packet"
- The probe contain only rate, Rate limiting Q ID
- The sender initializes the probes with rate=-1 ($\Rightarrow\infty$)
- Each switch computes an "advertised rate" based on its load
- The switches adjust the rate in probe packets down if necessary
- The receiver reflects the RD packets back to the source
- Source send at the rate received

Essential Components of Control



	BCN	FECN
Load	Rate	Rate
Regulator	Rate	Rate
Load sensor	Queue	Rate
Feedback	Queue	Rate

Strengths of FECN

1. Explicit feedback vs implicit (drift up)
2. Rate based feedback vs queue feedback
   Queue feedback from very different link rates are not comparable.
3. Rate based load sensor vs queue based sensor
   Instantaneous queue values are very noisy indicator of load.
4. Simple source algorithm
   No computation. No drifts. No RTT measurements. Single feedback signal (BCN, BCN0, BCNmax, …)
5. Very low overhead = 1/10th BCN [Cisco's simulations]
6. Fast rate increase vs drifting
7. Perfect fairness
8. Only feedback format needs to be standardized.
   Internal algorithms should be left to vendors and users.

Feedback: Desired Changes in FECN

Don't automatically start with a rate regulator

$\Rightarrow$ Start high

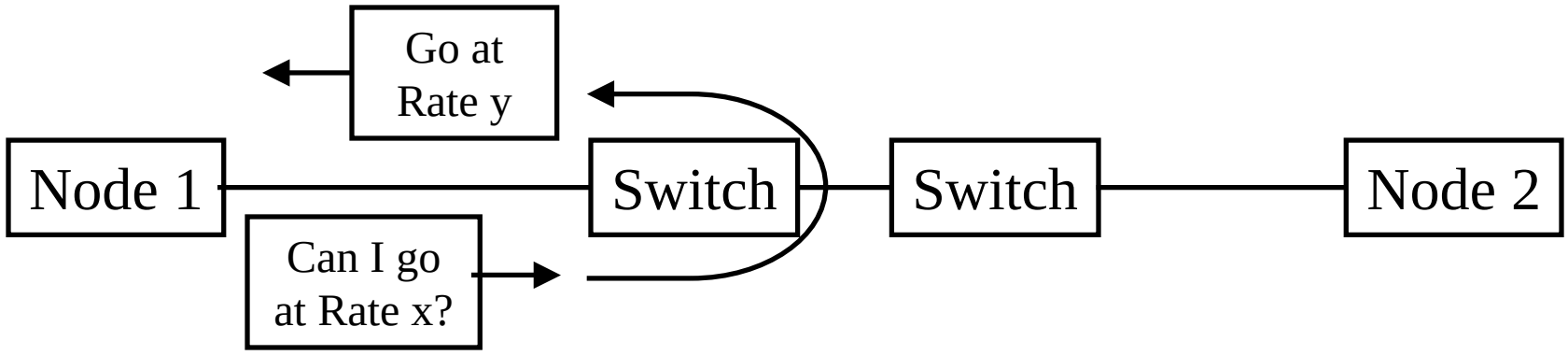
Enhanced FECN

1. Switch from congestion avoidance to congestion control $\Rightarrow$ Allow fast start
2. Combine the best of FECN and BCN

Congestion Control vs Avoidance

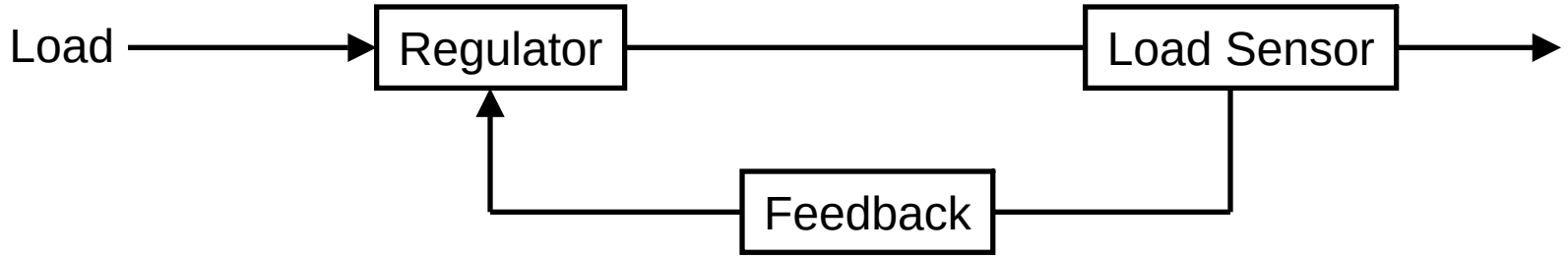
- Control = Reactive vs Avoidance = pro-active
- Pro-active ⇒ Slow start
- Fast start and regulate if congestion experienced
- Notes:
  - Fast start (with any scheme) lasts only as long as effectively there is a single flow
  - With two or more flows passing through a congestion point, there will be congestion

Rate Probe Reflection and Generation



- Enhancements:
  - Probes can be reflected by any switch
  - Probes can even by generated by the switches
  - Rate increase can occur only by destination (or final switch) reflected probes.
  - Rate decrease can be caused by any probe regardless of its origin or reflection point

Essential Components of Control (Cont)



	BCN	FECN	Enhanced
Load	Rate	Rate	Rate
Regulator	Rate	Rate	Rate
Load sensor	Queue	Rate	Rate+Queue
Feedback	Queue	Rate	Rate

FECN Tag Format

- All tags have the same format. Switch generated BCN00 is identical to source generated FECN tag
- Switch can only generate backward rate decrease signal to 0 = R_{min}
- Source behavior is same for reflected or switch generated FECN tags
- FECN Tags at Source:

- BCN00 from Switch:

Source Action

- When to start tagging:
  - *As soon as a flow starts (simple); T=1ms (Congestion avoidance)
  - Only when a flow receives a BCN00 from switch (Congestion Control)

Switch Action

- Switches update rate in the tag if their advertised rate is lower and update CPID.
- Under severe congestion: (q_> Q_sc threshold)
  - *Generate BCN00 by sampling if q >Q_sc
    - Set rate to R_{min} = C/N0
    - When q ⩽ Q_sc use normal FECN

Control Parameters

- Fast Start in all cases
- No pause
- Frequency of tagging = 1 ms
- Qsc = 80 packets of 1500 bytes = 120 kB
- R_{min} = 500 Mbps (To be studied further)
- Control Schemes:
  - FECN
  - BCN
  - FECN with BCN00

Simulation Parameters

- Network Configuration
  - Configurations: 1 Congestion Point (CP)
  - Link Capacity = 10Gbps (Default)
  - Congested link 10G
  - Switch port buffer size = 150 KB (100 pkts)
  - Switch latency (1us) + Prop delay (0.5us)
- Traffic Pattern
  - UDP
  - Workload = CBR 10Gbps
  - Frame size: Fixed 1500B
- Simulation duration: 100ms (4 flows start at 5ms and 2flows end at 80ms)

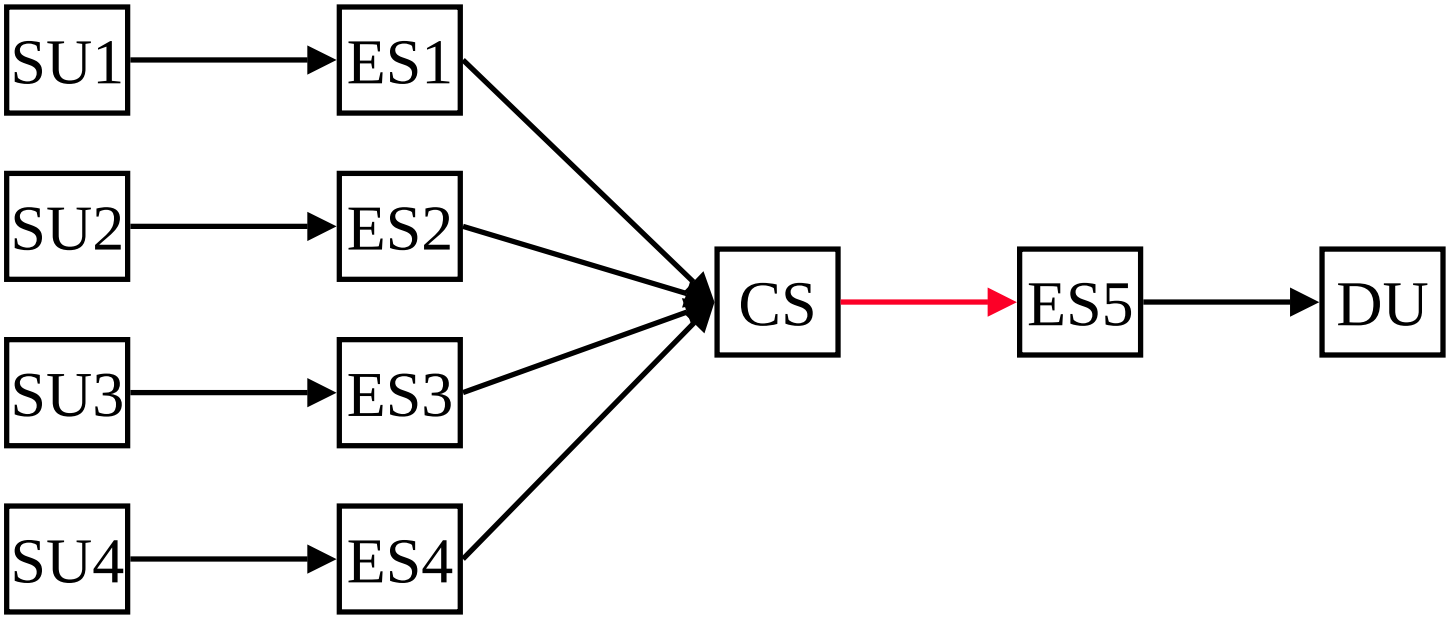
FECN Parameters

- T = 1ms and 0.05ms
- a = 1.1, b=1.002, c = 0.1
- Qeq = 16*1500B (100 packets)
- Initial source rate R0=10Gbs $\Rightarrow$ fast start

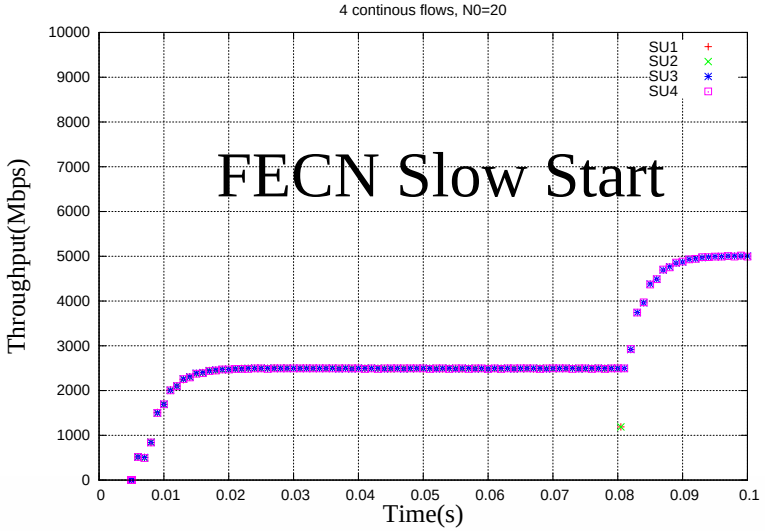
BCN Parameters

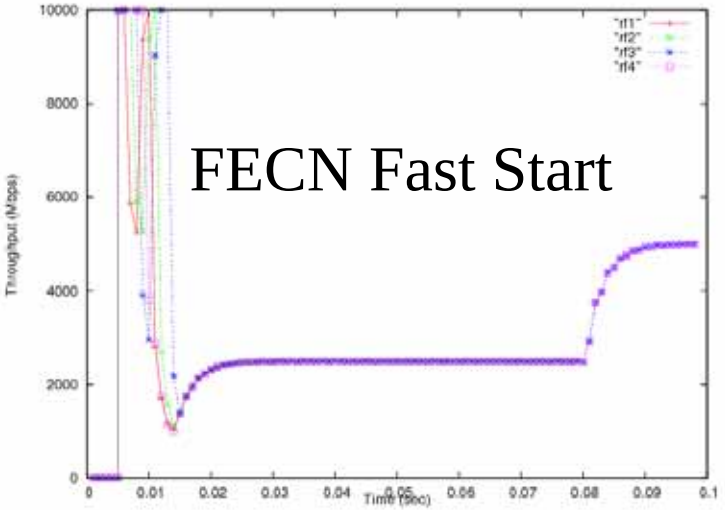
- Qeq = 16*1500B (100 packets)
  - W=2, Gi = 0.53, Gd = 0.0002667, Ru = 1Mbps.
- Fixed Sampling = 75000B (2%)
  - Over sampling (10%) if Q > Qsc (Qsc=80)
- BCN-Max used in lieu of BCN(0,0)

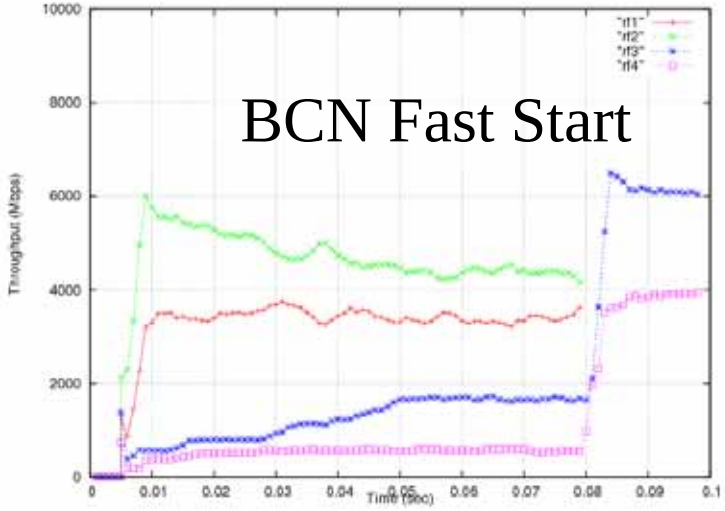
Baseline 4 source 1 CP (RR) (BCN, FECN, FECN with BCN)



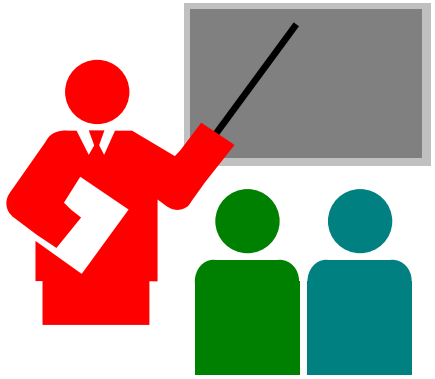
Source Throughput







Summary



1. Preliminary simulations show that FECN with BCN00 works better than FECN or BCN alone.
2. Combines the fast rise, low overhead, and fairness of FECN with fast decrease of BCN00.
3. This enhancement allows sources to fast start.
   Rate regulator can be installed after receiving BCN00.
4. Need to do more detailed simulations.
5. In particular, the min rate R_{min} needs further investigation.