Modeling of BCN V2.0

Jinjing Jiang and Raj Jain
Washington University in Saint Louis
Saint Louis, MO 63130
Jain@wustl.edu

IEEE 802.1 Congestion Group Meeting, San Diego, July 19, 2006

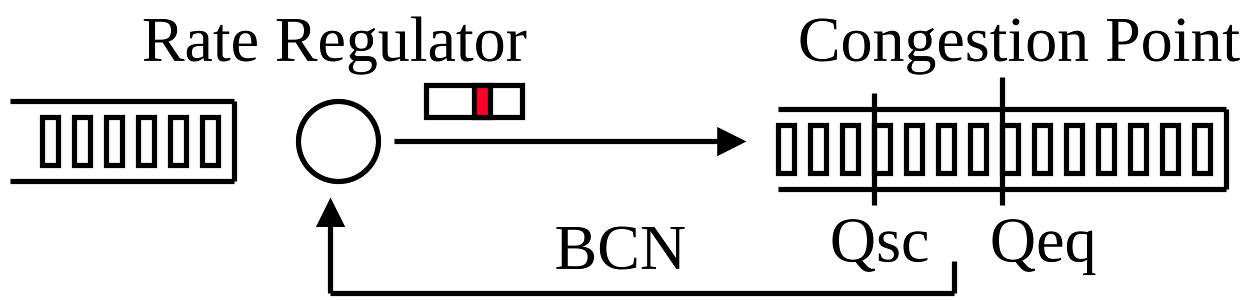
These slides are available on-line at:

http://www.cse.wustl.edu/~jain/ieee/bcn607.htm

au-jain-bcn-simulation-0706.ppt

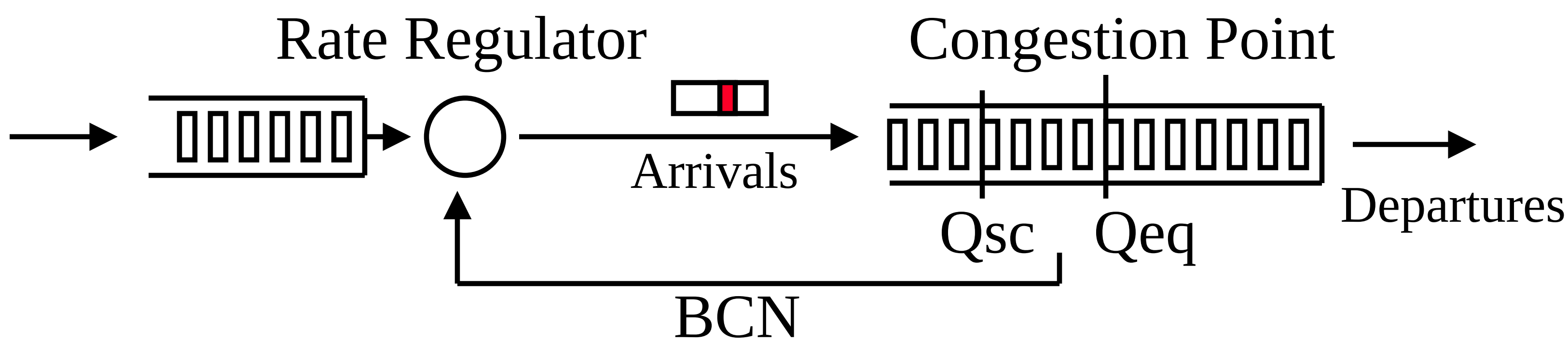
- **Goal**: Present *new results* since Denver/March 2006
- BCN Mechanism: Quick Review
- Action Items from Denver meeting
- New Analytical Results
- New Simulation Results

BCN Mechanism



- Backward Congestion Notification - Closed loop feedback
  - **Detection**: Monitor the buffer utilization at possible congestion point (Core Switch, etc)
  - **Signaling**: Generate proper BCN message based the status and variation of queue buffer
  - **Reaction**: At the source side, adjust the rate limiter setting according to the received BCN messages
    - Additive Increase Multiplicative Decrease (AIMD)
- Ref: new-bergamasco-backward-congestion-notification-0505.pdf

Parameters for BCN



- Key Parameters
  - Threshold for buffer:
    - *Qeq* (Equilibrium),
    - *Qsc* (Severe Congestion)
  - Sampling Probability Pm=0.01
- Queue Variation : *Qoff, Qdelta*
  - *Queue is sampled randomly with probability Pm*
  - *Qlen* (current length)
  - *Qoff = min{Qeq,Qeq-Qlen}, range [-Qeq, +Qeq]*
  - *Qdelta = max{min{2Qeq, #pktArrival-#pktDeparture},-2Qeq}, [-2Qeq, +2Qeq]*
    Note: Qoff is limited to –Qeq and Qeq. Qdelta is limited to –2Qeq and +2Qeq

AIMD Algorithm

- Source Rate *R*
- Feedback
  - $Fb = (Qoff - W \times Qdelta)$
- Additive Increase ($Fb > 0$)
  - $R = R + Gi \times Fb \times Ru$
- Multiplicative Decrease (Fb < 0)
  - $R = Min\{0, R \times (1 - Gd \times |Fb|)\}$
- Parameters used in AIMD:
  1. Derivative weight *W*
  2. Additive Increase gain *Gi*,
  3. Multiplicative Decease Gain *Gd*,
  4. Rate Unit *Ru*

Summary of Results (From March Meeting)

- BCN V2 simulation validate Cisco's results on throughput
- Time to Fairness and oscillation trade-off needs to be studied further
- Parameter setting needs more work
  Need to modify formula so that parameters are dimensionless
- Need to simulate more configurations:
  asymmetric, larger bandwidth delay, and multi-bottleneck cases

Issues to be Studied (From March Meeting)

1. Fix the dimensioning problem
2. Asymmetric Topology
3. Multi-bottleneck case
4. Larger/smaller Bandwidth×Delay product networks
5. Bursty Traffic
6. Non-TCP traffic
7. Interaction with TCP congestion mechanism
8. Effect of BCN/Tag messages getting lost

We present results on the first 6 issues + an analytical model + proportional fairness

Topics for Today

1. Analytical Model
2. Simulation Study: Convergence
3. Dimensioning Problem
4. Asymmetric Topology and Multiple Congestion Points
5. Max-min vs Proportional Fairness
6. Mixed TCP and UDP Traffic
7. Bursty Traffic
8. Other Issues
9. Bandwidth*Delay Product

1. Analytical Model

At the ith source, assume t_n is the time at nth rate update event. Then:

$$r_i(t_{n+1}) = \underbrace{e_i(t_n)^{\dagger}[1-|e_i(t_n)|G_d]r_i(t_n)}_{\text{Multiplicative Decrease}} + \underbrace{(1-e_i(t_n)^{\dagger})[r_i(t_n)+G_i|e_i(t_n)|R_u]}_{\text{Additive Increase}}, \quad (5)$$

where

$$e_i(t_n)^{\dagger} = \begin{cases} 1, \text{if } e_i(t_n) < 0, \\ 0, \text{otherwise.} \end{cases}$$

The above equation can be rewritten as:

$$r_i(t_{n+1}) - r_i(t_n) = |e_i(t_n)|\{G_i R_u - e_i(t_n)^{\dagger}(G_d r_i(t) + G_i R_u)\} \quad (6)$$

Following the stochastic approximation mentioned in [5] and assuming absolute value $|e_i(t)|$ is independent of the sign of $e_i(t)$, the above discrete time equation can be approximated into an ordinary differential equation (ODE)[5][10]:

$$\frac{dr_i(t)}{dt} = \frac{E\{|e_i(t)|\}\{G_i R_u - b_i(t)(G_d r_i(t) + G_i R_u)\}}{\mu_i(t)}. \quad (7)$$

$$\frac{dr_i(t)}{dt} = E\{|e_i(t)|\} r_i(t)(G_d r_i(t) + G_i R_u) \times \left\{ \frac{G_i R_u P_m}{S(G_d r_i(t) + G_i R_u)} - \frac{\partial G(\vec{r}(t))}{\partial r_i(t)} \right\} \quad (10)$$

The Lyapunov function for the ODE is:

$$V(\vec{r}) = \sum_{i=1}^{N} \int_0^{r_i} \frac{G_i R_u P_m}{S(G_d u_i + G_i R_u)} du_i - P(\vec{r}) = \frac{G_i R_u P_m}{S G_d} \log\left\{ \frac{G_d}{G_i R_u} r_i + 1 \right\} - P(\vec{r}) \quad (11)$$

Hence we can write the ODE as:

$$\frac{dr_i(t)}{dt} = E\{|e_i(t)|\} r_i(t)(G_d r_i(t) + G_i R_u) \frac{\partial V(\vec{r})}{\partial r_i} \quad (12)$$

Since $V(\vec{r})$ is strictly concave, it can reach a unique maximum over any bounded region. Also we have:

$$\frac{d}{dt} V(\vec{r}) = \frac{\partial V}{\partial r_i(t)} \frac{dr_i(t)}{dt} \quad (13)$$

$$= E\{|e_i(t)|\} r_i(t)(G_d r_i(t) + G_i R_u) \left(\frac{\partial V(\vec{r})}{\partial r_i} \right)^2$$

Hence V increases along any solution, which converges towards the unique maximum of V. This shows that the rates

- See Wash U technical report, which will be posted shortly

Analytical Model Results

Rate of Convergence:

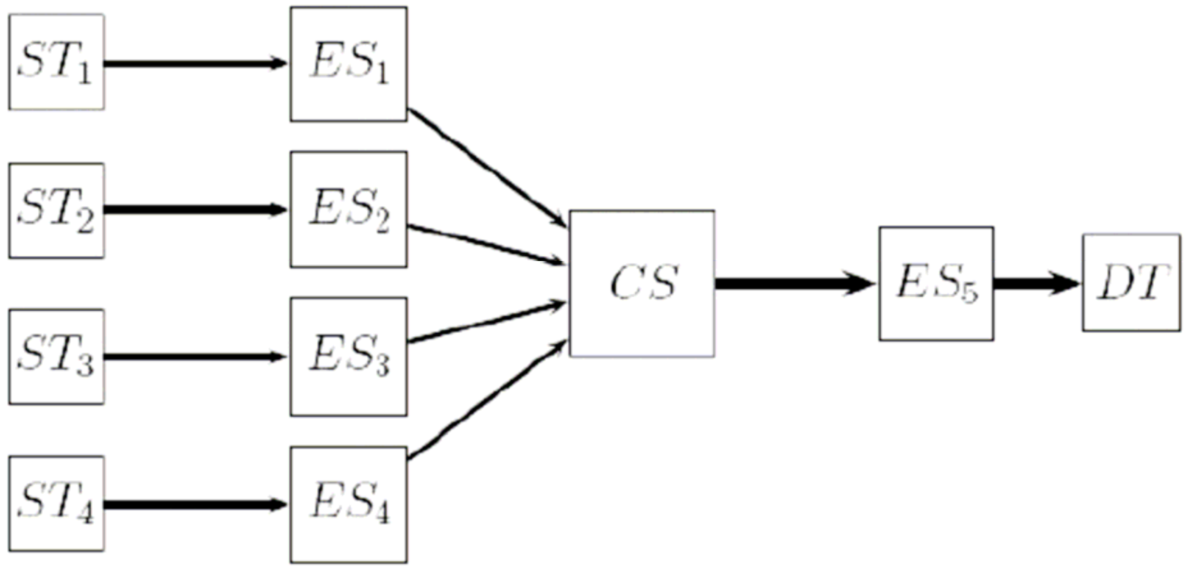
$$\Delta(t) = \frac{1}{S}G_i R_u P_m - (2G_d r_i + G_i R_u)h_i(t) - R_i(G_d R_i + G_i R_u)h_i'(t) \quad (17)$$

Conclusions:

- Increasing *Gi*, *Gd* and *Ru* will always increase the rate of convergence
- Feedback Delay= Sampling+Propagation+Switching+Reaction
  Sampling Delay = $\frac{S_p}{r_i(t)P_m}$ = Pkt Size/(input rate*Sampling P)
- Bandwidth*delay (delay=Propagation and switch delays) may not be related to the operation of the BCN mechanism
- Sampling probability P_m is the key parameter.
  Should be carefully selected considering current input rate r_i and packet size S_p

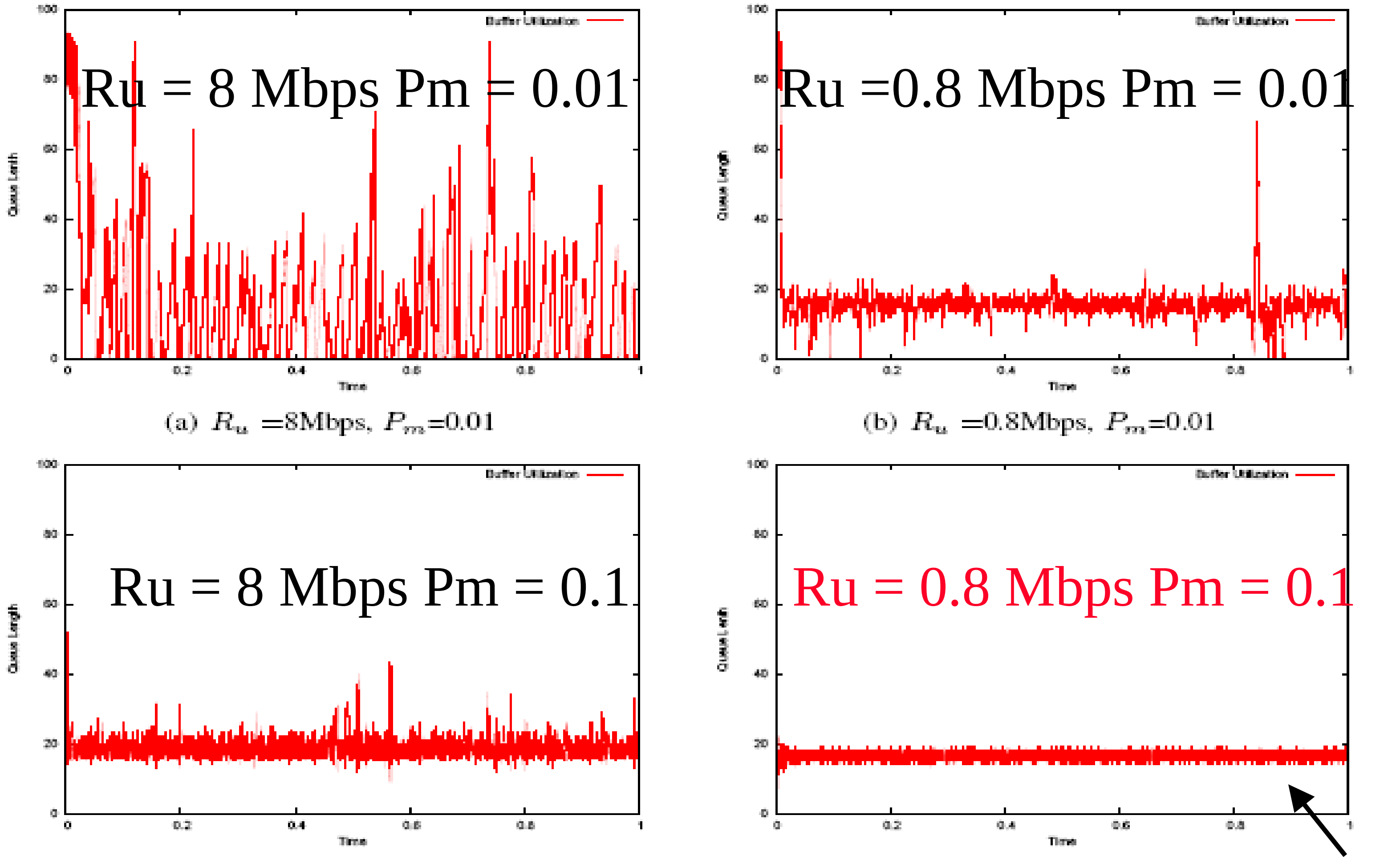
2. Simulation Study: Convergence

- Goal: To find optimal parameters for least oscillation
- Topology:



- Two Configurations: All links 1 Gbps or All links 10 Gbps
- Two Values for Rate Increment Ru
- Two values for Sampling Probability Pm
- A 2^2 Full factorial experimental design
  [See “Art of Computer Systems Performance Analysis”]

Simulation on Convergence – 1Gbps Link



(a) R_u =8Mbps, P_m=0.01

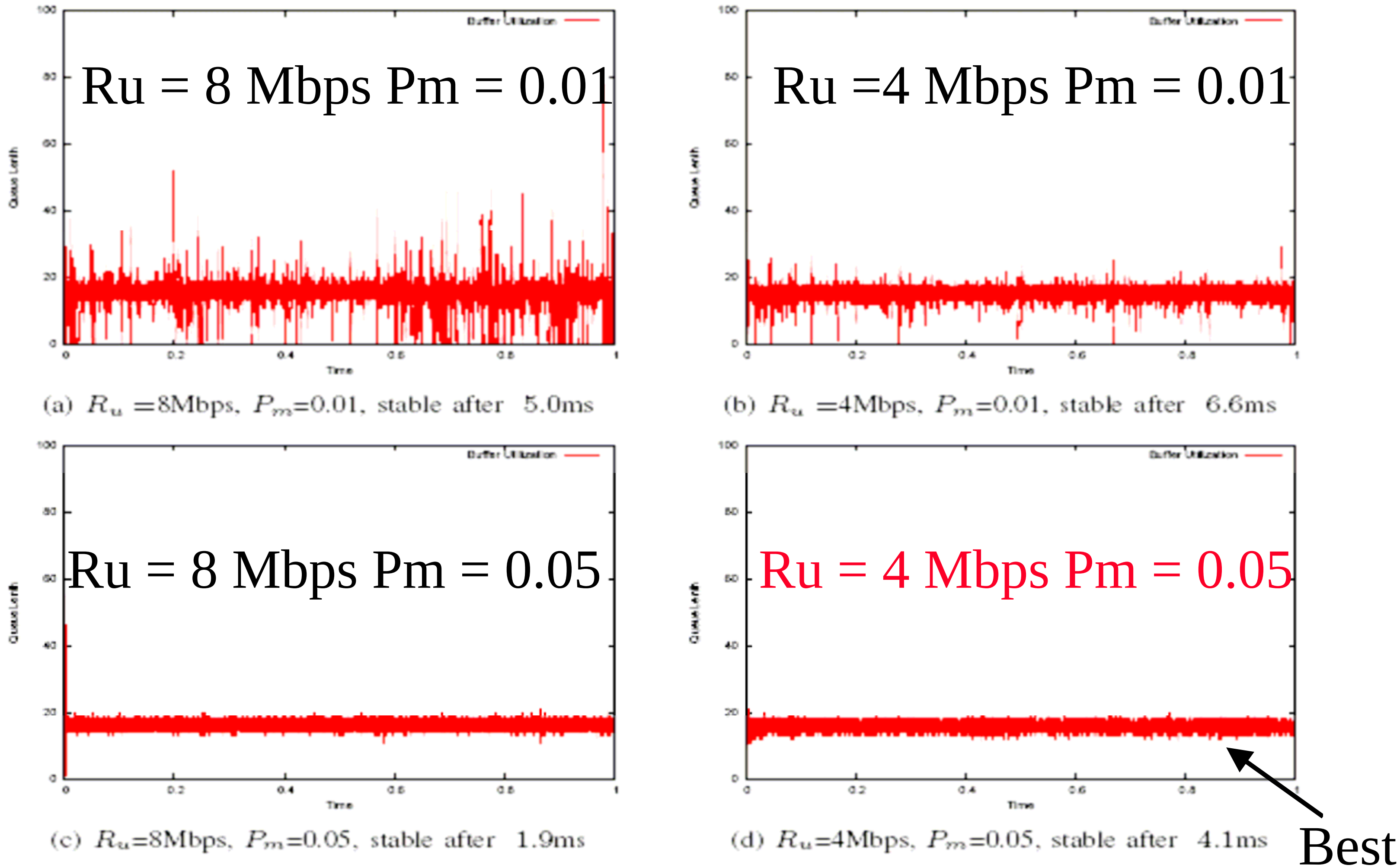
(b) R_u =0.8Mbps, P_m=0.01

(c) R_u = 8Mbps, P_m=0.1

(d) R_u = 0.8Mbps, P_m=0.1

Fig. 7. Performance of BCN for the symmetric topology with 1 Gbps bottleneck link

Simulation on Convergence - 10Gbps Link



(a) R_u =8Mbps, P_m=0.01, stable after 5.0ms

(b) R_u =4Mbps, P_m=0.01, stable after 6.6ms

(c) R_u=8Mbps, P_m=0.05, stable after 1.9ms

(d) R_u=4Mbps, P_m=0.05, stable after 4.1ms

Fig. 8. Performance of BCN for the symmetric topology with 10 Gbps bottleneck link

Simulation on Convergence: Conclusions

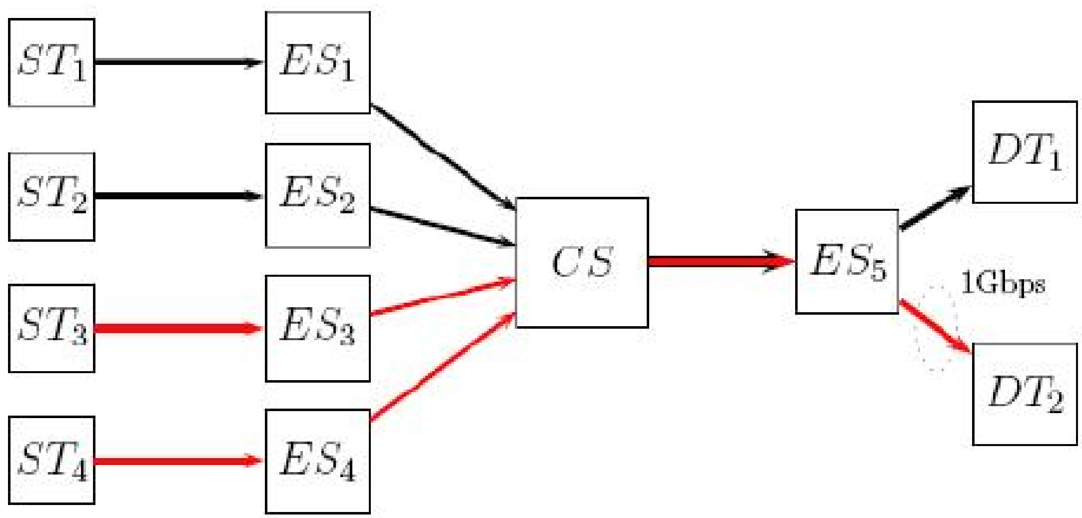
- Large P_m, small R_u make oscillations smaller in both cases
  - Larger $P_m \Rightarrow$ excessive signaling overhead
  - Small $R_u \Rightarrow$ long time to converge
- Parameters depend upon bottleneck link speed

3. Dimensioning Problem

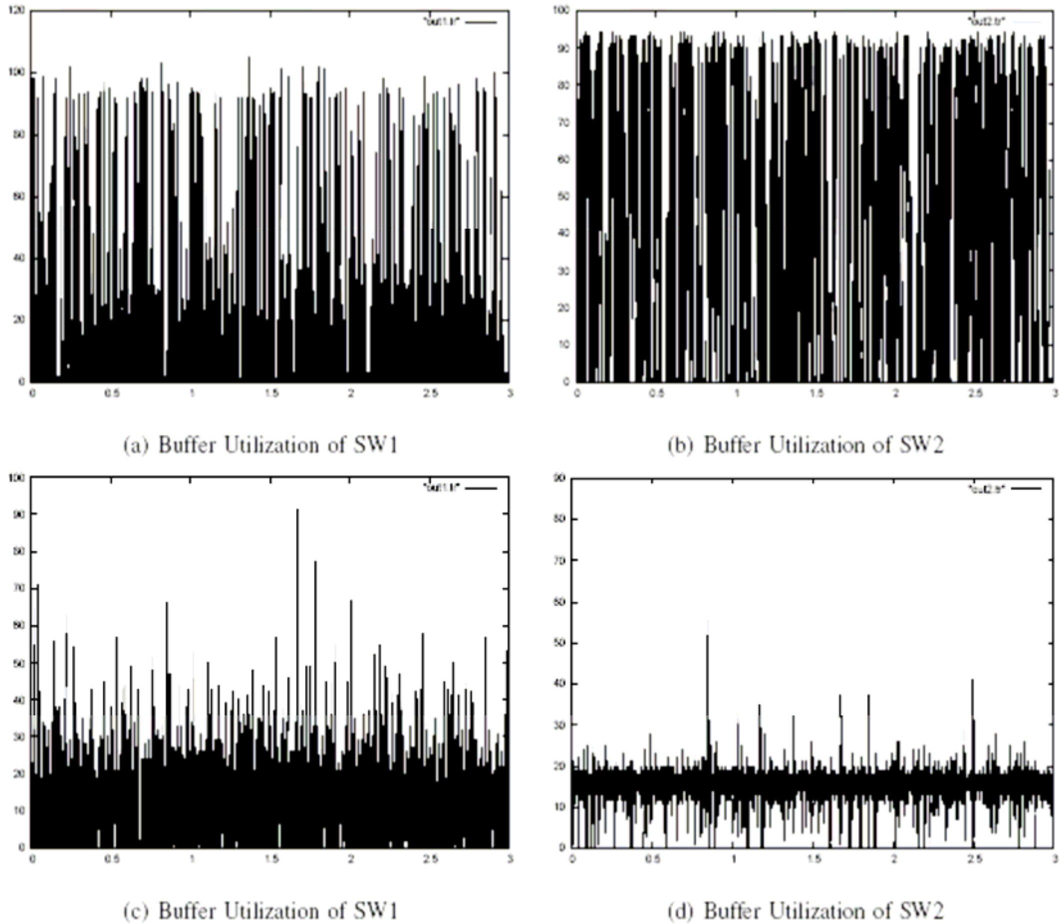
- 1 Gbps Link and 10 Gbps Link
- Same P_m and R_u leads to instability
- Sources need to know the bottleneck link capacity
  Need to add bottleneck rate to the BCN message.
- Current BCN mechanism sets 5 Gbps as the initial rate for rate limiter. If congested link capacity (1 Gbps) is not known at the source, it takes long time for the sources to decrease their rates to less than 1 Gbps.
- The rate increase unit R_u should be set as *C/N for some N*. If not, there are large oscillations

4. Asymmetric Topology and Multiple Congestion Points

Topology: Only one link is 1Gbps, others are all 10Gbps



A Simulation Result on BCN and the Enhanced Version



Not Stable

Stable with oscillations

Source Speed?

(a, b) $P_m = 0.01$, $R_u = 8$ *Mbps* for all links (c, d) $Pm = 0.01$, $Ru = 8$ *Mbps* for 10Gbps links, $Pm = 0.1$, $Ru = 0.8$ *Mbps* for 1Gbps link

5. Max-min vs Proportional Fairness

- Max-Min Fairness: Assumes linear utility of data rate
  Maximize the minimum allocation w/o exceeding the capacity

maximize $f_1(\vec{r}) = min\{r_1, r_2, r_3, r_4, r_5\}$

subject to

$$r_1 + r_2 + r_3 + r_4 + r_5 \leq C$$

- Proportional Fairness: Data rate has a log utility
  Maximize the sum of the logs w/o exceeding the capacity

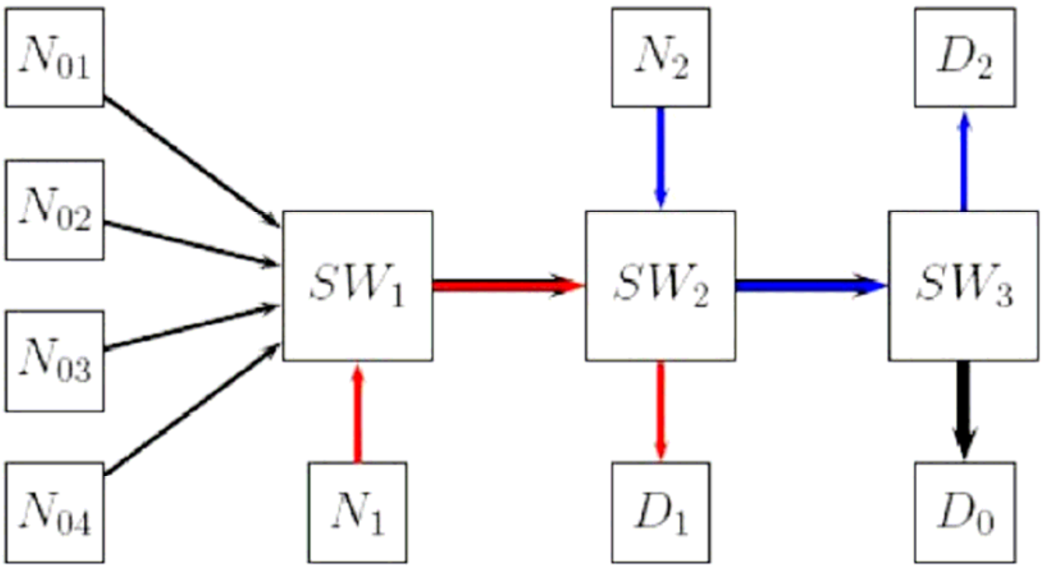
maximize $f_3(\vec{r}) = log(r_1) + log(r_2) + log(r_3) + log(r_4) + log(r_5)$

subject to

$$r_1 + r_2 + r_3 + r_4 + r_5 \leq C$$

Simulations for Fairness

- Simulation using Parking Lot Topology



- Max-min fairness: $R_{01} = \ldots = R_{04} = R_1 = R_2 = C/5$
- Proportional fairness: $R_{01} = \ldots = R_{04} = C/6$; $R_1 = R_2 = C/3$

Simulation on Fairness

- Simulation results for Parking Lot topology(Gbps):
  - $R_{01}=1.4643$, $R_{02}=1.4532$
  - $R_{03}=1.5430$, $R_{04}=1.7291$
  - $R_1=3.0795$, $R_2=3.0185$
  - $2(R_1+R_2)/(R_{01}+\ldots+R_{04})=1.97 \approx 2$

6. Mixed TCP and UDP Traffic

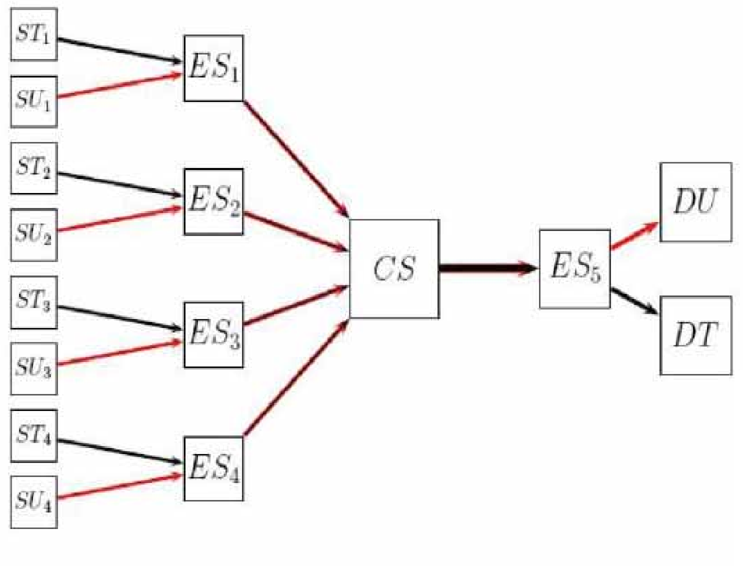
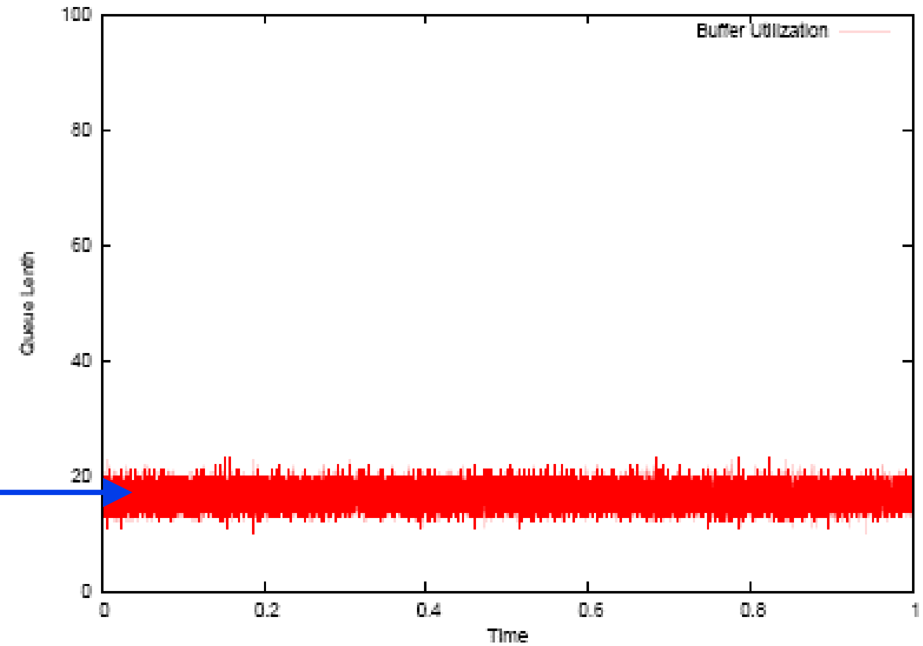


Fig. 7. Topology for Mixed Traffic

Stable

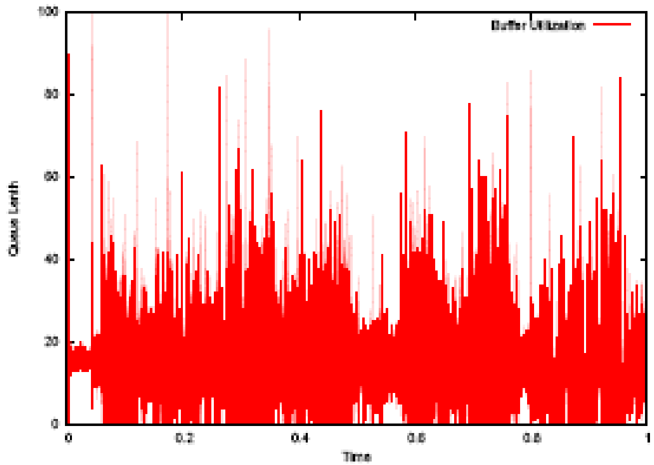


Mixed TCP and UDP Traffic: Results

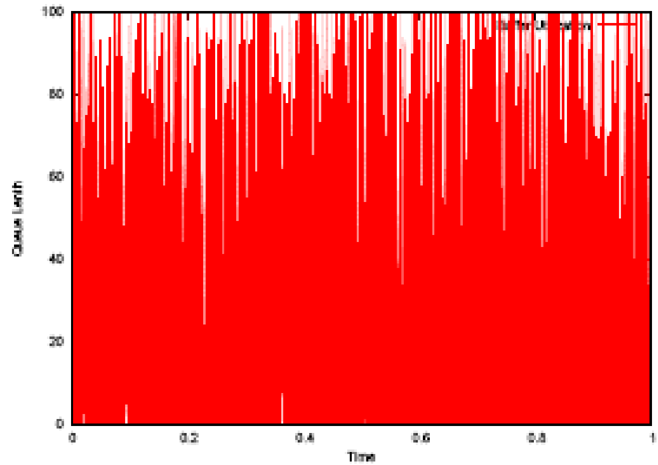
- TCP average throughput:1.16 Gbps
  UDP average throughput:1.34 Gbps
  No significant performance difference compared with TCP-only workload
- Since rate limiter is implemented at the sources, UDP rate is also controlled
- UDP has slightly higher throughput than TCP
- TCP has its own congestion control mechanism, rate limiter's rate is the peak rate is can achieve.

7. Bursty Traffic

- If the burst period is much longer than the settling time of the system, the system is still stable.
  If not, the system tends to be unstable.
- Settling time ≈ 4 ms for the above simulations
- UDP is bursty with Pareto distribution for Burst size, Topology for mixed traffic



(a) Average burst period: 100ms



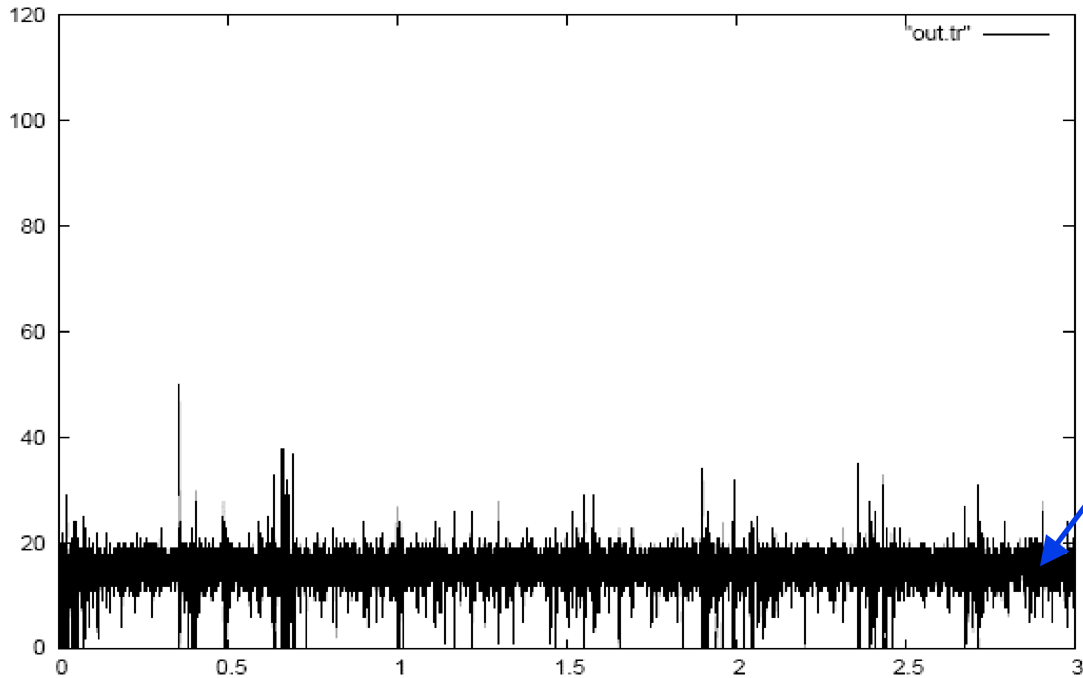
(b) Average burst period: 1ms

8. Other Issues

- BCN(0,0) is sent when the queue is severely congested. It asks source to stop and restart at 1/100th of the link capacity after a some random interval.
- This leads to low throughput.
- In the original BCN message, sending back $Q_{off}=0$, $Q_{delta}=0$ to indicate the severe congestion, which may cause low link utilization
  - $Q_{off}=0$, $Q_{delta}=0$ is very likely when the queue operates at the equilibrium
  - Our results in March presentation have larger oscillation is purely because of the different use of BCN(0,0) message

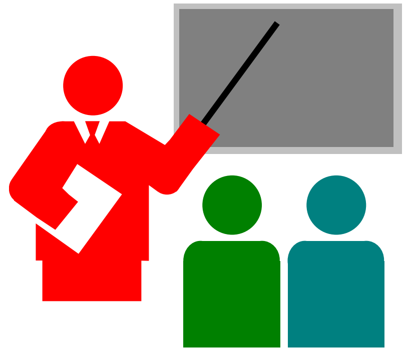
9. Bandwidth*Delay Product

- In this simulation, the symmetric topology is used, propagation delay is 9.5 us (originally it is 0.5 us), which to some extent, we use this to simulate 7 hops network from the source to the congested switch



The queue is stable. As aforementioned, propagation delay have small effect on the feedback delay

Summary



1. We have developed an analytical model of BCN that allows us to study the effect of various parameters on convergence and stability
2. BCN achieves Proportional Fairness (vs max-min fairness)
3. Need to feedback bottleneck capacity in the BCN message
4. Optimal parameters depend upon the bottleneck capacity
5. Performance of BCN (including bottleneck rate feedback):
    1. TCP and UDP mixed traffic
    2. Performance with Multiple Congestion Point
    3. Bursty Traffic
    4. Different Bandwidth*Delay product networks

Action Items for Next Time

- Indicate:
  - Number of flows
  - Details of TCP traffic
  - Packet drops
  - End-to-end latency
- Investigate Burst to settling time ratio (reduce the range)
- Try Mix of TCP+UDP traffic w/o BCN
- Interaction between BCN and latency based TCP flow controls
- Transients at the end of congestion bursts.
  How fast the rates pick up?
- Larger probability of sampling with two way traffic
- Two switches sending BCNs to the same source
  The BCNs giving conflicting increase/decrease to a source.