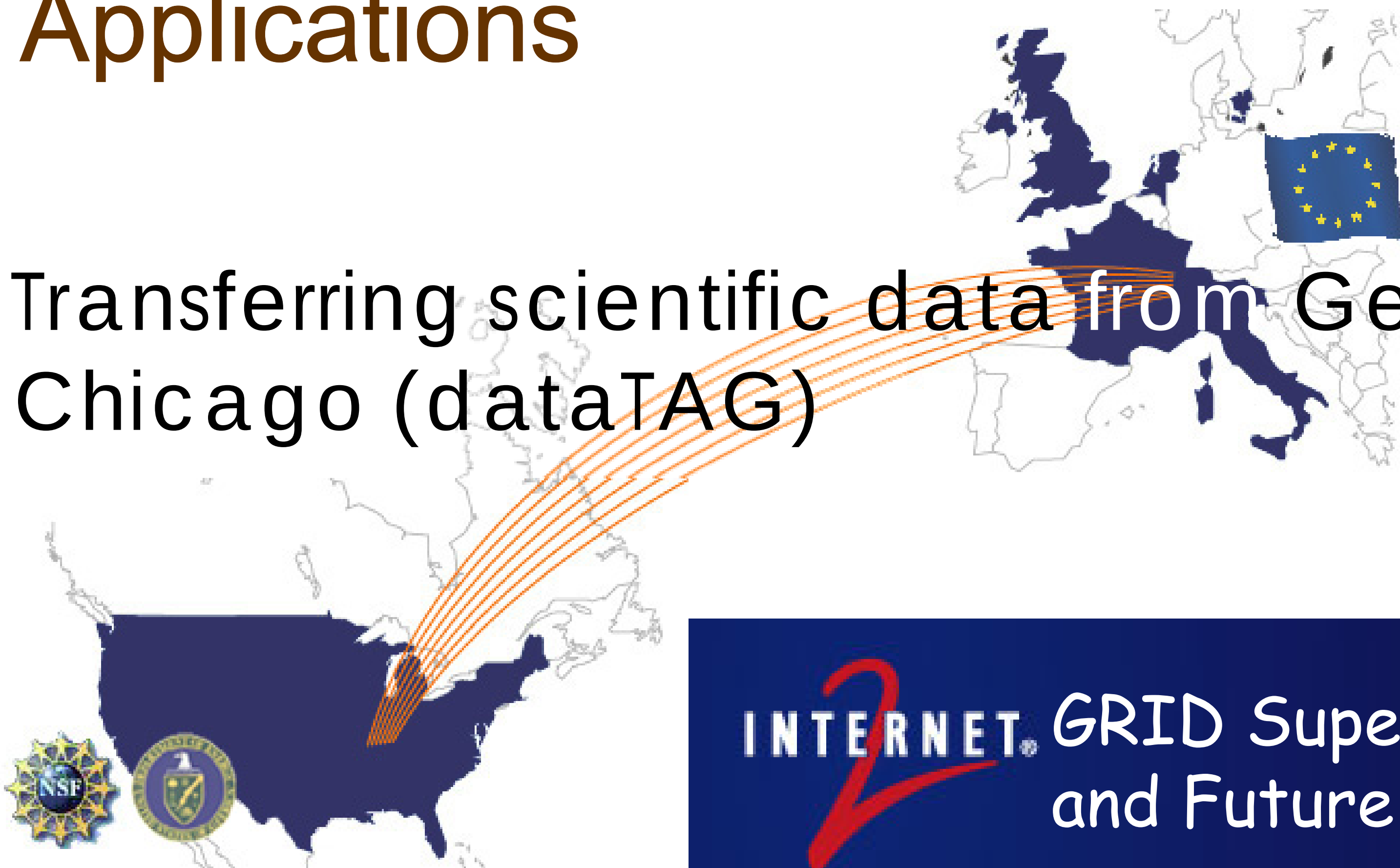
The Effect of Router Buffer Size on HighSpeed TCP Performance

Dhiman Barman

Joint work with Georgios Smaragdakis and Ibrahim Matta

Applications

Transferring scientific data from Geneva to Chicago (dataTAG)



INTERNET2 GRID Super Computing and Future Internet

Human Genome Project and Future Analysis Project

TCP in High Speed Networks

Regular TCP is *too slow* in High Speed Networks

Throughput (Mbps)	**RTT between Losses**	**Window size**	**Losses**
1	5.5	8.3	0.02
10	55.5	83.3	0.002
100	555.5	833.3	0.0002
1000	5555.5=9 mins	8333.3	0.00002
10000	55555.5=1.5h	83333.3	0.000002

Packet size = 1500 bytes, RTT = 0.1s

A High Speed Protocol has to be:

- Scalable
- Stable
- Responsive
- Intra protocol fair

High Speed Protocols

- **HighSpeed TCP**
  - **First standardized HighSpeed TCP**
  - **AIAD in log-scale**
- Scalable TCP
  - MIMD in linear scale
- BIC TCP
  - Binary search for available bandwidth
- FAST TCP
  - Reacts to both packet loss and queueing delay

Motivation

- All High Speed variants of TCP scale over Gigabit linespeed with rule-of-thumb buffer size
  - Rule-of-thumb: Buffer Size = RTT * BW
  - Ideal static transmission window
    - high utilization and minimal loss

- Focus has been given to improve other High Speed TCP features.

Our Contribution

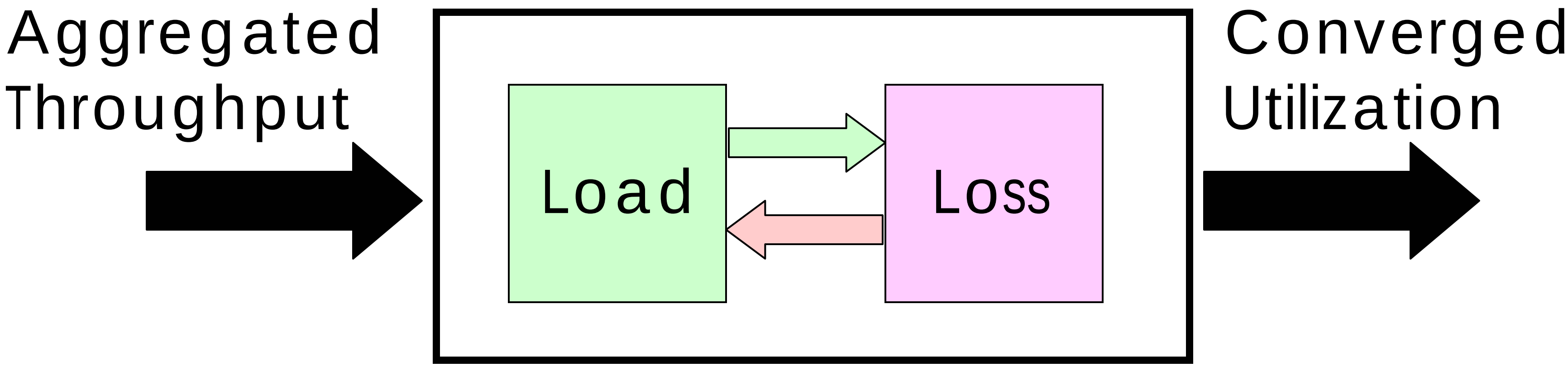
We investigate:

- Interaction between protocol and router buffer size and linespeed
- Do we need to increase buffer size linearly with linespeed ?
- How Router Buffer affects HighSpeed TCP behavior ?

Why Router Buffer is a concern

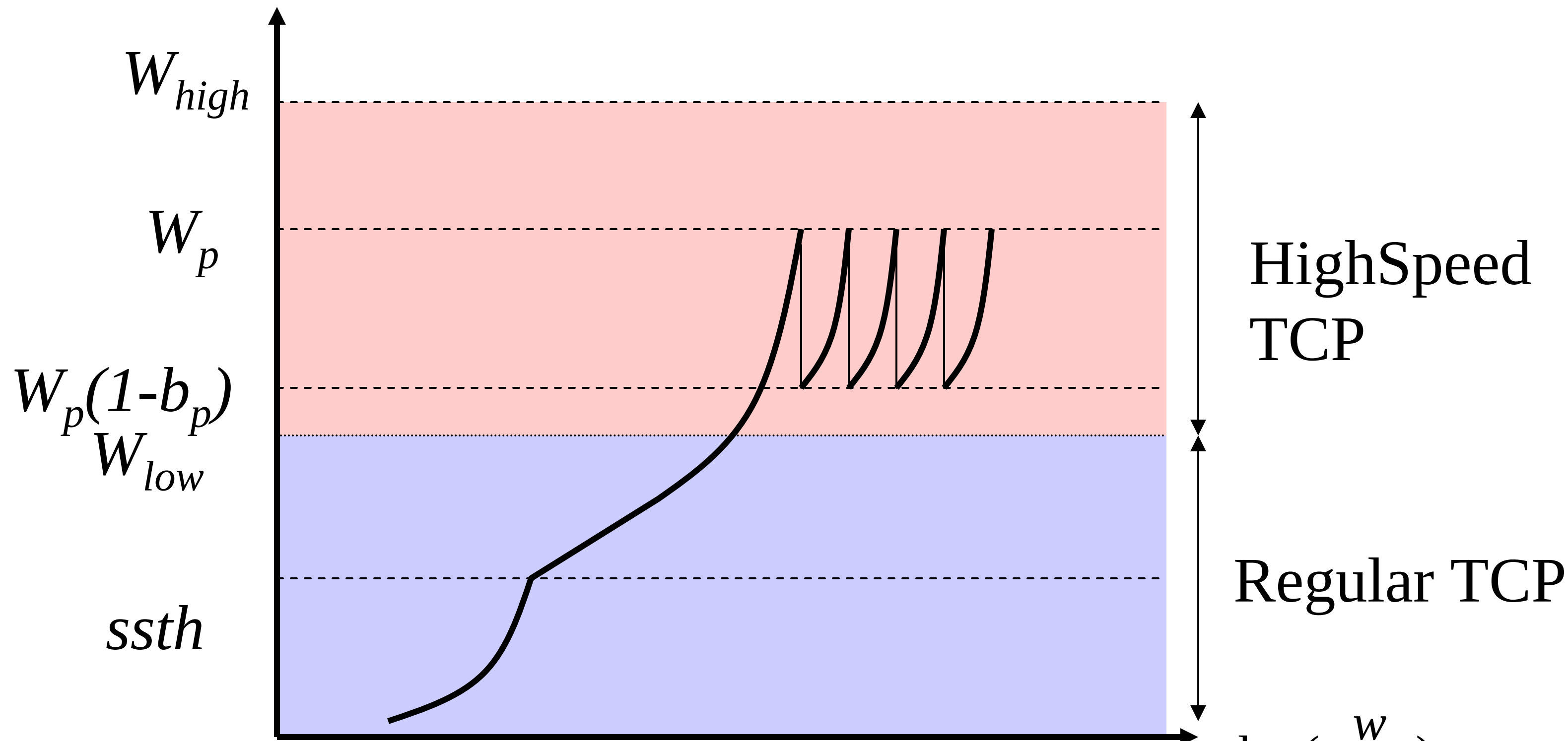
- **10Gb/s linecard**
  - Rule-of-thumb about 300Mbytes of buffer size
  - Read and write 40 byte packet every 32ns
- **Memory technologies**
  - SRAM: fast, require more devices, expensive, energy consuming
  - DRAM: slow, require less devices, cheap, less energy consuming
- **Problem gets severe**
  - At 40Gbps, 100Gbps, …
  - In an all-optical router for a backbone network where buffer size is limited to 5-10 packets in delay lines

Joint Model of HighSpeed TCP and Router Buffer Size



- Compute Throughput as function of loss rate
- Compute Loss rate as function of Throughput
- Iterate to compute fixed-point

HighSpeed TCP Window Dynamics

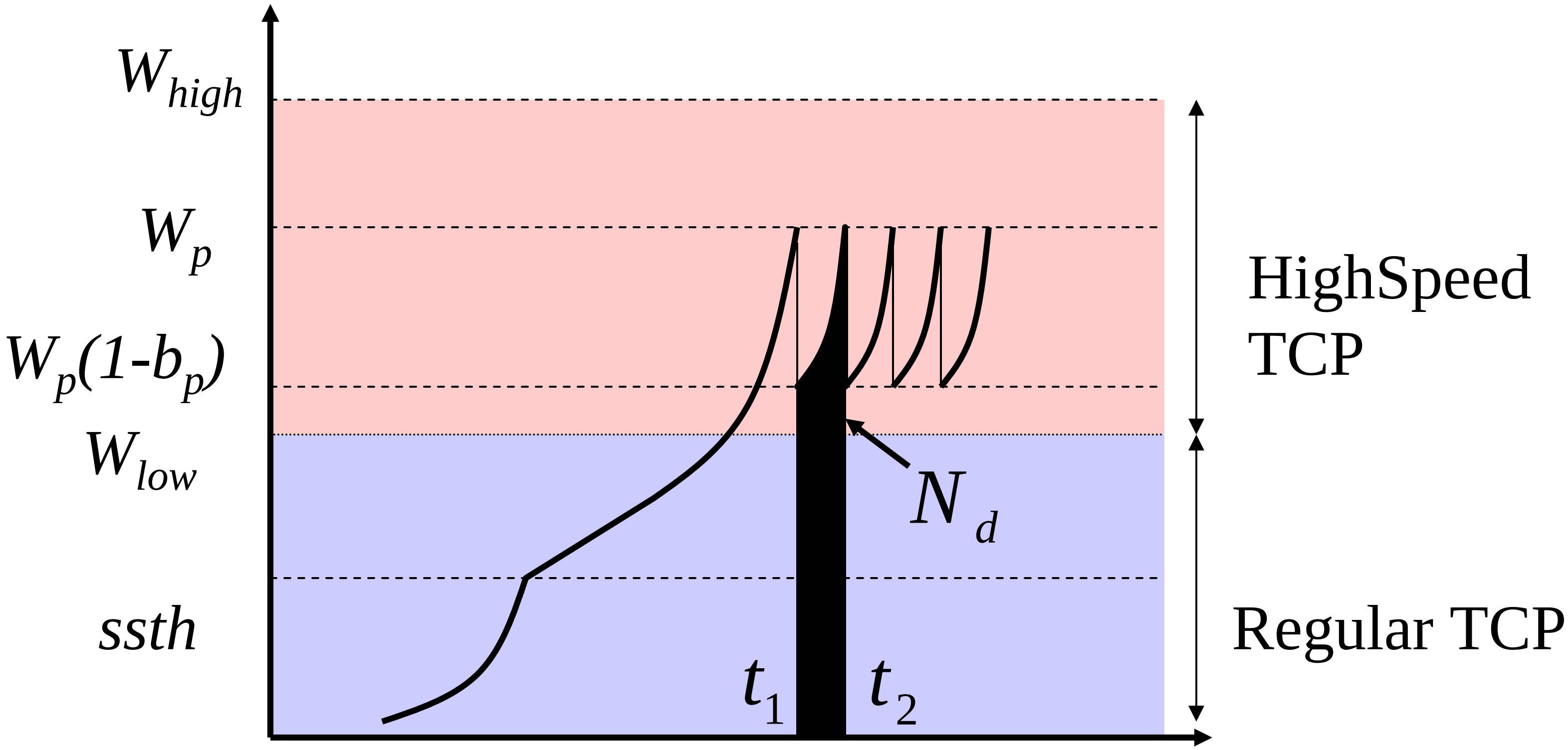


$$a(w) = \frac{2w^2 b(w) p(w)}{2 - b(w)}; b(w) = \frac{\log(\frac{w}{w_{low}})}{\log(\frac{w_{high}}{w_{low}})}(b_{high} - 0.5) + 0.5$$

$$w_{t+R} = \begin{cases} 2w_t, w_t < ssth \\ w_t + 1, ssth \le w_t \le w_{low} \\ w_t + a(w_t), w_t > w_{low} \end{cases}$$

$$p(w) = e^{\left(\frac{\log(\frac{w}{w_{low}})}{\log(\frac{w_{high}}{w_{low}})}\log(\frac{p_{high}}{p_{low}}) + \log(p_{low})\right)}$$

HighSpeed TCP Window Dynamics



We derived Analytical Closed Form Expression for Throughput

$$\frac{dw(t)}{dt} = \frac{a(w(t))}{RTT} = \frac{2w^2 b(w)\, p(w)}{RTT\,(2 - b(w))}$$

$$\lambda = \frac{N_d}{t_2 - t_1} \approx \frac{\omega_l}{R}\left(\frac{p}{p_l}\right)^{\frac{\log(w_h / w_l)}{\log(p_h / p_l)}}$$

Analytical Loss Rate Derivation

$$\rho = \frac{1}{C}\sum_{i=0}^{N} \lambda_i(p, RTT_i),$$

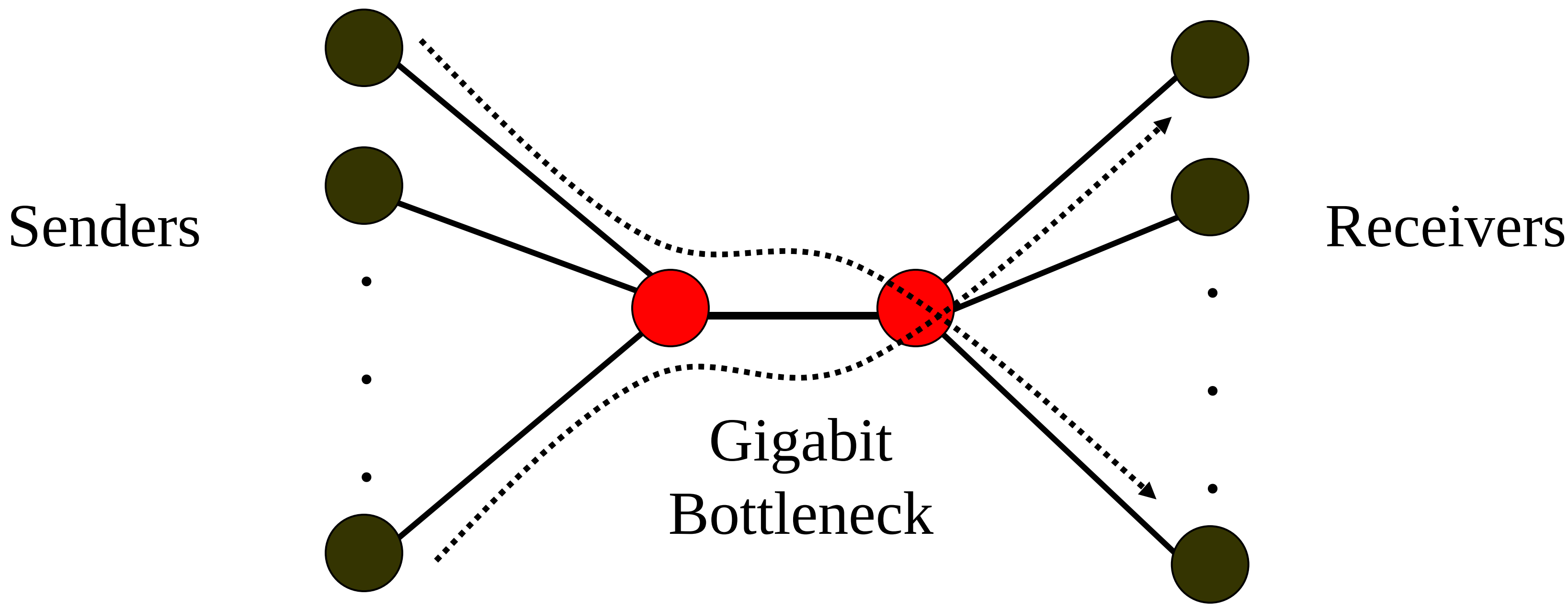
Drop Tail [M/M/1/K/FCFS]

$$p = \frac{(1-\rho)\rho^K}{1-\rho^{K+1}}, R_i(k) = 2d_i + \frac{MSS}{C}\left(\frac{1}{1-\rho} - K\frac{\rho^K}{1-\rho^K}\right)$$

RED [M/M/1/K/RED]

$$p_{red} = \sum_{i=0}^{K} \pi(i)d(i), \quad \pi(j) = \frac{(1-\rho)\rho^j}{1-\rho^{K+1}}, d(k) = \begin{cases} 0, k \le \min_{th} \\ \dfrac{(k-\min_{th})p_{\max}}{(\max_{th}-\min_{th})}, k > \min_{th} \end{cases}$$

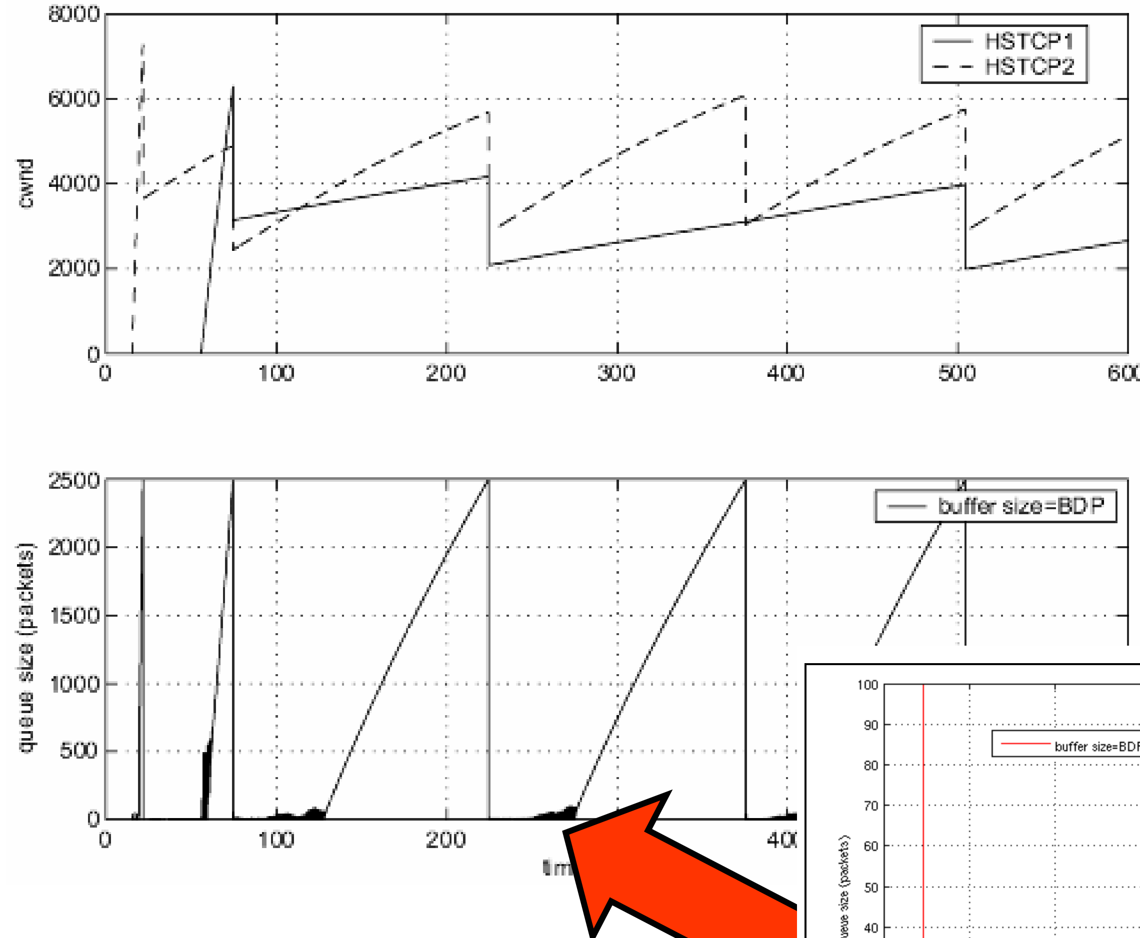
Simulation Setup

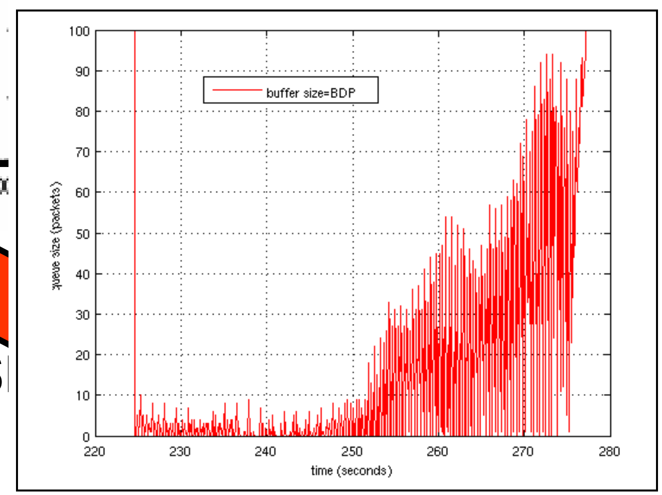


- In traffic mix, long flows (elephants) dominate buffer requirements

- Buffer size varies as fraction of bandwidth-delay product

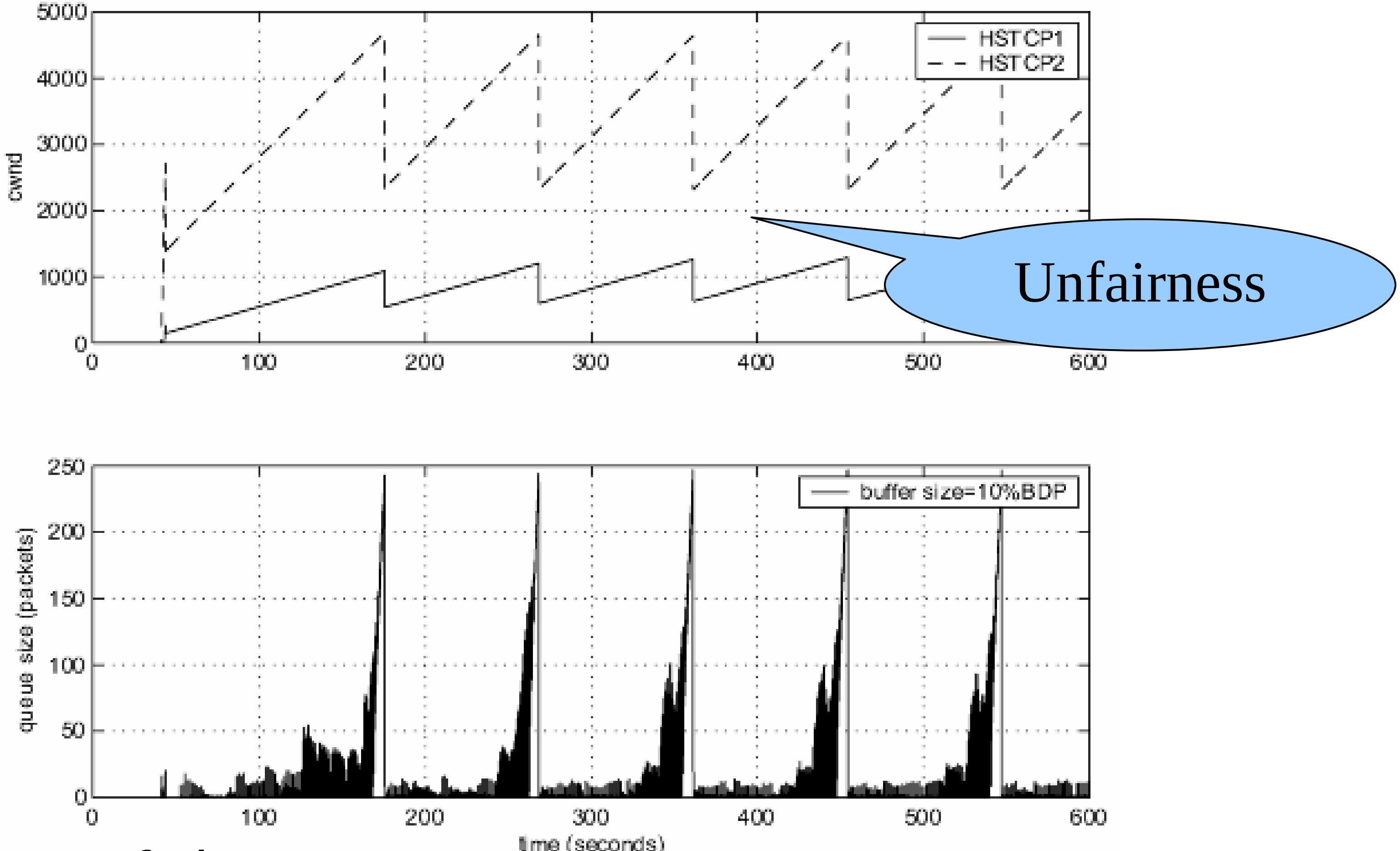
Rule-of-Thumb Buffer Sizing





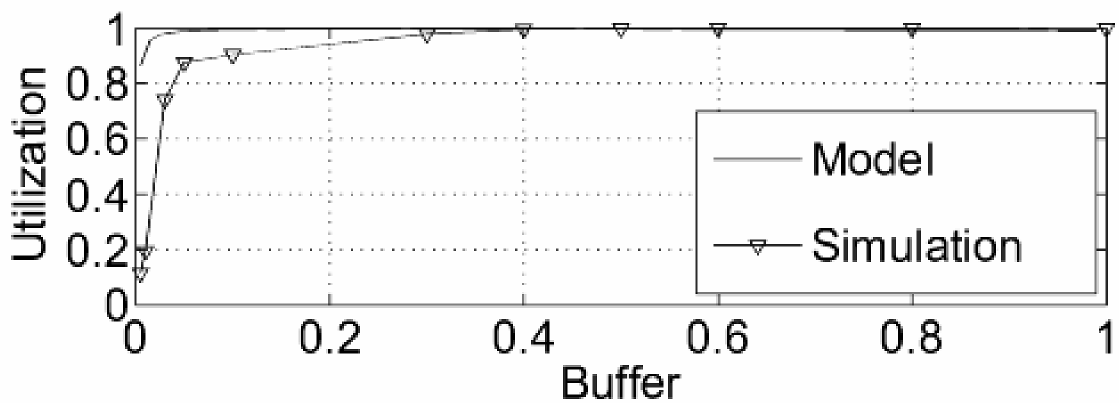
High utilization is achieved for a s
synchronized flows

Reduced Buffer Sizing

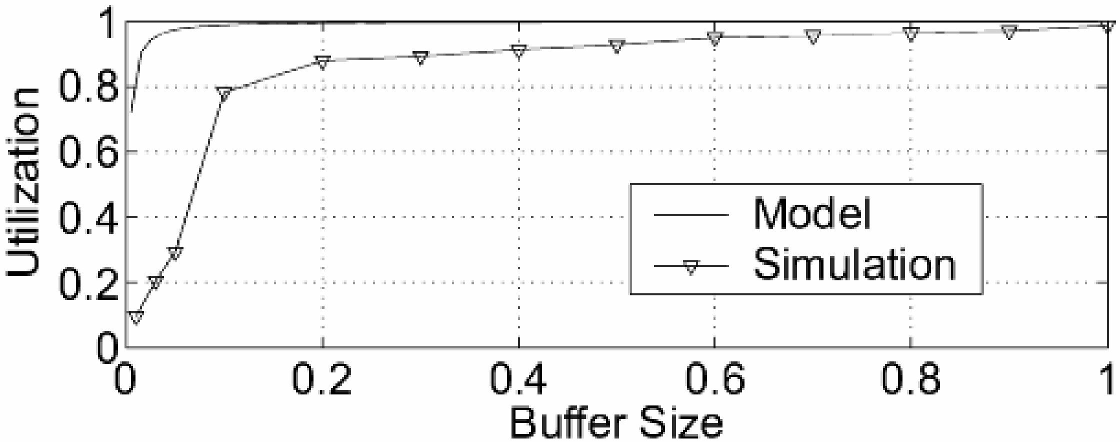


- Worse RTT fairness
- Increased synchronization
- High Utilization

Reduced Buffer Size (DropTail)



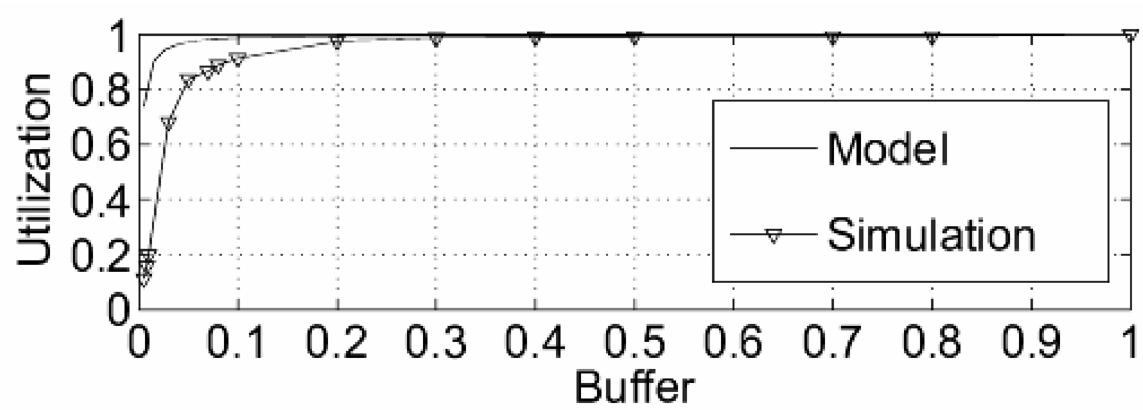
1Gbps bottleneck
10 desynchronized flows



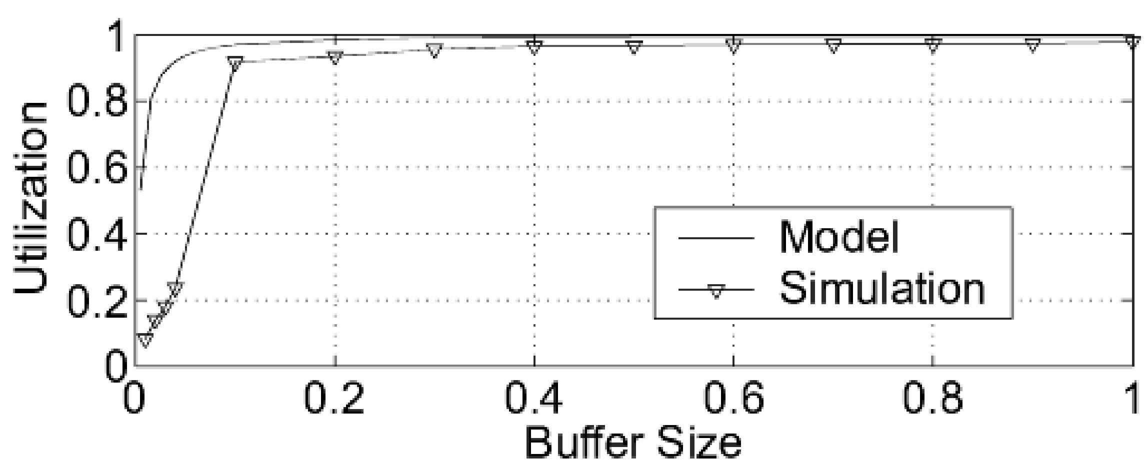
2.5 Gbps bottleneck
10 desynchronized flows

High Utilization even with small buffer size is due to aggregation

Reduced Buffer Size (RED)



1Gbps bottleneck
10 desynchronized flows



2.5 Gbps bottleneck
10 desynchronized flows

High Utilization is still achieved even with conservative behavior of RED.

Summary

- Buffer size ≈ 10% of BDP suffices to maintain high Utilization > 90%
- Under high aggregation the results are insensitive to
  - Queue Management
  - Router Linespeed
- Small buffer size
  - worsens fairness among flows (due to less delay variability)
  - leads to synchronization (unlikely due to high aggregation and path diversity)

Future Work

- Analytically quantify degree of aggregation for our results to hold
- Improve model considering unsynchronized losses
- Experiments with other HighSpeed TCP variants
  – e.g., FAST TCP which reacts to both packet loss and queueing delay
- Model the effect of router buffer size on other High Speed TCP features
  – Stability, Responsiveness

Thank you for your attention