Switch Buffers Experiments:
How much buffer do you need to support 10G flows?

Michael Smitasin (mnsmitasin@lbl.gov), Lawrence Berkeley National Laboratory
Brian L Tierney (bltierney@es.net), ESnet

Buffer Experiment #1: Juniper MX80

TCP Test flows, 50ms path

2Gbps UDP background data

Host 1

Host 2

Host 3

Host 4

Juniper EX9204

Juniper MX80

Modify this egress Buffer size
Experiment #1 Setup

- Try various buffer size on Juniper MX80 using ‘scheduler-map’
- Maximum queue buffer = 125MB
- 2Gbps UDP background traffic from host 4 to host 3; 9000 B MTUs
- Added 50ms latency to TCP flows from host 2 using ‘tc’:
  - tc qdisc add dev ethN root netem delay 50ms

- Paced traffic using ‘tc’, eg:
  - tc qdisc add dev ethN handle 1: root htb
  - tc class add dev ethN parent 1: classid 1:1 htb rate 7gbit
  - tc qdisc add dev ethN parent 1:1 handle 10: netem delay 50ms
  - tc filter add dev ethN protocol ip prio 1 u32 match ip dst 10.1.2.0/24 flowid 1:1
Results: Small Buffers Kill Performance!

<table>
<thead>
<tr>
<th>Buffer Size</th>
<th>Packets Dropped</th>
<th>TCP Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>120 MB</td>
<td>0</td>
<td>8Gbps</td>
</tr>
<tr>
<td>60 MB</td>
<td>0</td>
<td>8Gbps</td>
</tr>
<tr>
<td>36 MB</td>
<td>200</td>
<td>2Gbps</td>
</tr>
<tr>
<td>24 MB</td>
<td>205</td>
<td>2Gbps</td>
</tr>
<tr>
<td>12 MB</td>
<td>204</td>
<td>2Gbps</td>
</tr>
<tr>
<td>6 MB</td>
<td>207</td>
<td>2Gbps</td>
</tr>
</tbody>
</table>

30 Second test, 2 TCP streams
Results with Pacing, 12MB Egress Buffer
(note: many low cost switches have 9MB buffers)

2Gbps background traffic

<table>
<thead>
<tr>
<th>Paced Send Rate</th>
<th>Packets Dropped</th>
<th>TCP Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 Gbps</td>
<td>0</td>
<td>8 Gbps</td>
</tr>
<tr>
<td>4 Gbps</td>
<td>0</td>
<td>8 Gbps</td>
</tr>
<tr>
<td>5 Gbps</td>
<td>0</td>
<td>8 Gbps</td>
</tr>
<tr>
<td>6 Gbps</td>
<td>212</td>
<td>1.5 Gbps</td>
</tr>
</tbody>
</table>

3Gbps background traffic

<table>
<thead>
<tr>
<th>Paced Send Rate</th>
<th>Packets Dropped</th>
<th>TCP Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 Gbps</td>
<td>0</td>
<td>8 Gbps</td>
</tr>
<tr>
<td>4 Gbps</td>
<td>0</td>
<td>8 Gbps</td>
</tr>
<tr>
<td>5 Gbps</td>
<td>53</td>
<td>2.5 Gbps</td>
</tr>
<tr>
<td>6 Gbps</td>
<td>203</td>
<td>1.5 Gbps</td>
</tr>
</tbody>
</table>
Buffer Experiment #2: Cisco 6506

TCP Test flows, 50ms path

2Gbps UDP background data
Experiment #2 Setup

- Try various ‘hold queue’ buffer size on Cisco 6506  
  Maximum queue buffer = 125MB
- 2Gbps UDP background traffic from host 3 to host 4, 9000 Byte MTU
- Added 50ms latency to TCP flows from host 1 to host 2 using ‘tc’:
  - `tc qdisc add dev ethN root netem delay 25ms` (on both sender and receiver)

  Paced sender traffic using ‘tc’, eg:
  - `tc qdisc add dev ethN handle 1: root htb`
  - `tc class add dev ethN parent 1: classid 1:1 htb rate 7gbit`
  - `tc qdisc add dev ethN parent 1:1 handle 10: netem delay 25ms`
  - `tc filter add dev ethN protocol ip prio 1 u32 match ip dst 10.1.2.0/24 flowid 1:1`
### Results for 6506

<table>
<thead>
<tr>
<th>Hold Queue Size</th>
<th>Packets Dropped</th>
<th>TCP Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000 packets (18 MB)</td>
<td>700-800</td>
<td>5 Gbps</td>
</tr>
<tr>
<td>4096 packets (36 MB)</td>
<td>250-550</td>
<td>8.5 Gbps</td>
</tr>
</tbody>
</table>

With Pacing; hold-queue = 4096

<table>
<thead>
<tr>
<th>Paced Send Rate</th>
<th>Packets Dropped</th>
<th>TCP Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 Gbps</td>
<td>5-10</td>
<td>8 Gbps</td>
</tr>
<tr>
<td>6 Gbps</td>
<td>200-400</td>
<td>8 Gbps</td>
</tr>
<tr>
<td>7 Gbps</td>
<td>200-400</td>
<td>8 Gbps</td>
</tr>
<tr>
<td>8 Gbps</td>
<td>200-400</td>
<td>1.5 Gbps</td>
</tr>
</tbody>
</table>

Conclusion: 32MB buffer is not enough, but pacing helps
Brocade MLX results

Configuration:
default-max-frame-size 9212
qos queue-type 0 max-queue-size VALUE

Brocade NI-MLX-10Gx8-M
30 second averages of 15 tests per configuration (omitting first 5 seconds of TCP stream)
with 50ms simulated RTT + 2Gbps UDP Background Traffic
Full LBL test results

Comparison of Linecards & Devices
Averages of 15 tests, 30 seconds each with 50ms simulated RTT + 2Gbps UDP Background Traffic
Full LBL test results

Comparison of Linecards & Devices
30 second averages of 15 tests (omitting first 5 secs)
with 50ms simulated RTT + 2Gbps UDP Background Traffic

- Brocade MLX (1M default)
- Arista 7150
- Cisco 6704
- Cisco 6716 (Oversub Mode)
- Brocade MLX (64M)
- Cisco 6716 (Perf Mode)
- Arista 7500
More Information

• Page describing test methodology:

• Page with known device buffer sizes:
  – http://people.ucsc.edu/~warner/buffer.html

• email: mnsmitasin@lbl.gov, BLTierney@es.net