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 requires
  - $2T$ 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
  \[
  \frac{q_i(t)}{dt} = -1(q_i(t) > 0)C_i + \sum_{i=1}^{n_l} A_i^i(t)
  \]

  - Sum of the arrival rates of all flows at bottleneck queue
  \[
  ARsum_1 = \sum_{i=1}^{n_l} A_i^i(t)
  \]

  - DT queue generates the loss probability
  \[
  p_i(t) = \begin{cases} 
  0, & q_i(t) < q_i^{max} \\
  \max(\frac{ARsum_1 - C_i}{ARsum_1}, 0), & q_i(t) = q_i^{max}
  \end{cases}
  \]

  - This loss probability is proportionally divided among all flows
  \[
  \lambda_i(t) = \sum_{l \in F_i} A_i^i(t)p_l(t)
  \]

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 $P^k = [D_{k1}, D_{k2}, ..., D_{kN}]$
    - $D_{ki} > D_{kj}$ indicates that packets in flow $i$ have higher drop probability than flow $j$

- **All the flows satisfy**
  - Every loss is accounted and distributed among the flows
    \[
    \sum_{i \in P_l_k} \lambda_i(t) = ARsum_l - C
    \]
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
  $$U = \frac{\sum_s \sum_{i=1}^{n_l} Dep_i(t_s)}{C_l \times \sum_s} \times 100$$
  - sample the departure rate ($= (dep_i)$ of all the flows $i$ at the bottleneck link
**Fluid Model Equations for high speed TCP-Variants**

<table>
<thead>
<tr>
<th>TCP-Variant</th>
<th>(a)</th>
<th>(b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCP-Reno</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>STCP</td>
<td>(0.01w)</td>
<td>0.125</td>
</tr>
<tr>
<td>HSTCP</td>
<td>(\frac{2^{\frac{a}{w-b}}}{2-b})</td>
<td>((0.1 - 0.5)\frac{\log(w) - \log(w_{low})}{\log(w_{high}) - \log(w_{low})} + 0.5)</td>
</tr>
<tr>
<td>CUBIC-TCP</td>
<td>(\text{Min}(target_w - w, S_{\text{max}}R))</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>Where, (target_w) = (\text{origin-point} + c(\Delta t_K - K)^3)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(K = (b \cdot \text{prevMax}_w / c)^{\frac{1}{3}})</td>
<td></td>
</tr>
<tr>
<td>H-TCP</td>
<td>(1 + 10(\Delta i - \Delta th) + (\Delta i - \Delta th)^2)</td>
<td>(1 - \frac{R_{\text{min}}}{R_{\text{max}}})</td>
</tr>
<tr>
<td>FAST-TCP</td>
<td>(\text{Min}(w, \gamma(2\text{baseR}) - \text{avgRTT} \cdot \frac{w}{\text{RTT}} + \alpha)</td>
<td>0.5</td>
</tr>
</tbody>
</table>

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

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 \( X \)
Fluid simulation with synchronization model gives more accurate and realistic results than the Boston model.
Simulation Setup for 10Gbps Networks

- Network Topology = 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

Reserved Nodes

<table>
<thead>
<tr>
<th>Node ID</th>
<th>Name</th>
<th>Type</th>
<th>Default OSID</th>
<th>Node Status</th>
<th>Hours [1]</th>
<th>Startup Status [2]</th>
<th>SSH</th>
<th>Console</th>
<th>Log</th>
</tr>
</thead>
<tbody>
<tr>
<td>pc1</td>
<td>node1</td>
<td>pcSUN4240</td>
<td>UBUNTU10-64.BETA-10K</td>
<td>possibly down</td>
<td>29.03?</td>
<td>none</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pc3</td>
<td>node2</td>
<td>pcSUN4240</td>
<td>UBUNTU10-64.BETA-10K</td>
<td>possibly down</td>
<td>34.97?</td>
<td>none</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pc4</td>
<td>lbdelay1</td>
<td>pcSUN4240</td>
<td>FBSD81-04-DELAY-BETA</td>
<td>up</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pc5</td>
<td>lbdelay2</td>
<td>pcSUN4240</td>
<td>FBSD81-04-DELAY-BETA</td>
<td>up</td>
<td>0.08</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pc6</td>
<td>node3</td>
<td>pcSUN4240</td>
<td>UBUNTU10-64.BETA-10K</td>
<td>possibly down</td>
<td>16.78?</td>
<td>none</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pc7</td>
<td>router</td>
<td>pcSUN4240</td>
<td>UBUNTU10-64.BETA-10K</td>
<td>up</td>
<td>16.36</td>
<td>none</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pc9</td>
<td>lbdelay0</td>
<td>pcSUN4240</td>
<td>FBSD81-84-DELAY-BETA</td>
<td>up</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

Visualization, NS File, and Details

Experiment CRONtest/Test
Experimental Results and Analysis

Link Utilization - Two flow

Queue size in % of BDP

Link Utilization - 4 flows

Queue size in % of BDP
Questions ?