Advantages of TCP pacing using FQ

Brian Tierney, ESnet
bltierney@es.net
http://fasterdata.es.net
TCP option: Fair Queuing Scheduler (FQ)

Available in Linux kernel 3.11 (released late 2013) or higher
   – available in Fedora 20, Debian 8, and Ubuntu 13.10
   – Backported to 3.10.0-327 kernel in v7.2 CentOS/RHEL (Dec 2015)

To enable Fair Queuing (which is off by default), do:
   – tc qdisc add dev $ETH root fq

Or add this to /etc/sysctl.conf:
   net.core.default_qdisc = fq

To both pace and shape the traffic:
   – tc qdisc add dev $ETH root fq maxrate Ngbit
     • Can reliably pace up to a maxrate of 32Gbps on a fast processors

Can also do application pacing using a ‘setsockopt(SO_MAX_PACING_RATE)’ system call
   – iperf3 supports this via the “—bandwidth’ flag
Advantages of Pacing

• The following plots show a clear benefit from pacing TCP
  – Proper pacing of flows can completely eliminate TCP ‘sawtooth’
• The advantage is even greater with parallel flows
• We recommend all Data Transfer Nodes (DTNs) use FQ-based pacing
• Pacing also helps with eliminate packet loss due to under-buffered network hardware, and under-powered security devices.
New York to Texas: With Pacing

TCP performance: BNL to Pantex; CentOS 6.5 vs CentOS 7.2
10G Host to 1G Host, rtt = 88ms

Bandwidth (Gbits/second)

CentOS6
CentOS7
CentOS7 with 800M pacing

0  0.2  0.4  0.6  0.8  1.0  1.2  1.4  1.6
0  10  20  30  40  50  60  70

Time (seconds)
More examples of pacing helping

TCP performance: CentOS6 vs CentOS7
LBL-to-lut2-net2-0.1u.edu

TCP performance: CentOS6 vs CentOS7
LBL-to-sdm00-rcc.uchicago.edu
Parallel Stream Test 1

Left side:
- sum of 4 streams

Right side:
- tput of each stream

Streams appear to be much better balanced with FQ, pacing to 2.4 performed best.
Parallel Stream Test 2

Left side: sum of 4 streams

Right side: throughput of each stream

Streams appear to be much better balanced with FQ
100G Host, Parallel Streams: no pacing vs 20G pacing

We also see consistent loss on the LAN with 4 streams, no pacing
Packet loss due to small buffers in Dell Z9100 switch?
100G Host to 10G Host

TCP performance: 100G to 10G, default settings, rt = 92ms

TCP performance: 100G to 10G, maxrate = 2.5G, rt = 92ms
Fast Host to Slow host

Throttled the receive host using ‘cpupower’ command:

/bin/cpupower -c all frequency-set -f 1.2GHz
FQ Packets are much more evenly spaced
tcptrace/xplot output: FQ on left, Standard TCP on right
Run your own tests

• Find a remote perfSONAR host on a path of interest
  – Most of the 2000+ worldwide perfSONAR hosts will accept tests
    • See: http://stats.es.net/ServicesDirectory/

• Run some tests
  – bwctl -c hostname -t60 --parsable > results.json

• Convert JSON to gnuplot format:
  – https://github.com/esnet/iperf/tree/master/contrib