A Fluid-based Simulation Study:

The Effect of Loss Synchronization on Sizing Buffers over 10Gbps High Speed Networks

Suman Kumar, Mohammed Azad, Seung-Jong Park*

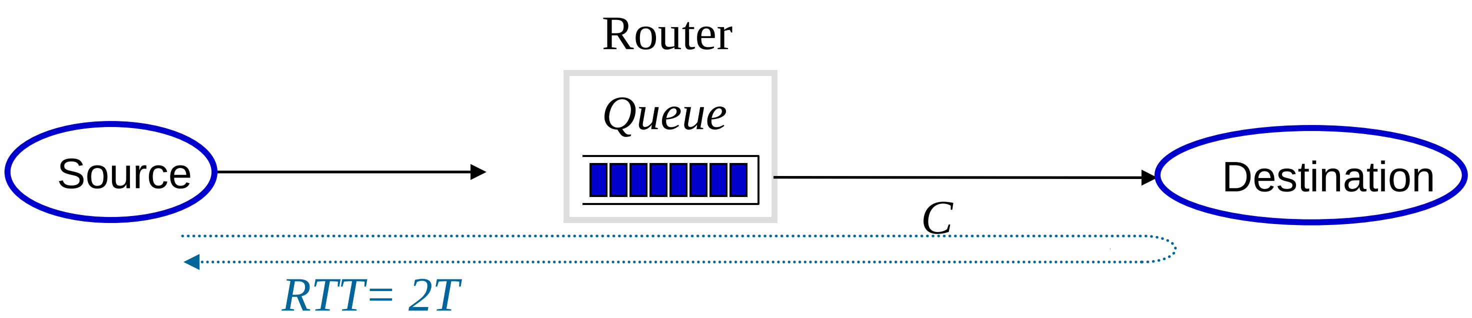
Computer Science Department and
Center for Computation and Technology
Louisiana State University

Outline

- Background
- Problem and Motivation
- Fluid Model for High Speed Networks
- Performance Evaluation on 10Gbps High Speed Networks
- Conclusion and Future Research Direction

Background: Initial Work

- Packet switching networks need a buffer at routers to
  - ✓ Absorb the temporary bursts to avoid packet losses
  - ✓ Keep the link busy during the time of congestion



- Classic rule of thumb for sizing buffers to achieve full link utilization requre
  - ✓ $2T$ is the two-way propagation delay
  - ✓ C is capacity of bottleneck line

$$B = 2T \times C$$

*Villamizar and Song: “High Performance TCP in ANSNET”, CCR, 1994

Background: Recent Works

- Small size buffers are enough to achieve high link utilization [Appenzeller 2004, Raina 2005, etc]

$$B = \frac{2T \times C}{\sqrt{n}}$$

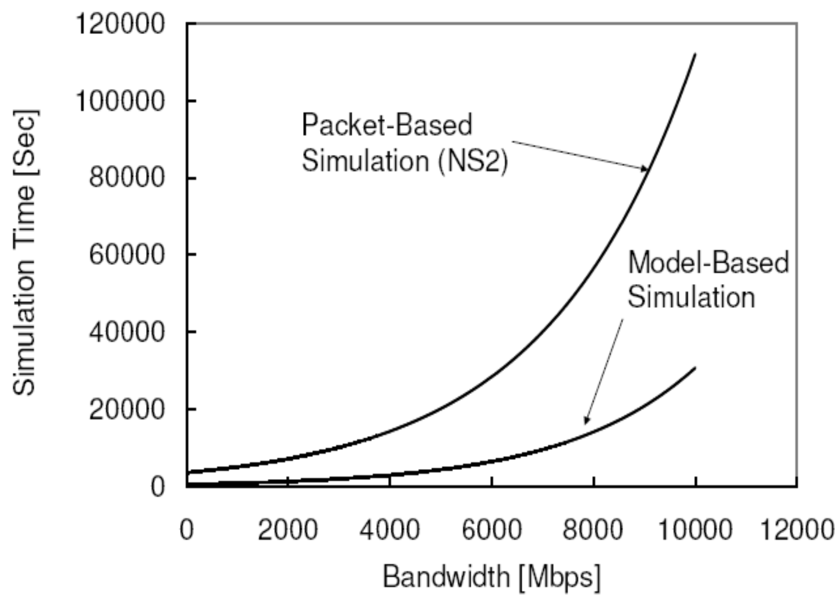
- ✓ Based on assumptions:
  - Larger number of flows than 100 or 1,000 flows
  - Desynchronized and long-lived flows
  - Non-burst traffic flows

Motivation to Revisit

- Different characteristics of high speed networks
  - ✓ A few number of users sharing high speed networks
  - ✓ Most of applications over 10Gbps high speed networks
    - Create a few number of parallel TCP flows
  - ✓ Most of TCP variants for high speed networks
    - Produce high burst traffic
  - ✓ Larger buffer than BDP is not feasible for high speed networks
- Reconsideration on the sizing buffer over 10Gbps high speed networks
  - ✓ Step 1: Find an efficient simulation method for 10Gbps networks
  - ✓ Step 2: Evaluate the performance as a function of buffer size
  - ✓ Step 3: Analyze the impact of synchronization of TCP flows

Comparison of Simulation Methods

- NS2/NS3 Simulation
  - ✓ Only Gigabit results are available
  - ✓ Does not scale to bandwidth of the order of 10Gbps
- Queuing Model [Raina 2005, Barman 2004]
  - ✓ Produces statically stable averaged results
- Fluid Simulation [Liu 2003]
  - ✓ Describes dynamic nature of TCP flows, buffer occupancy, etc.



Scope of this work

- ❑ Network operator's Dilemma
  - ✓ How much buffering to provide
- ❑ Network Users Dilemma
  - ✓ Which high speed TCP variants to use

- ❑ **Goal:**
  - ✓ Understand the impact of loss synchronization on sizing buffers
  - ✓ The effect of these two on the performance of high speed TCPs on 10Gbps high speed networks

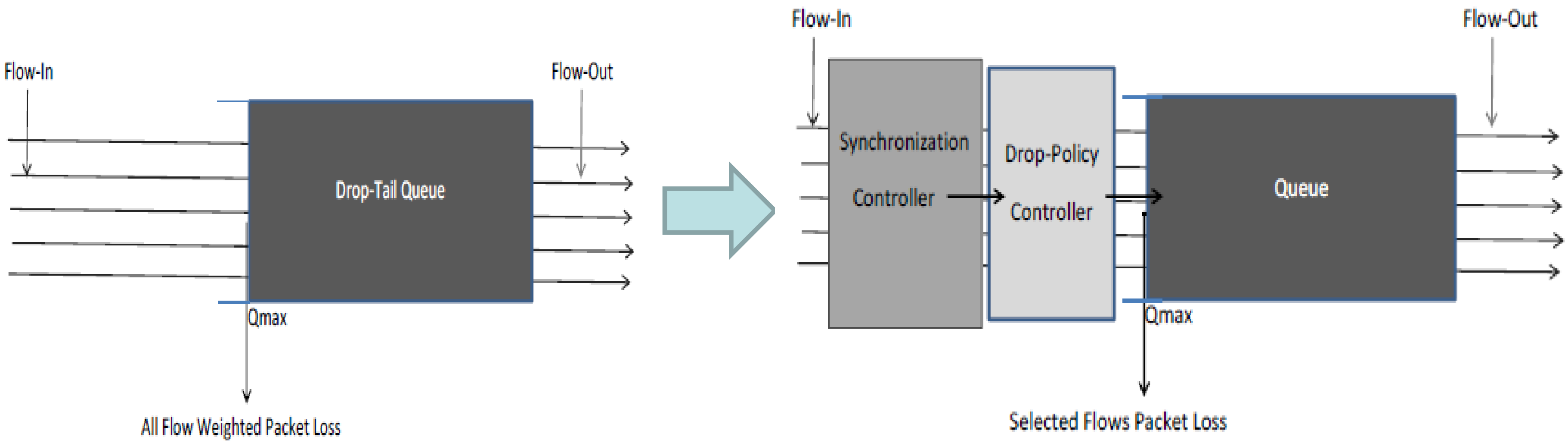
A General Fluid Model

- Traffic is modeled as fluid. [Fluid model -Misra et al]

  - TCP congestion window: $$\frac{dW_i(t)}{dt} = \frac{1(W_i(t) < M_i)}{R_i(t)} - \frac{W_i(t)}{2}\lambda_i(t)$$
  - Queue dynamics $$\frac{q_l(t)}{dt} = -1(q_l(t) > 0)C_l + \sum_{i=1}^{n_l} A_l^i(t)$$
  - Sum of the arrival rates of all flows at bottleneck queue $$ARsum_l = \sum_{i=1}^{n_l} A_l^i(t)$$
  - DT queue generates the loss probability $$p_l(t) = \begin{cases} 0, & q_l(t) < q_l^{max} \\ max(\frac{ARsum_l - C_l}{ARsum_l}, 0), & q_l(t) = q_l^{max} \end{cases}$$
  - This loss probability is proportionally divided among all flows $$\lambda_i(t) = \sum_{l \in F_i} A_l^i(t) p_l(t)$$

Above model do not capture loss synchronization

Loss-Synchronization Model



- Synchronization controller
  - Controls the loss synchronization factor (= m_k) at the time of congestion.
- Drop Policy controller
  - Selects those m_k under some policy

Loss Synchronization Model

- Synchronization Controller
  - ✓ selects m_k flows to drop

- Drop policy controller
  - ✓ At k^{th} congestion, the packet-drop policy controller determines the priority matrix $P^k = [\, D_k^1, D_k^2 \ldots\ldots\ldots, D_k^N]$
    - $D_k^i > D_k^j$ indicates that packets in flow i has higher drop probability than flow j

- All the flows satisfy $$\sum_{i \in Pl_k} \lambda_i(t) = ARsum_l - C$$
  - ✓ every loss is accounted and distributed among the flows

High-Speed Network Simulation Set-up

- Congestion events occur when bottleneck buffer is full.
- Highest rate flows are more prone to record packet losses.
  - ✓ Drop highest rate flows first
- High Speed TCP flow's burstiness induces higher level of synchronization.
  - ✓ Select random m_k at any congestion event k, we define a synchronization ratio parameter X.
    - Ratio of synchronized flows (i.e. experiencing packet losses) and total number of flows is no less than X
    - Selection of X satisfies a least certain level of drop synchronization

- Performance Matrix
  - ✓ %link utilization denoted as $$U = \frac{\sum_s \sum_{i=1}^{n_l} Dep_l^i(t_s)}{C_l \times \sum s} \times 100$$
    - – sample the departure rate (= (dep_l^i) of all the flows *i* at the bottleneck link

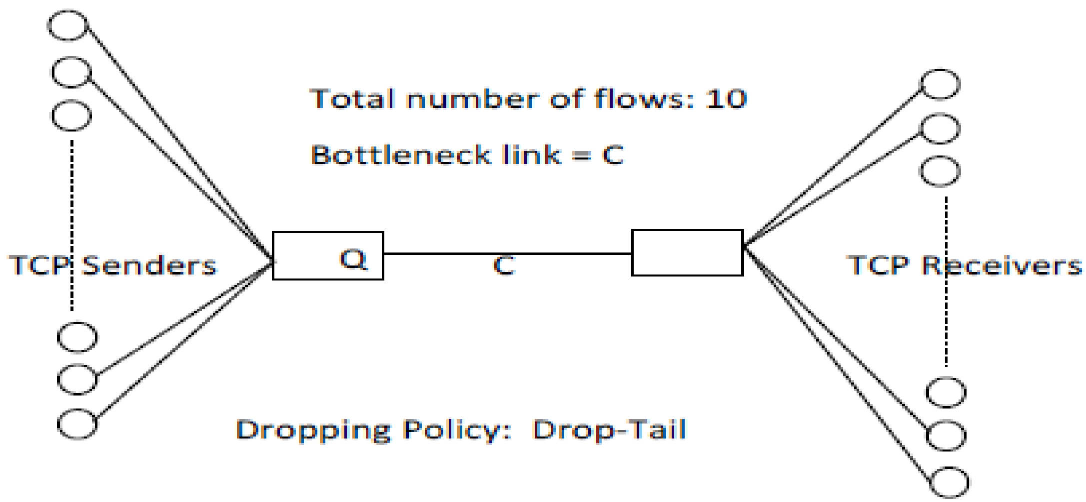
Fluid Model Equations for high speed TCP-Variants

TCP-Variant	a	b
TCP-Reno	1	0.5
STCP	$0.01w$	0.125
HSTCP	$2\frac{w^{0.8}b}{2-b}$	$(0.1-0.5)\frac{log(w)-log(w_{low})}{log(w_{high})-log(w_{low})}$ $+0.5$
CUBIC-TCP	$Min(target_w - w, S_{max}R)$ Where, $target_w$ $= origin\_point + c(\Delta_{th} - K)^3$ $K = (b.prevMax_w/c)^{\frac{1}{3}}$	0.2
H-TCP	$1 + 10(\Delta_i - \Delta_{th})$ $+(\frac{\Delta_i - \Delta_{th}}{2})^2$	$1 - \frac{R_{min}}{R_{max}}$
FAST-TCP	$Min(w, \gamma(2baseR)$ $-avgRTT)\frac{w}{RTT} + \alpha$	0.5

$$\frac{dW_i(t)}{dt} = \frac{a(t)}{R_i(t)} - W_i(t)b(t)\lambda_i(t)$$

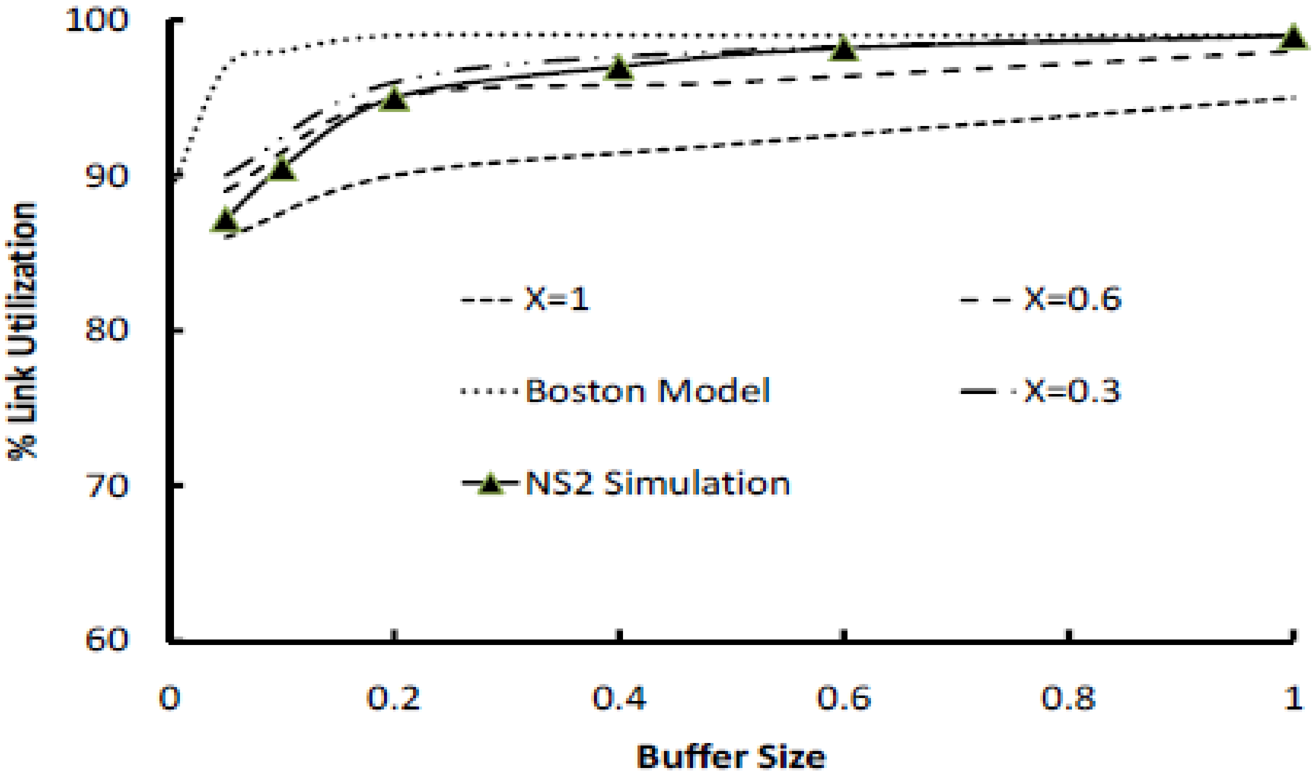
*** Kumar et. al. "A loss-event driven scalable fluid-based simulation method for high-speed networks," Journal of Computer Networks, Elsivier, 2010 Jan**

Simulation Setup



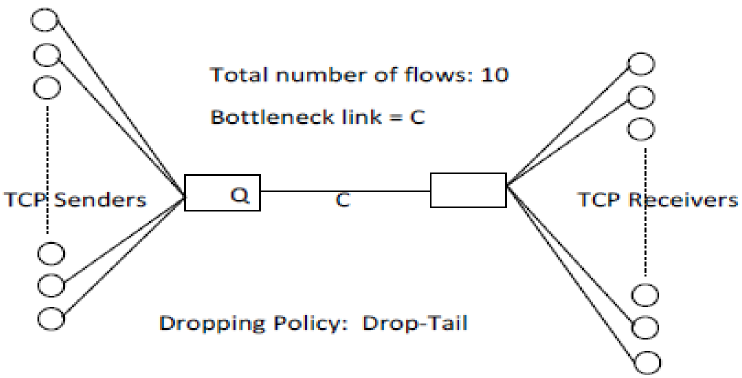
- Unfair drop-tail with the support of loss-synchronization
  - ✓ Two level of Synchronization
  - ✓ Low, *X*=0.3
  - ✓ High, *X*=0.6
- m is drawn from normal distribution and bounded by above values of *X*

Simulation Model Verification



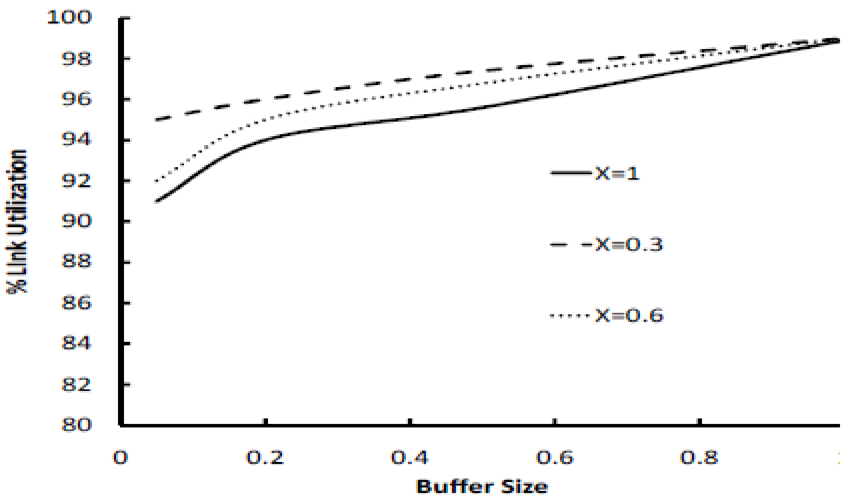
✓ Fluid simulation with synchronization model gives more accurate and realistic results than the Boston model.

Simulation Setup for10Gbps Networks

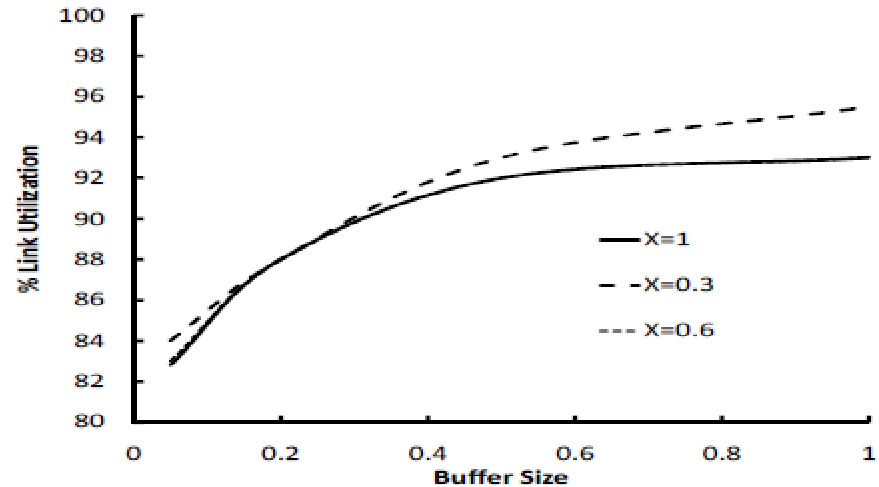


- Network Topolgy = Dumb-bell
- Number of flows = 10
- Bottleneck Link = 10Gbps,
- Link delay = 10ms
- RTTs of 10 flows are ranging from 80ms ~ 260ms
- Maximum buffer size = 141,667 of 1500Byte packets (calculation based on average RTT of 170ms)

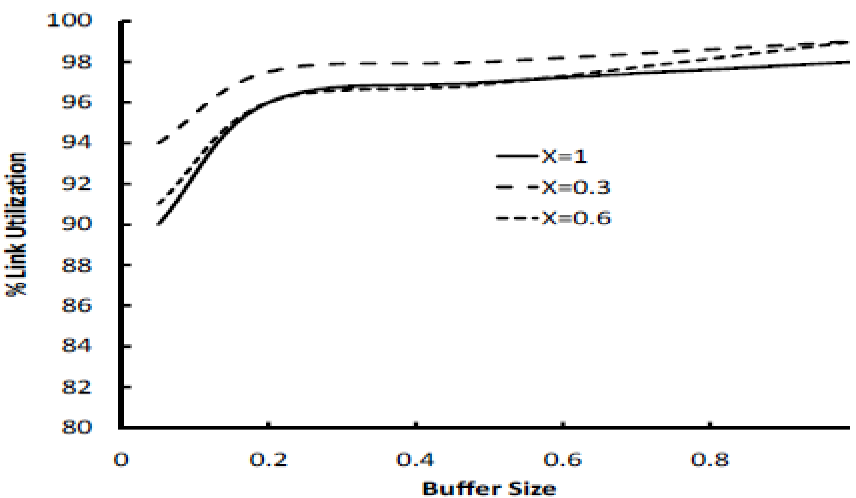
Simulation Results



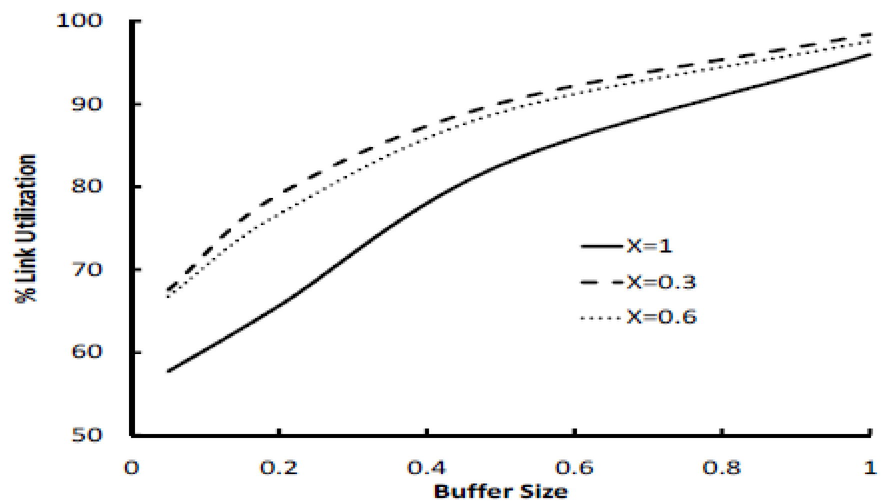
(a) HSTCP



(b) CUBIC



(c) AIMD



(d) HTCP

Observations

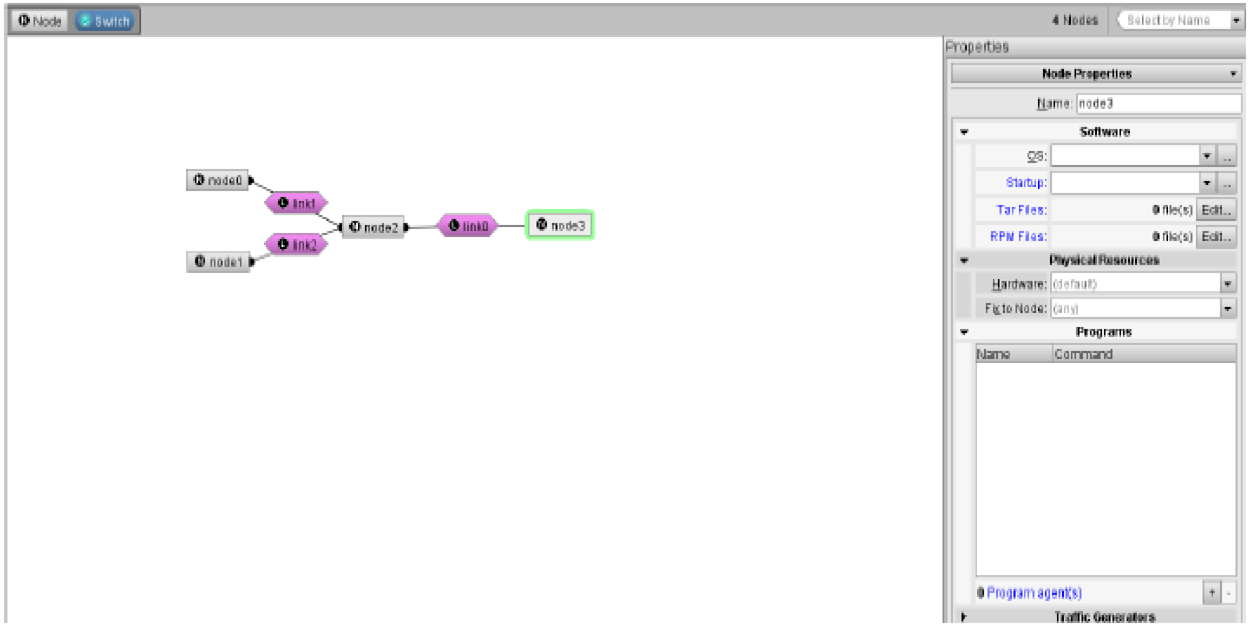
- Measured throughputs of high speed TCP variants were lower than that of TCP Reno especially for high level of synchronization
- For HSTCP, more than 90% link utilization can be achieved with buffer size fraction of 0.05
- Main reason for the poor performance of CUBIC and HTCP as compared to AIMD and HSTCP is attributed to its improved fairness
- Lower synchronization (= Higher desynchronization) further improves the link utilization for HSTCP and AIMD.

Conclusion and Future Work

- A loss synchronization module for fluid model simulation is proposed
- Simulation results for HSTCP, CUBIC and AIMD are presented to show the effect of different buffer sizes on link utilization.
- Loss synchronization module as a black box, where loss synchronization data can be fed from real experiments or one can utilize some theoretical distribution models.
- Future work
  - ✓ Exploration of more accurate models for drop synchronization
  - ✓ Proposing desynchronization methods

Experiment with CRON

- Experimental design with Java based GUI of Emulab
  - Additional features such as tracing, Link Queuing policy, traffic generators, availability of TAR files etc.

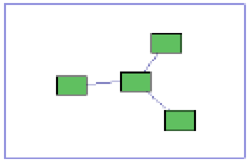


Experiment with CRON contd…

Experiment Options

- View Activity Logfile
- Swap Experiment Out
- Terminate Experiment
- Modify Experiment
- Modify Traffic Shaping
- Modify Settings
- Link Tracing/Monitoring
- Event Viewer
- Update All Nodes
- Reboot All Nodes
- Run LinkTest
- Show History
- Duplicate Experiment

0 Free PCs, 0 reloading	
pcSUN4240	0
SUN4240pc2only	0



Settings | Visualization | NS File | Details

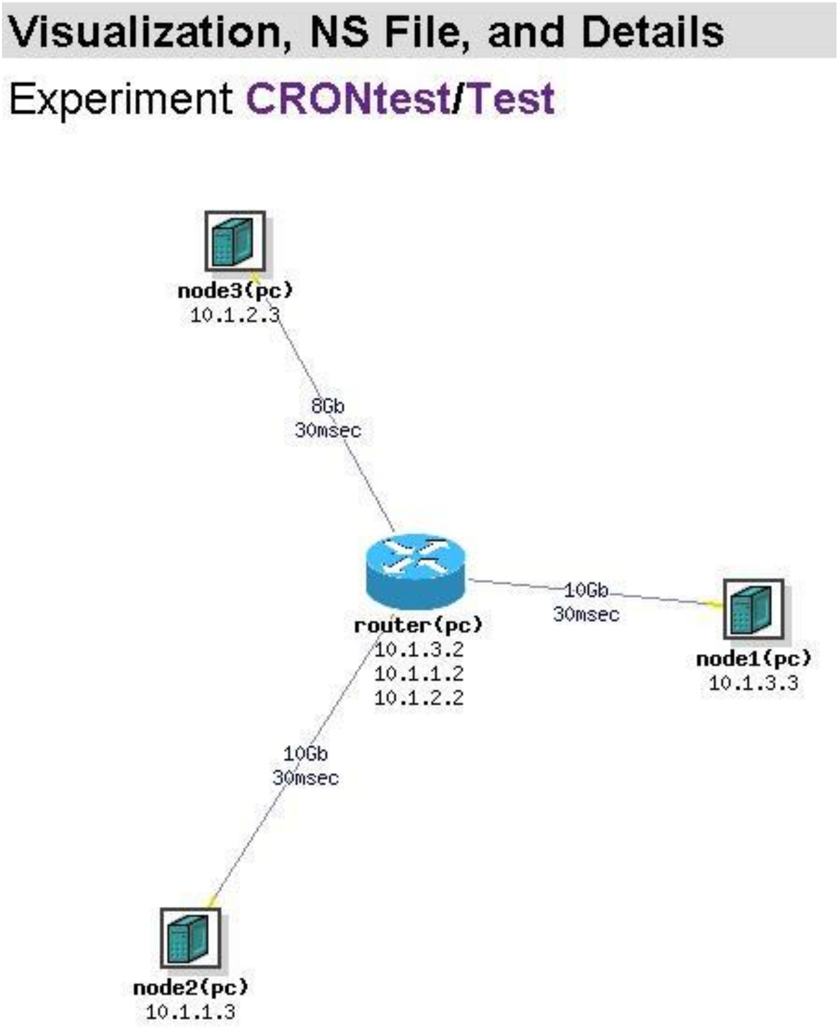
Name:	Ytopology
Description:	Link test on Y topology
Project:	CRONtest
Group:	CRONtest
Experiment Head:	userccui
Created:	2010-10-23 22:24:02
Last Swap/Modify:	2010-10-29 17:13:47 (userccui)
Idle-Swap:	**No** (test)
Max. Duration:	No
Save State:	No
Path:	/proj/CRONtest/exp/Ytopology
Status:	active
Linktest Level:	0
Reserved Nodes:	7 (pc)
Mem Usage Est:	0
CPU Usage Est:	3
Last Activity:	2010-11-04 12:09:41
Idle Time:	0 hours (stale)
Locked Down:	No (Toggle)
Sync Server:	node1
Index:	168

Reserved Nodes

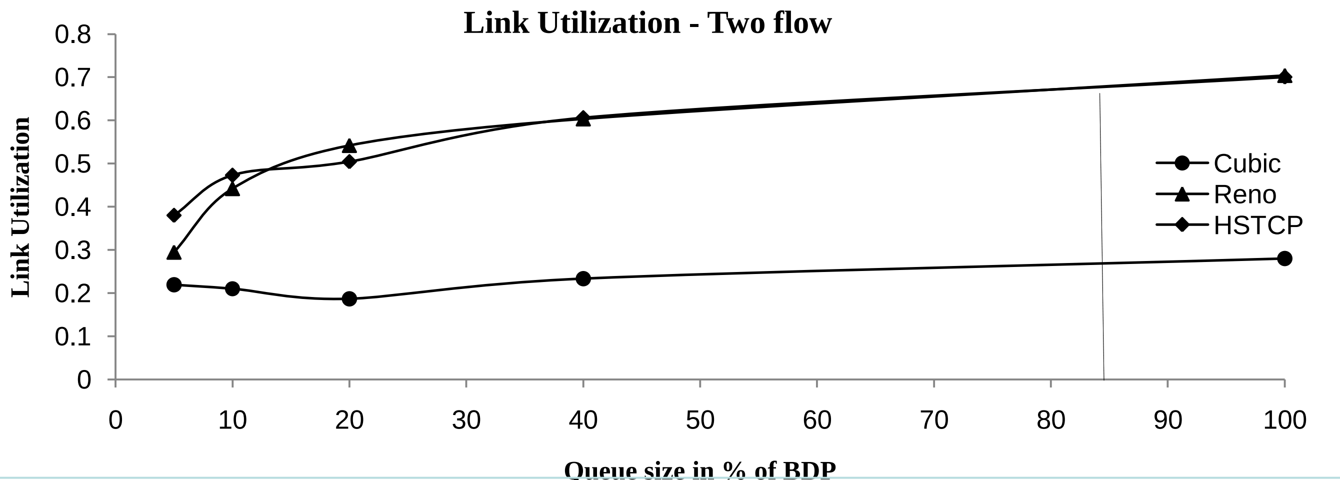
Node ID	Name	Type	Default OSID	Node Status	Hours Idle[1]	Startup Status[2]	SSH	Console	Log
pc1	node1	pcSUN4240	UBUNTU10-64-BETA-10K	possibly down	29.03?	none			
pc3	node2	pcSUN4240	UBUNTU10-64-BETA-10K	possibly down	34.97?	none			
pc4	tbdelay1	pcSUN4240	FBSD81-64-DELAY-BETA	up	0	0			
pc5	tbdelay2	pcSUN4240	FBSD81-64-DELAY-BETA	up	0.08	0			
pc6	node3	pcSUN4240	UBUNTU10-64-BETA-10K	possibly down	16.78?	none			
pc7	router	pcSUN4240	UBUNTU10-64-BETA-10K	up	16.36	none			
pc9	tbdelay0	pcSUN4240	FBSD81-64-DELAY-BETA	up	0	0			

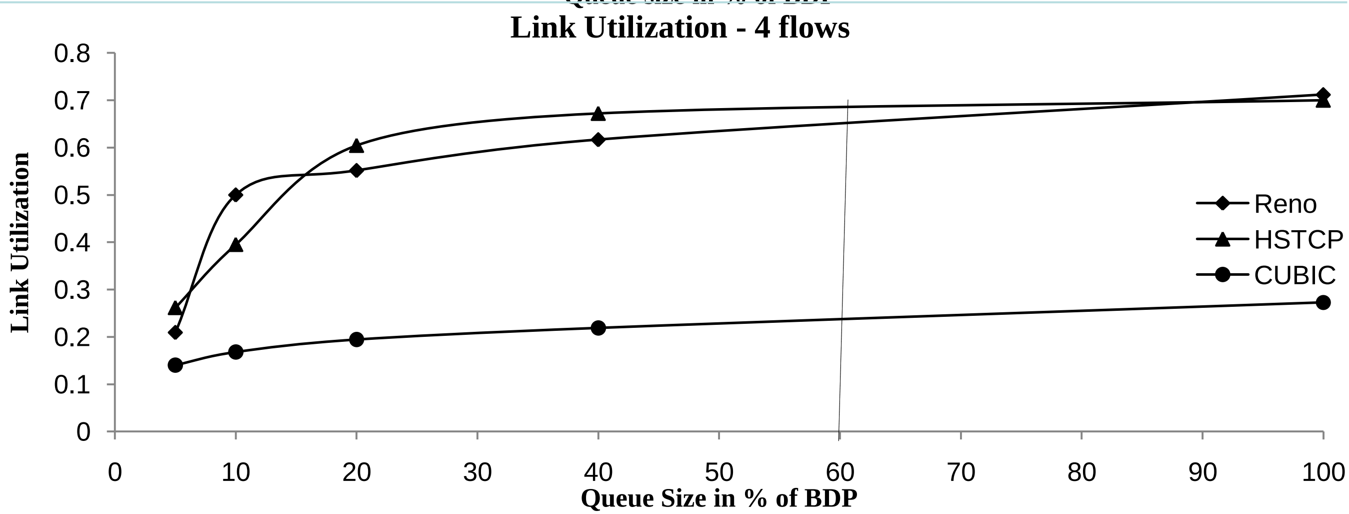
Experiment with CRON contd…

- Y-topology similar to Dumbbell
- Dummynet software emulators were used to emulate large size buffers
- Bottleneck link has 8Gbps bandwidth and 30msec
- CRON testbed webpage
  - ✓ http://cron.cct.lsu.edu



Experimental Results and Analysis





Questions ?