Congestion Notification for Data Center Ethernet Networks: Key Principles

Raj Jain, Jinjing Jiang, Chakchai So-In
Washington University in Saint Louis
Saint Louis, MO 63130

Jain@cse.wustl.edu

IEEE 802.1au Meeting, San Francisco, July 18, 2007

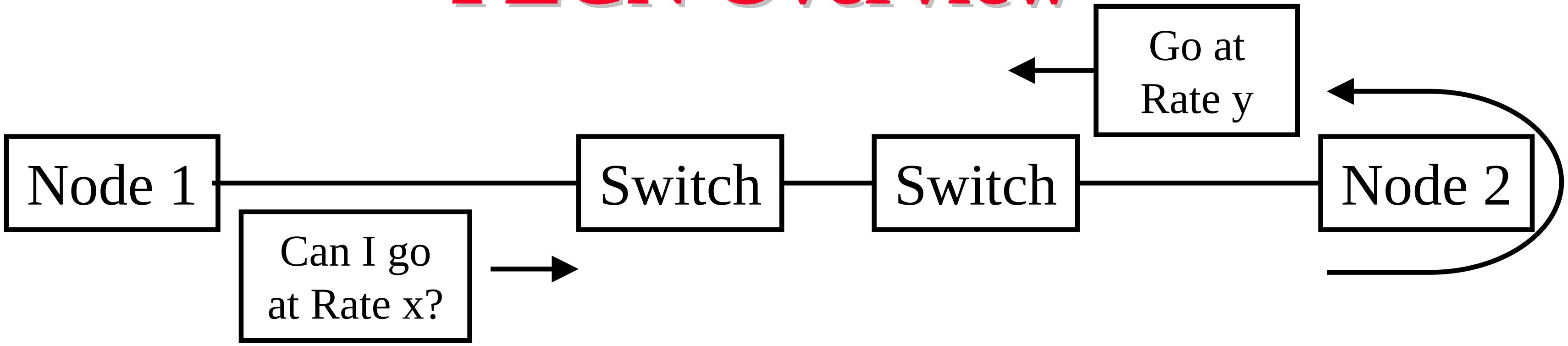
These slides are also available at:

http://www.cse.wustl.edu/~jain/ieee/fecn707b.htm

Overview

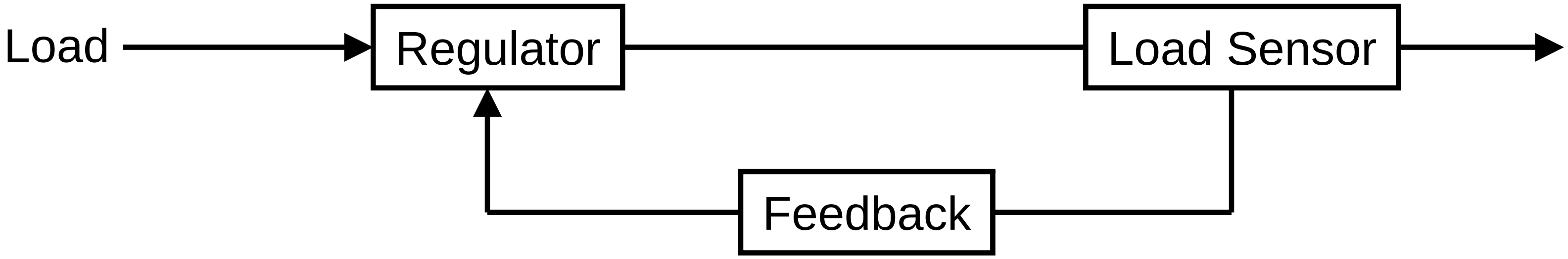
1. Explicit vs Implicit Feedback
2. Rate vs Queue Load Sensor
3. Rate vs Queue Feedback
4. Data Plane vs Control Plane
5. Source Complexity
6. Network Overhead
7. Random vs Predictable Behavior

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 (⇒∞)
- 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	E2CM	QCN
Load	Rate	Rate	Rate	Rate
Regulator	Rate	Rate	Rate	Rate
Load sensor	Queue	Rate	Queue	Queue
Feedback	Queue	Rate	Queue	Queue

Explicit vs Implicit Feedback

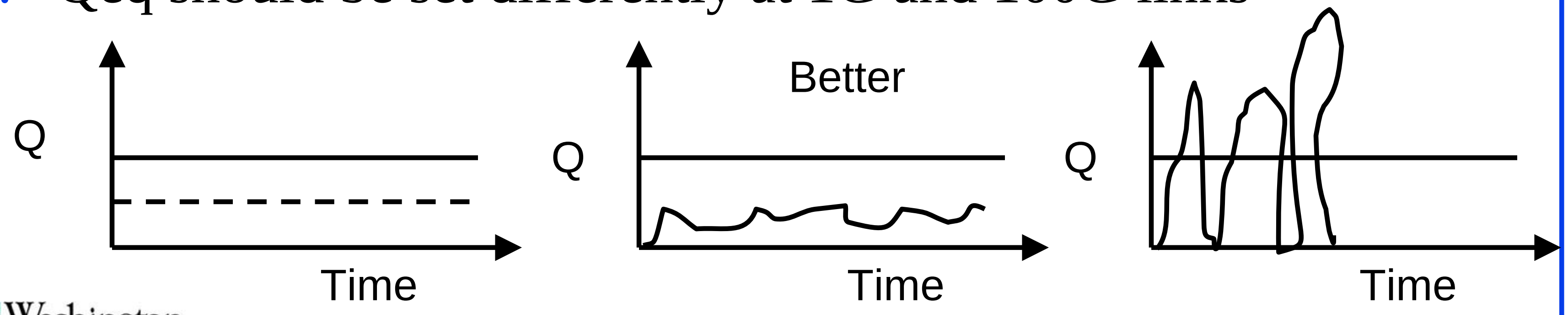
1. Explicit is better than implicit
2. All schemes have explicit negative feedback
3. BCN has sampled positive explicit feedback.
   => Increase probability decreases for lower rate sources => Main cause of unfairness
4. QCN-2P has no positive feedback.
   => Increase is by trial (implicit)
   => Slow transient as shown by Davide’s simulations
5. QCN-3P has one-bit positive feedback but it is ignored 96% of the time and sent back only 4% of the time

Rate vs Queue Load Sensor

1. Queue is a random quantity. For the same load, the instantaneous queue length can vary a lot. Simple M/M/1:

$$P(Q = n) = (1 - \rho)\rho^n$$

2. Queue length depends upon the queue service architecture
3. Queue length is highly related to the bottleneck link rate. Ten 1500 byte packet at a 1 kbps link are a "big" queue while would be negligible at 10Gbps link.
4. Optimal: Mean Q = 1 (Q includes the packet in service)
5. Qeq should be set differently at 1G and 100G links



Q vs Rate Load Sensor (Cont)

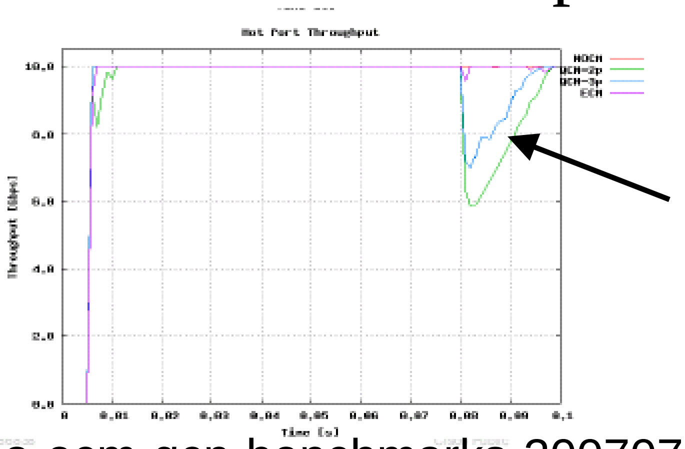
1. Rate based Load Sensor
   1. Utilization = Arrival rate/service rate
   2. Desired utilization is same at 1G and 100G links
2. Rate = packets serviced per unit of time
   => Stable quantity (low variance)
3. Rate is a better measure of link utilization. Managers can easily set the goal.

Rate vs Queue Feedback

- Queue length feedback from different links can not be compared
- Again, Ten 1500 byte packet at a 1 kbps link are a “big” queue while would be negligible at 10Gbps link.
- Rate feedback from different link speeds has exactly the same meaning. => When a source is told to send at 1 kbps, it does not matter whether the bottleneck is 10Gbps, or 1 Mbps, the source should send at 1 kbps.
- Queue feedback should indicate the link capacity, burstineess of traffic, queueing structure, …
  Fb = 10 from 1 Mbps link is extremely bad news
  Fb = 10 from 100 Gbps link is mildly bad news
  QCN will decrease by the same amount for both of these feedbacks => increased transient time

Transient Response Time

- Wrong feedback => Multiple attempts to reach goal
- Example: Correct rate = 5 Gbps,
  1, 2, 3, 4, 5, …
- Time to reach the optimal increases by a few RTTs.
- Shows up as slow transient response time



Missed/delayed information

Ref: au-bergamasco-ecm-qcn-benchmarks-20070717.pdf

Data Plane vs Control Plane

1. FECN operates entirely in control plane
   There are no bits/no tags in the data packet headers
2. BCN sources add a tag for increase request to all data packets
3. QCN-3P requires header and CRC modification in the data plane

Source Complexity

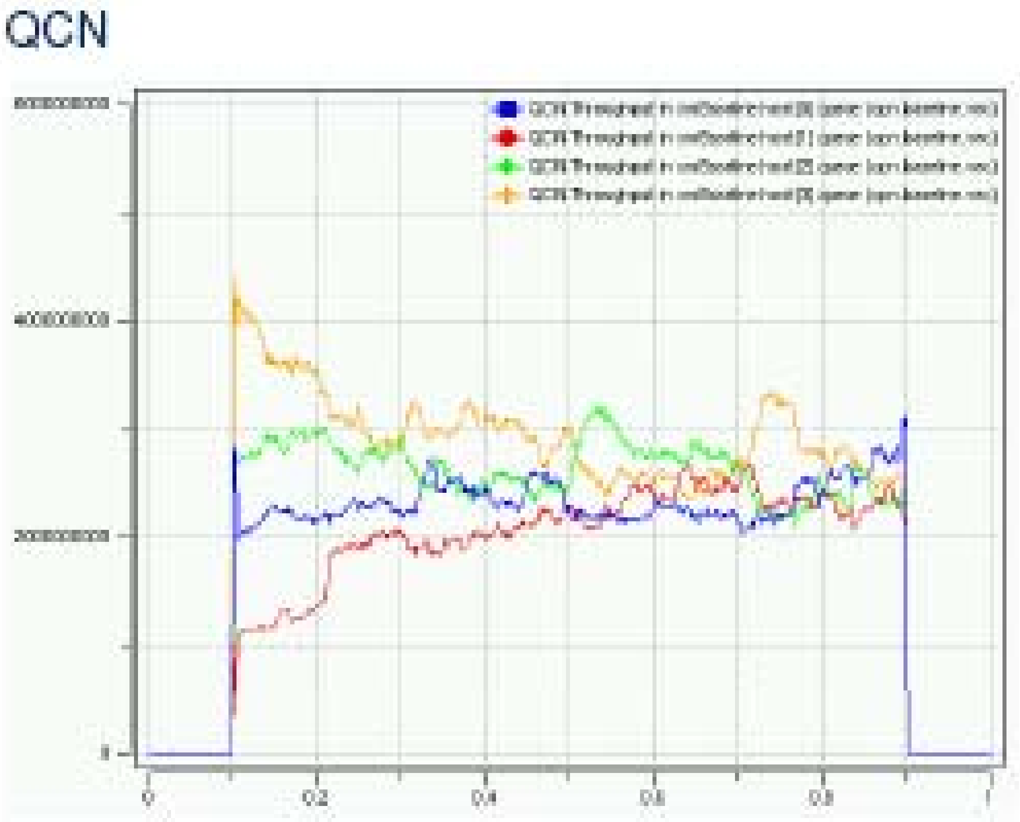
1. FECN source algorithm is simple.
   Source Rate <- Rate in FECN probes
2. No computation. No drifts. No RTT measurements.
   Single feedback signal (BCN, BCN0, BCNmax, …)
3. No Time based drift, byte based increase, jitter,…
4. High cost NICs => No deployments

Network Overhead

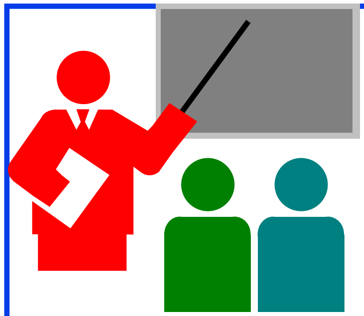
1. 10% sampling
   => 10% extra traffic in the reverse direction
   => Significant for high speed links
2. Queue feedback includes lot of bytes

Random vs Predictable Behavior

1. Persistent Unfairness => Random performance



Ref: au-roeck-simulation-results-071707.pdf



Summary

1. NIC and switch implementation complexity is important.
2. Explicit feedback is better than Implicit Feedback if done properly
3. Rate based load sensor is more stable (less variance than) Queue-based Load Sensor
4. Accurately interpreting queue feedback requires knowledge of link speed and queueing architecture. Otherwise increase transient time.
5. Modification to data packets or turning around tags in the switches is not desirable for very high-speed networks.
6. Network overhead during congestion is important
7. Fairness shows up as predictable behavior