Increasing TCP initial window

draft-hkchu-tcpm-initcwnd-01.txt

Nandita Dukkipati
Yuchung Cheng
Jerry Chu
Matt Mathis
{nanditad, ycheng, hkchu, mattmathis}@google.com

30 July, 2010
78th IETF, Maastricht
Overview of prior results for IW10

- Our proposal: increase TCP IW to 10 MSS
- IW10 improves average TCP latency by ~10%
- Large scale data-center experiments demonstrate latency improves across network and traffic properties:
  - Varying network BW, RTTs, BDP, HTTP response sizes, mobile networks
  - Small overall increase in retransmission rate (~0.5%), with most from multiple connections
- Prior work:
  - http://ccr.sigcomm.org/online/?q=node/621
New contributions and the questions addressed

- A framework for running experiments with different IWs in the same data-center
- Primary concern from IETF-77: how does IW10 perform on highly multiplexed links such as in Africa and South America?
- What is the impact on latency due to losses in IW?
- Evaluated the impact of different IWs [3, 10, 16] on latency and retransmission rate
  - Reinforced the prior experiment results with IW10
- Testbed experiments for IW study in controlled environment
  - Preliminary results on fairness
Improved methodology for experiments

Previous methodology:
- Change IW for entire data-center every week
  - Less apples-to-apples: changes in server software and user base
  - Takes weeks to collect data

New methodology:
- Serve different IWs based on IP address in one data-center simultaneously for weeks
  - Same IW for connections from the same IP
  - More apples-to-apples: similar load across server software update and user churn
Analysis of IW10 on Africa traffic

Experiment for 1 week in June 2010
# Impact of IW10 on Africa traffic

Web search latency (ms) and retransmission rate %

## All of Africa

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Avg.</th>
<th>50</th>
<th>75</th>
<th>90</th>
<th>99</th>
</tr>
</thead>
<tbody>
<tr>
<td>IW=10</td>
<td>988.4</td>
<td>503</td>
<td>795</td>
<td>1467</td>
<td>5042</td>
</tr>
<tr>
<td>IW=3</td>
<td>1123.9</td>
<td>538</td>
<td>878</td>
<td>1710</td>
<td>5923</td>
</tr>
<tr>
<td>Impr.</td>
<td>135.5</td>
<td>35</td>
<td>83</td>
<td>243</td>
<td>881</td>
</tr>
<tr>
<td>% Impr.</td>
<td>12%</td>
<td>6.5%</td>
<td>9.5%</td>
<td>14.2%</td>
<td>14.9%</td>
</tr>
</tbody>
</table>

## Africa with low QPS

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Avg.</th>
<th>50</th>
<th>75</th>
<th>90</th>
<th>99</th>
</tr>
</thead>
<tbody>
<tr>
<td>IW=10</td>
<td>1870.5</td>
<td>733</td>
<td>1363</td>
<td>3146</td>
<td>11579</td>
</tr>
<tr>
<td>IW=3</td>
<td>2340.7</td>
<td>857</td>
<td>1773</td>
<td>4110</td>
<td>14414</td>
</tr>
<tr>
<td>Impr.</td>
<td>470.2</td>
<td>124</td>
<td>410</td>
<td>964</td>
<td>2835</td>
</tr>
<tr>
<td>% Impr.</td>
<td>20.1%</td>
<td>14.5%</td>
<td>23.1%</td>
<td>23.5%</td>
<td>19.7%</td>
</tr>
</tbody>
</table>

## Retransmission %

<table>
<thead>
<tr>
<th></th>
<th>IW=10</th>
<th></th>
<th>IW=3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrans. %</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IW=10</td>
<td>3.77%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IW=3</td>
<td>3.35%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Increase</td>
<td>0.42</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Why does latency improve in Africa?

- Large network round-trip time
- Larger IW helps faster recovery of packet losses
- Experiments on testbed demonstrate latency improves in spite of increased packet losses
Why does latency improve in Africa?

- Testbed experiment: 20Mbps, RTT 300ms, BDP buffer, offered load 0.95, 50KB response size
- Motivating example: Makerere University, Uganda
Analysis of IW10 on South America traffic

Experiment for 1 week in June 2010
Latency improvement across services in South America

- Latency improves across a variety of services
- Services with multiple connections experience:
  - Least latency benefits
  - Most increase in retransmission rate

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Web</th>
<th>iGoogle</th>
<th>News</th>
<th>Blogger Photos (multiple connections)</th>
<th>Maps (multiple connections)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>18 [6%]</td>
<td>30 [10%]</td>
<td>4 [2.5%]</td>
<td>2 [1.1%]</td>
<td>6 [3.8%]</td>
</tr>
<tr>
<td>50</td>
<td>38 [6.6%]</td>
<td>198 [26%]</td>
<td>45 [9.9%]</td>
<td>98 [12.7%]</td>
<td>12 [3.2%]</td>
</tr>
<tr>
<td>90</td>
<td>154 [11%]</td>
<td>430 [16%]</td>
<td>336 [15%]</td>
<td>251 [4.5%]</td>
<td>37 [2.6%]</td>
</tr>
<tr>
<td>99</td>
<td>561 [12%]</td>
<td>986 [9.7%]</td>
<td>1827 [19%]</td>
<td>691 [2.9%]</td>
<td>134 [2.9%]</td>
</tr>
<tr>
<td>Delta in Retrans %</td>
<td>0.51</td>
<td>0.52</td>
<td>0.35</td>
<td>2.93</td>
<td>1.28</td>
</tr>
</tbody>
</table>

entry: latency improvement (ms) [% improvement]
Impact of latency under packet losses

Latency of traffic with retransmissions > 0 improves with IW10 as compared to IW3

<table>
<thead>
<tr>
<th>Traffic Type</th>
<th>IW3</th>
<th>IW10</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>6.6%</td>
<td>6.8%</td>
</tr>
<tr>
<td>Web Search</td>
<td>6.11%</td>
<td>6.57%</td>
</tr>
</tbody>
</table>
Retransmissions of IW3 vs IW10

IW10 has ~0 increase in #timeouts, but has more
- fast-retransmit
- post-RTO retransmits
Experiments with higher IWs

Does IW > 10 show better latency?

Try IW = \{3, 10, 16\} at
- **DC 1**
  - 20% in US east coast (RTT < 100ms)
  - 80% in south America (RTT > 100ms)
- **DC 2**
  - 97% in Europe (RTT < 100ms)
Comparison of IW = 3, 10, 16 (DC 1)

Small improvement for larger IWs (>10); mostly for mid-size flows
Comparison of $IW = 3, 10, 16$ (DC 2)

Small improvement for larger IWs (>10); mostly for mid-size flows
Testbed topology
All results are preliminary!

- Traffic generator – enhanced netperf dispatched based on poisson arrival
- Offered load - # of conn/sec (λ) with fixed response size, no pipelining
- Tests parameters - bottleneck b/w, RTT, buffer space, response size
- Test metrics - user completion time (UCT), retransmission rate, link utilization
- Measurement & Diagnosis tools
Fairness between IW10 and IW3 flows

50/50 mix of IW3 and IW10 traffic
BDP buffer, load 0.95, 15KB response size
Fairness between IW10 and IW3 flows

Same as previous slide except response size is 50KB

![Graphs showing user completion time and retransmission rate for different initcwnd values (mix 3, mix 10, all 3, all 10).]
Conclusion

● Take away summary
  o IW10 improves latency even in Africa and South America
  o IW10 helps in quicker recovery from packet losses
    ■ A higher retransmission rate does not necessarily increase latency
  o IW16 shows marginal latency improvement over IW10

● Next steps
  o Ongoing work: fairness between IW3 and IW10 in the transition phase
  o For any pending issues with IW10, join us in solving the problems!
Steps to configure IW on Linux

Changing TCP IW on Linux (kernel version >= 2.6.30)

On your server, do
$ ip route show

select the outgoing route then do
$ ip route change default via <gateway> dev eth0 initcwnd <iw>

If the server process explicitly set SNDBUF, then SNDBUF value >= IW*MSS. Otherwise increase
the initial socket buffer if IW*MSS > /proc/sys/net/ipv4/tcp_wmem[1]

$ cat /proc/sys/net/ipv4/tcp_wmem
4096 16384 4194304
$ echo '4096 IW*MSS 4194304' > /proc/sys/net/ipv4/tcp_wmem

Must restart server process to use new tcp_wmem[1]
Acknowledgements

- We acknowledge the following people at Google for their contributions towards the large scale IW experiments:
  - Ethan Solomita
  - Elliott Karpilovsky
  - John Reese
  - Yaogong Wang
  - Roberto Peon
  - Arvind Jain
Why does latency improve in Africa?

- (from testbed experiment results)

Utilization

Fast Retransmits

Timeouts

Spurious Timeouts
More preliminary results from testbed: latency improves across all transaction sizes with BDP buffer & < 90% offered load
But retransmission rates can be quite different with BDP buffer and > 95% offered load
Insufficient buffer can hurt IW10 latency
40% BDP buffer, 75% offered load