Transport Mechanisms for Data Centers: The Averaging Principle

Mohammad Alizade, Berk Atikoglu,
Abdul Kabbani, Ashvin Lakshmikantha,
Rong Pan, Balaji Prabhakar, Mick Seaman

Overview

- Paper on QCN with same authors recently written; has two main parts
  - QCN: Algorithm and theoretical model
    - This has been presented to the WG in July '08 at the Denver meeting
  - The Averaging Principle
    - A control-theoretic idea which can be applied to general control systems, ***not just*** congestion control systems, and which makes them more robust to increases in loop delays
    - Underlies the reason for the good stability of the QCN and BIC algorithms

- We describe the AP, apply it to BCN
  - We have also applied it to other algorithms in the Internet context

Background to the AP

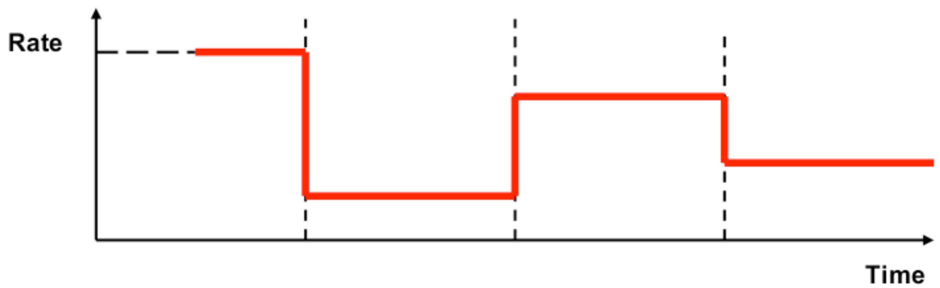
- When the lags in a control loop increase, the system becomes oscillatory and eventually becomes unstable

- Feedback compensation is applied to restore stability; the two main flavors of feedback compensation in are:
  1. Determine lags (round trip times), apply the correct “gains” for the loop to be stable (e.g. XCP, RCP, FAST).
  2. Include higher order queue derivatives in the congestion information fed back to the source (e.g. REM/PI, BCN).

- The Averaging Principle is a different method
  - It is suited to Ethernet where round trip times are unavailable
  - It is also a simpler method of coping with increasing lags than sending higher order derivatives
    - E.g. think of BCN v1.0 and v2.0

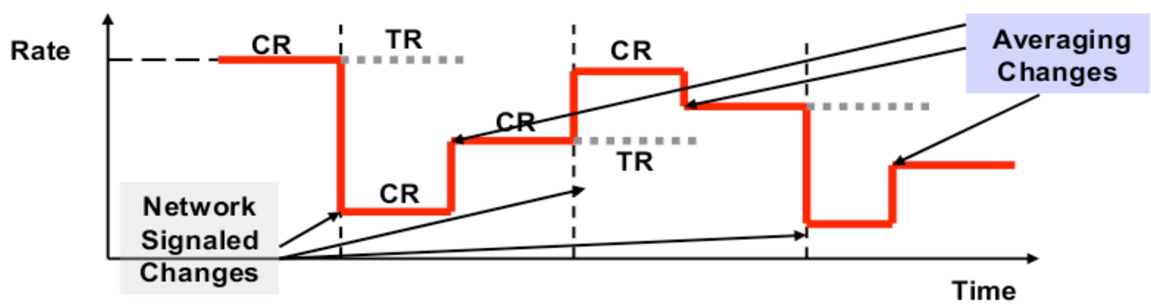
The Averaging Principle (AP)

- A source in a congestion control loop is instructed by the network to decrease or increase its sending rate (randomly) periodically



- AP: a source obeys the network whenever instructed to change rate, and then ***voluntarily*** performs ***averaging*** as below

TR = Target Rate
CR = Current Rate



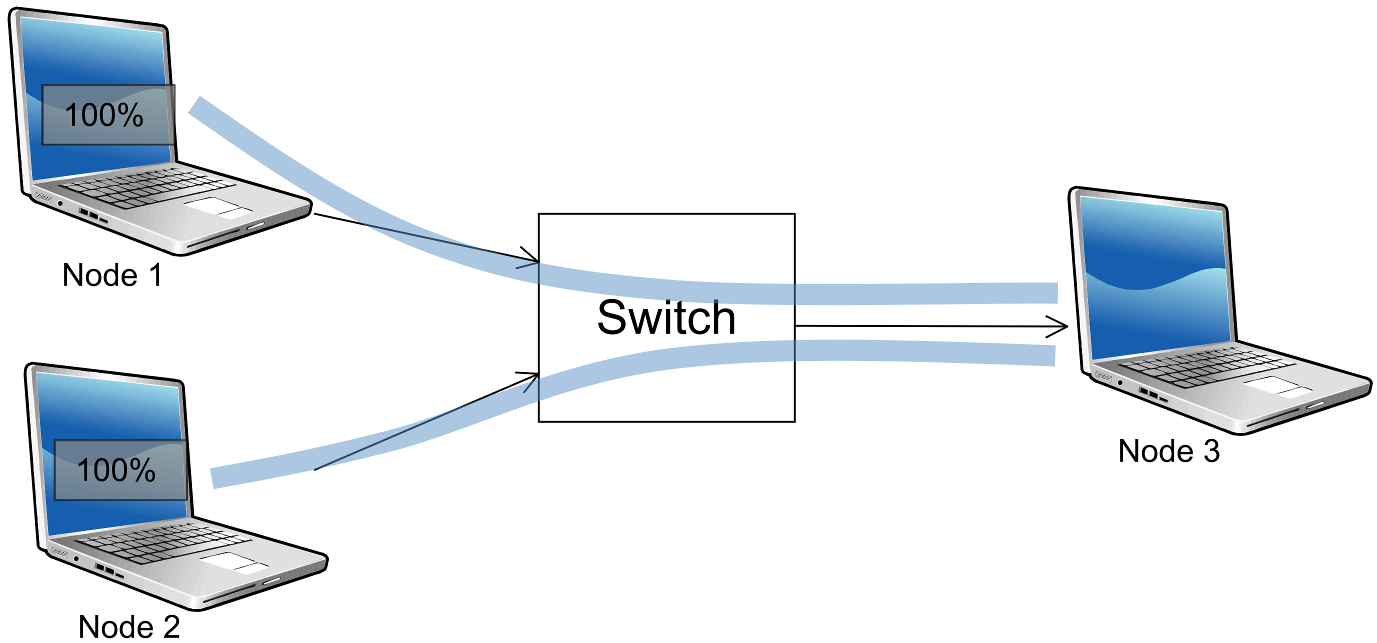
Averaging applied to BCN

- Algorithm
  - When Fb (positive or negative) is received:
    - Apply Fb to modify Current Rate
    - Set Target Rate = old Current Rate
  - Apply averaging after 50 packets are sent:
    - Current Rate = *alpha* * Target Rate + (1 – *alpha*) * Current Rate

- Recall: Rule for modifying current rate in BCN

$$R \leftarrow \begin{cases} R + G_i R_u F_b & \text{if } F_b \geq 0 \\ R(1 - G_d |F_b|) & \text{if } F_b < 0 \end{cases}$$

BCN with AP: Scenario and Workload

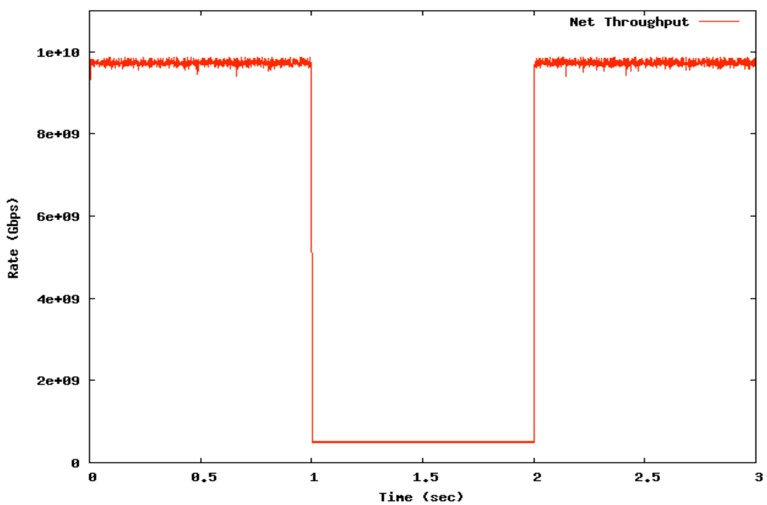


- 2 flows destined to node destined node 3 (First flow from Node 1, second flow from Node 2)
- Each flow is at maximum rate (10Gbps)
- Traffic: uniform
- Duration: 3s
- Service rate at switch is decreased to 0.5G from 1s to 2s
- RTT: varying

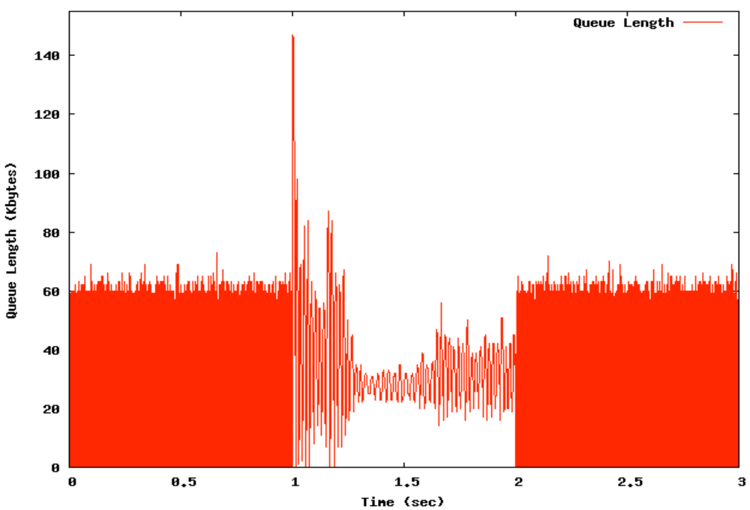
BCN parameters

- – Qeq = 375
- – Qsc = 1600
- – Qmc = 2400
- – Qsat disabled
- – Ecm00 disabled
- – Gi = 0.53333 (varies with RTT)
- **– W= 0 or 2**
- – Gd = 0.00026667
- – Ru = 1,000,000
- – Rd = 1,000,000
- – Td = 1ms
- – Rmin = 1,000,000

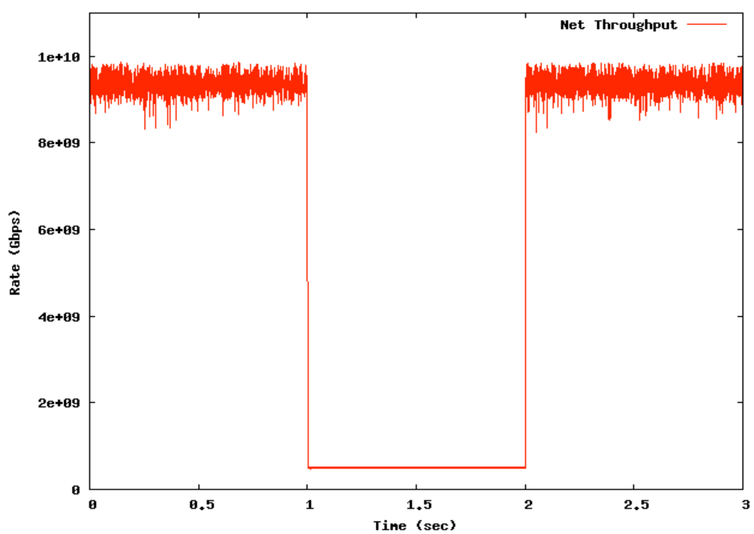
BCN v1.0
Fb = Qoff



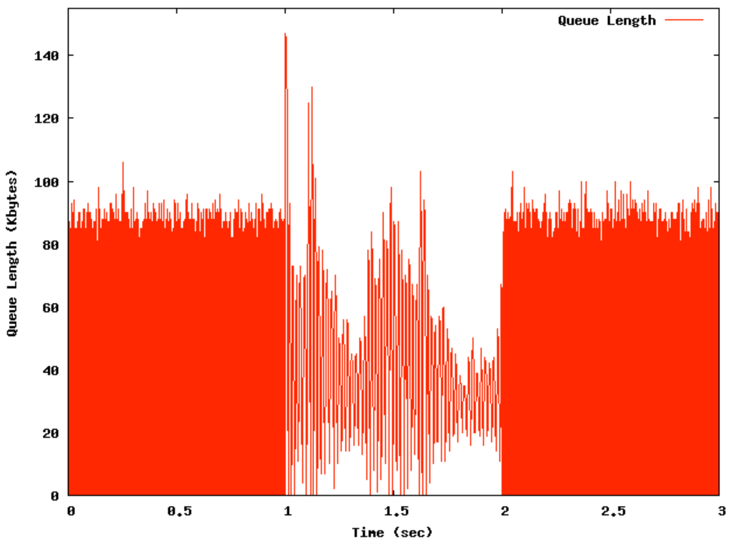
BCN v1.0
50 us



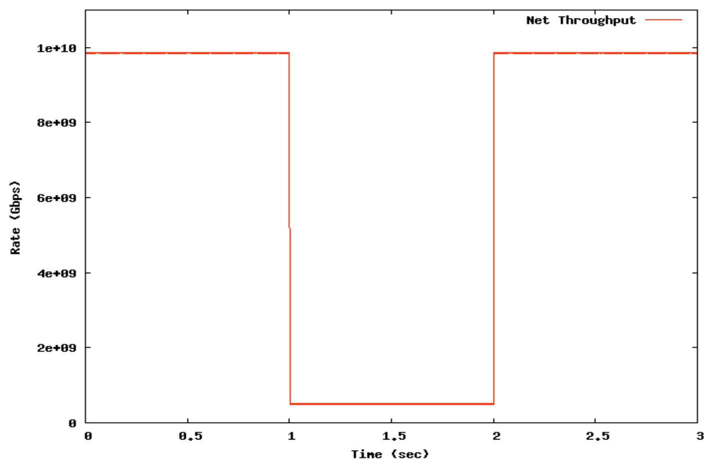
Fb = Qoff



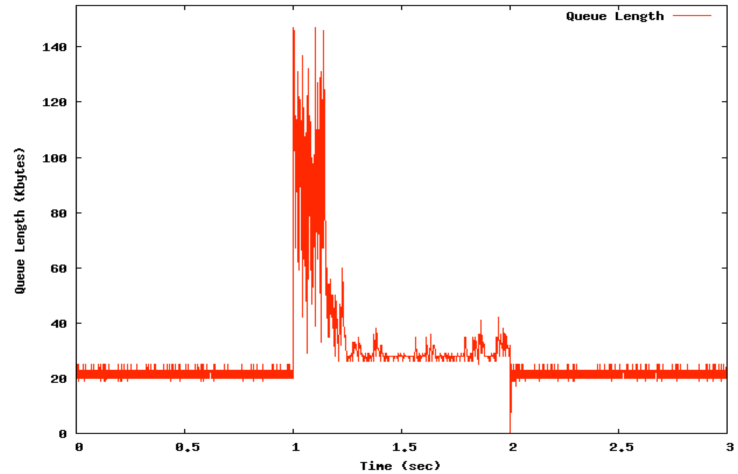
BCN v1.0
100 us



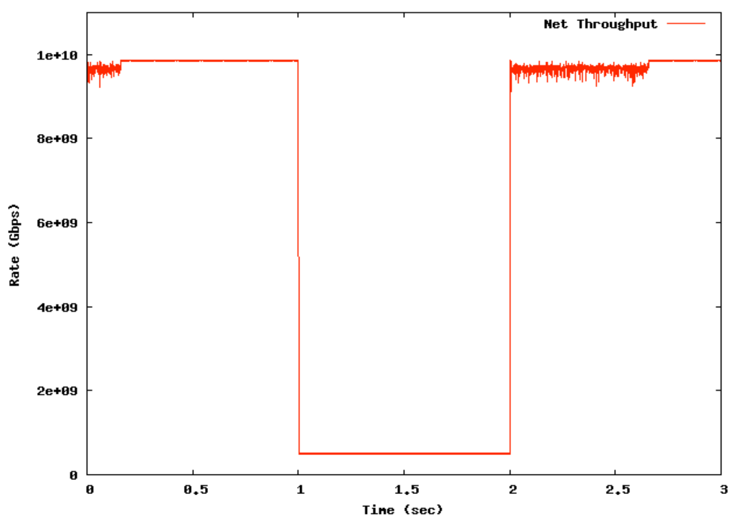
BCN v2.0
Fb = Qoff + 2 Qdelta



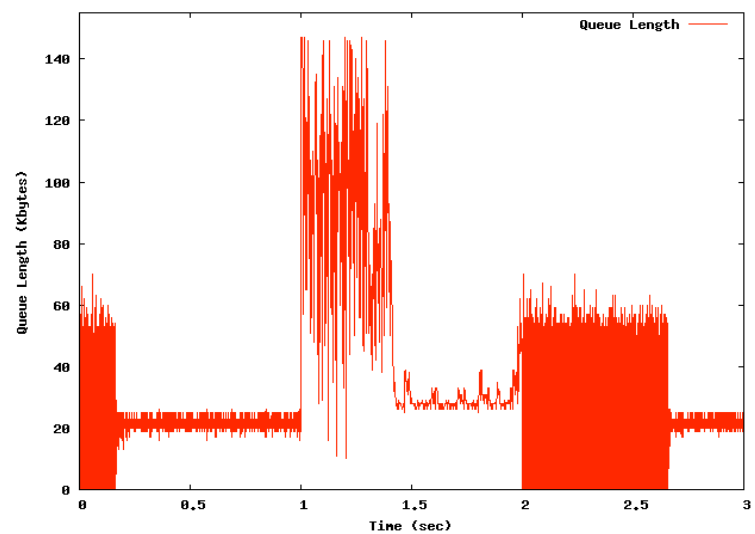
BCN v2.0
50 us



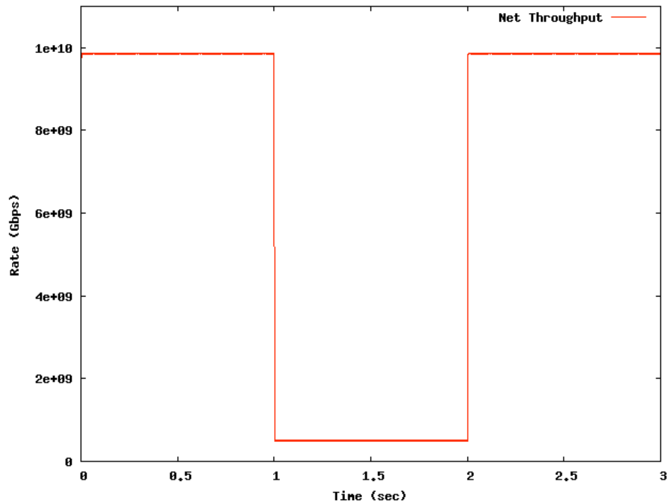
Fb = Qoff + 2Qdelta



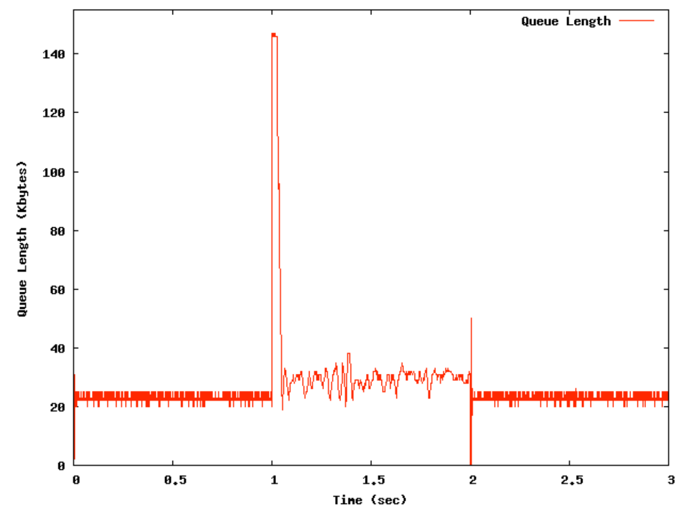
BCN v2.0
100 us



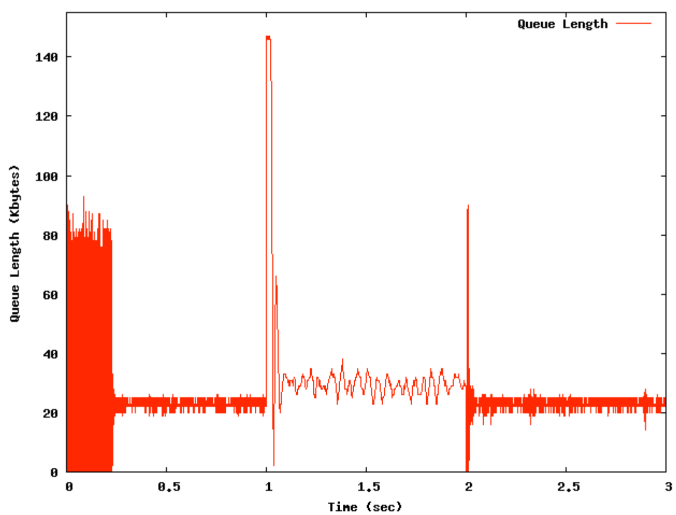
BCN v1.0 with AP
Fb = Qoff



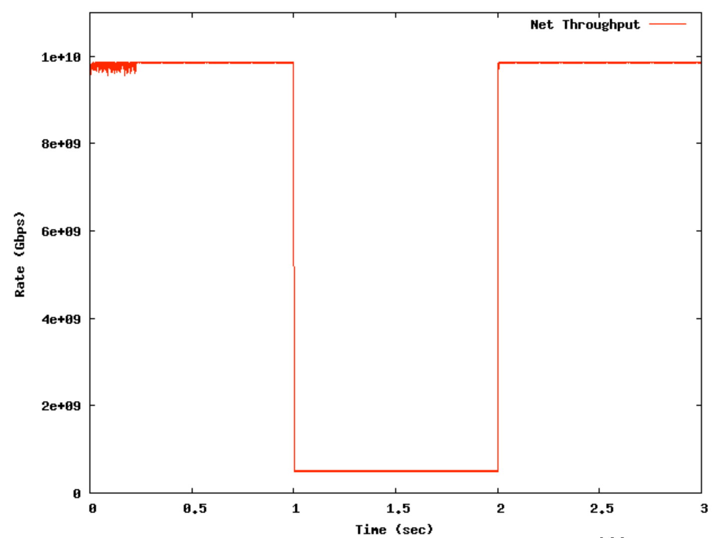
BCN *v1.0* with AP 50 us



Fb = Qoff



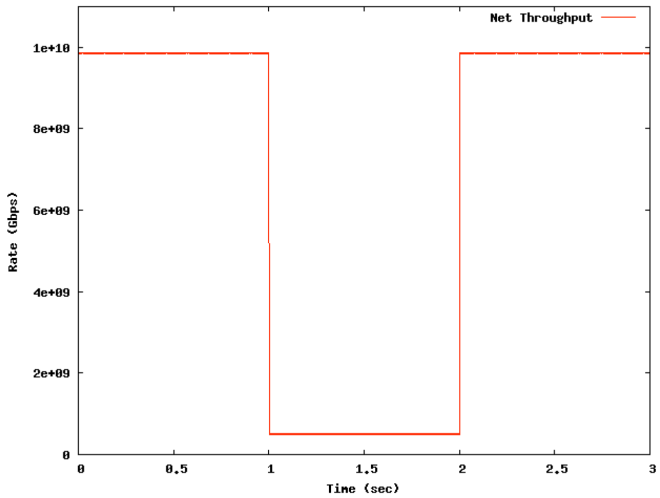
BCN *v1.0* with AP 100 us



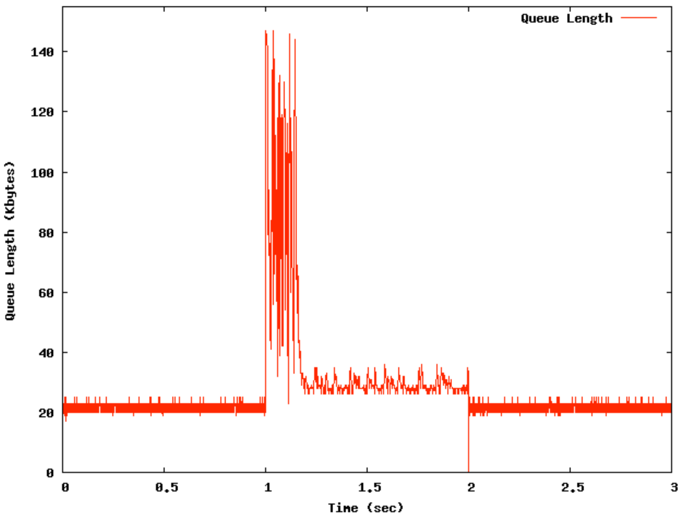
Summary of BCN v1.0 with AP

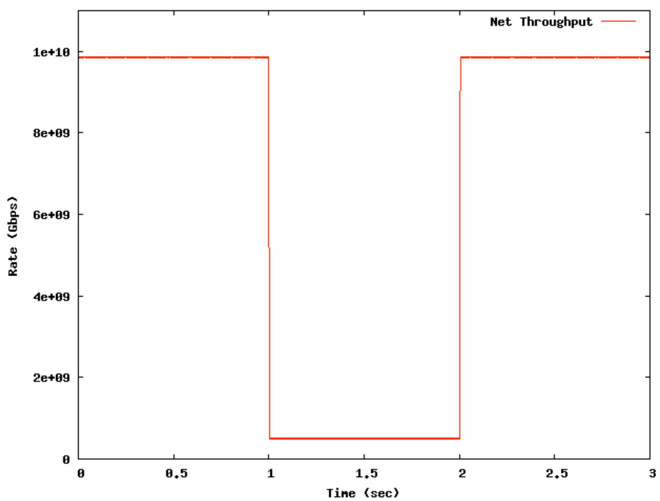
- We see that the AP provides an automatic stabilization to BCN v1.0 which is at least as good as that provided by BCN v2.0
  - The difference is that the AP does not require Qdelta
  - Qdelta requires a change at all switches, which we can avoid using the AP
- Now, suppose we take BCN v2.0, where Qdelta is already available
  - We saw in the Los Gatos meeting in Jan 08 that BCN v2.0 needs gain adjustments to be stable at large RTTs (200 us or so)
  - Can the AP be applied to BCN v2.0 to improve its stability?

BCN v2.0 vs BCN v2.0 with AP
RTT = 10 us

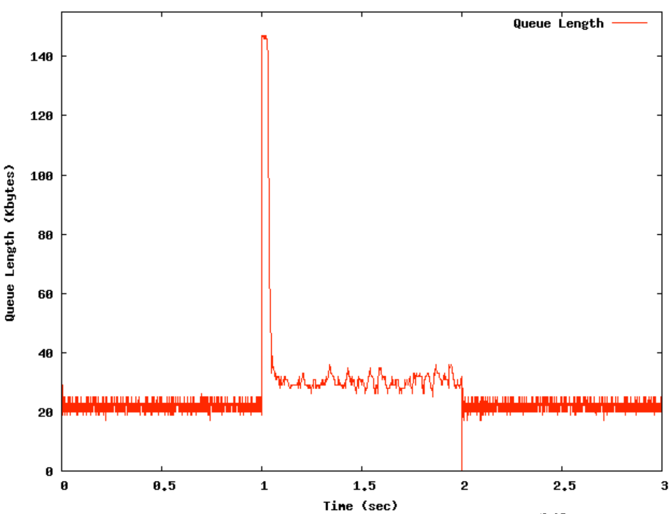


BCN *v2.0*

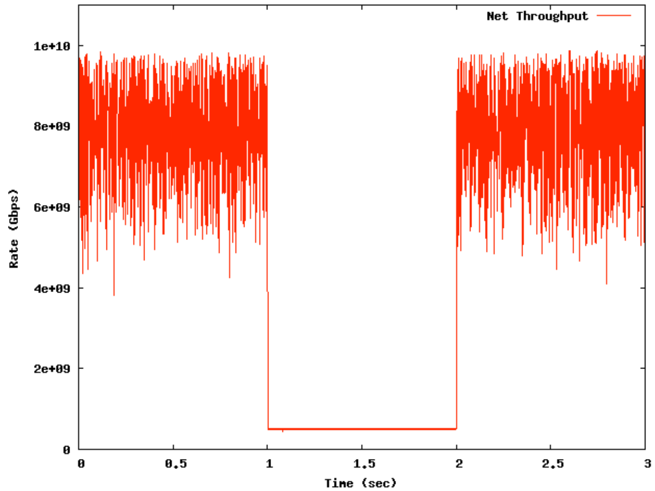




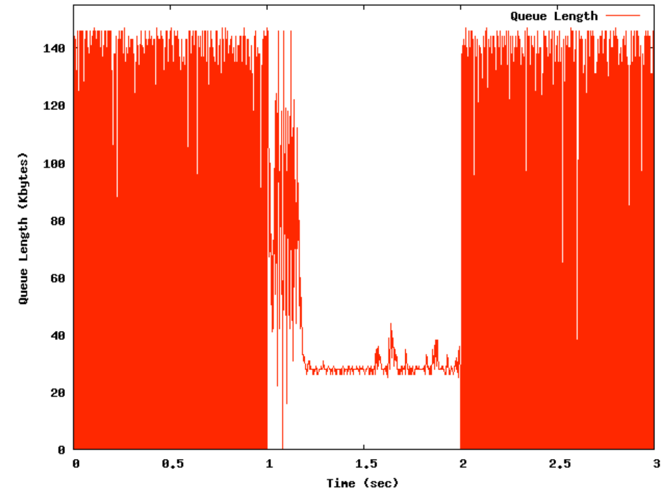
BCN *v2.0*
with AP

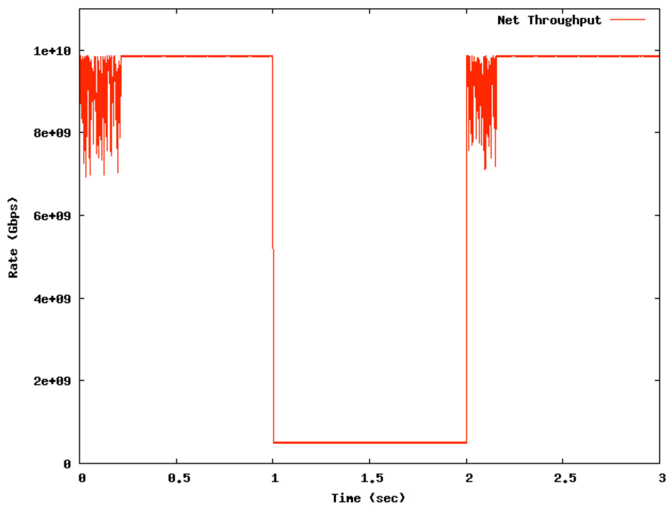


BCN v2.0 vs BCN v2.0 with AP
RTT = 250 us

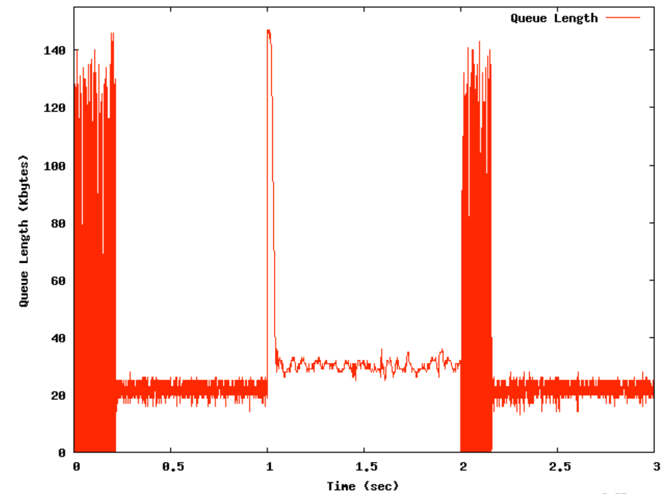


BCN *v2.0*
(Los Gatos, Jan ‘08)





BCN *v2.0*
with AP

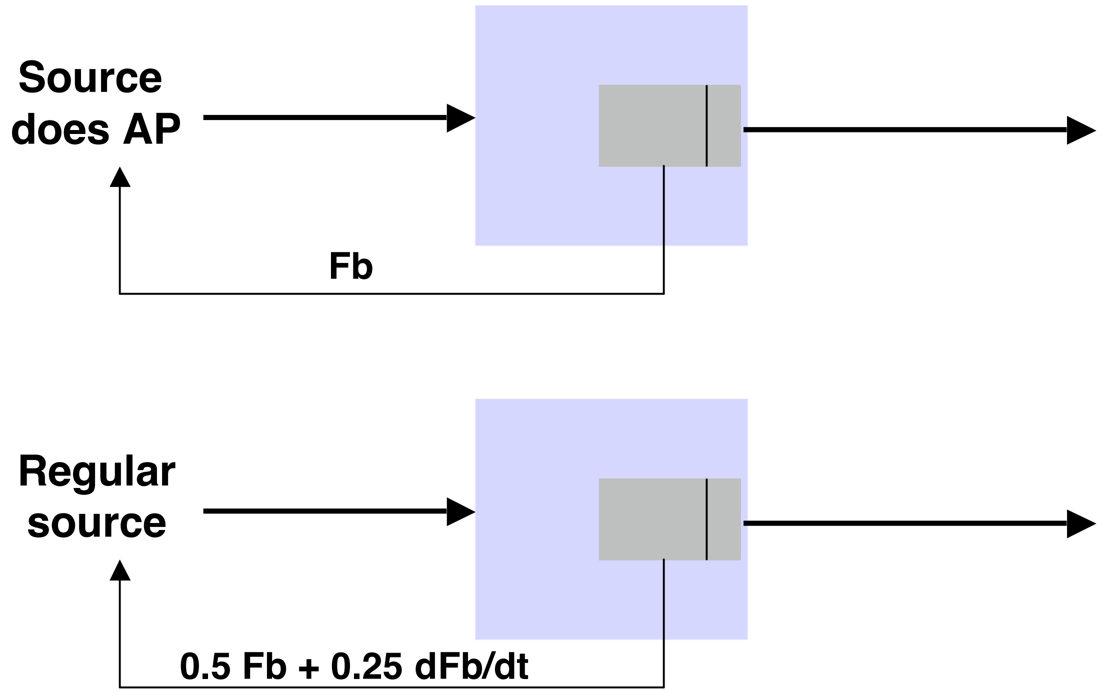


Understanding the AP

- As mentioned earlier, the two major flavors of feedback compensation are:
  1. Determine lags, chose appropriate gains
  2. Feedback higher derivatives of state

- We prove that the AP is in a sense equivalent to the second option above!
  – This is great because we don't need to change network routers and switches
  – And the AP is really very easy to apply; no lag-dependent optimizations of gain parameters needed

AP Equivalence: Single Source Case



- Systems 1 and 2 are discrete-time models for an AP enabled source, and a regular source respectively.
- ***Main Result:*** Systems 1 and 2 are algebraically equivalent. That is, given identical input sequences, they produce identical output sequences.
  - Therefore the AP is equivalent to adding a derivative to the feedback!
  - This is exactly what was done to BCN from v1.0 to v2.0

Conclusion

- The AP is a simple method for making many control loops (not just congestion control loops) more robust to increasing lags
- Gives a clear understanding as to the reason why the BIC-TCP and QCN algorithms have such good delay tolerance: they do averaging repeatedly
  - There is a theorem which deals explicitly with the QCN-type loop
- Variations of the basic principle are possible; i.e. average more than once, average by more than half-way, etc
  - The theory is fairly complete in these cases