ESnet

ENERGY SCIENCES NETWORK

Packet Pacing for TCP

Eli Dart and Brian Tierney

Energy Sciences Network (ESnet)

Lawrence Berkeley National Laboratory

RNTWG Meeting

Virtual

25 May, 2023

U.S. DEPARTMENT OF ENERGY

Office of Science

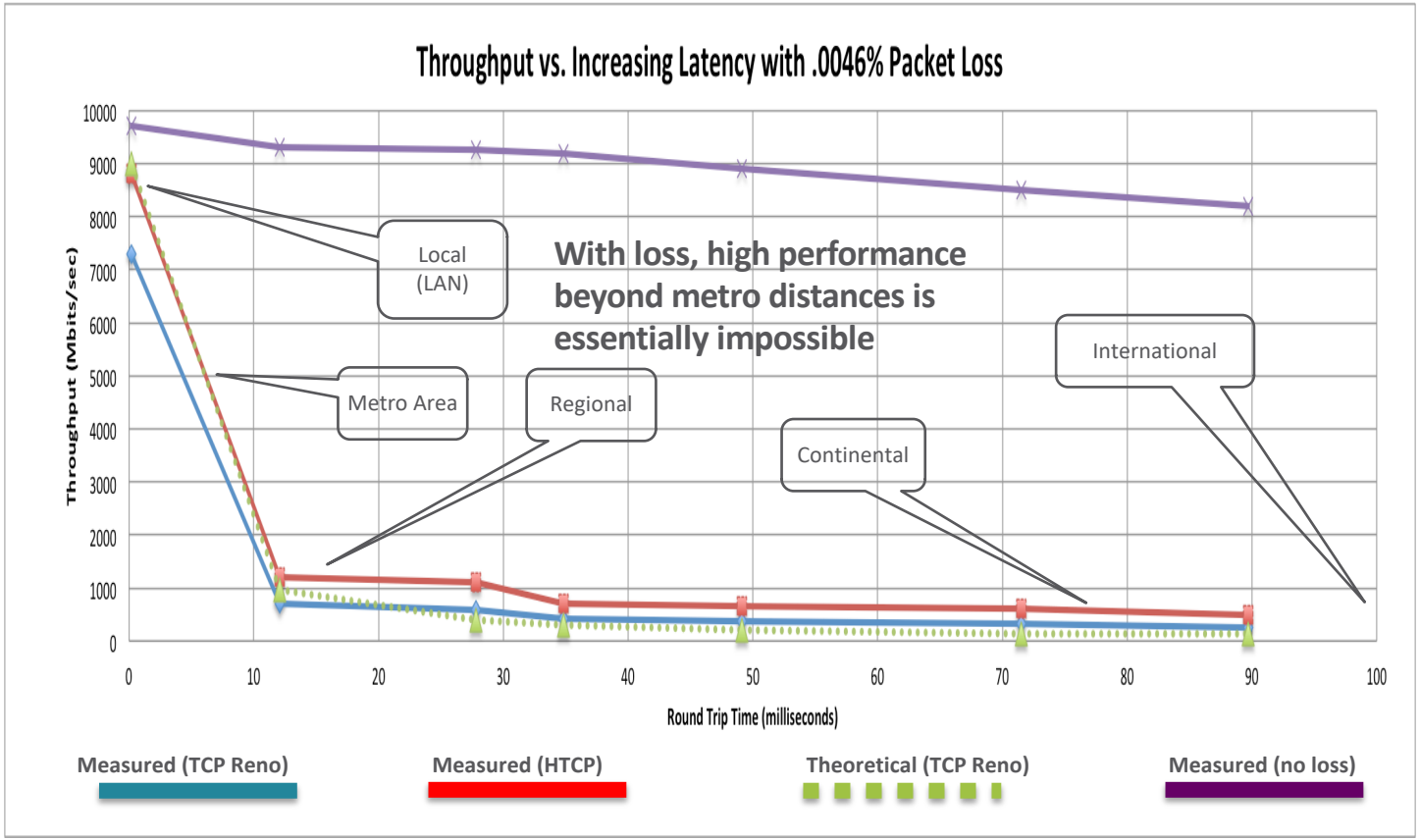
BERKELEY LAB

Framing and Context

- TCP has been and continues to be the workhorse protocol used by data transfer applications
  - Internet's Reliable byte-stream delivery protocol (in contrast to UDP)
  - Underlying mechanism used by HTTP, Globus, FTP, etc.
- TCP performance is badly impacted by even minor packet loss
- TCP's bursty behavior contributes to packet loss
  - This is why we deploy deep buffers
  - Deep buffers are expensive, and will go away in the future
- Is there something we can do about all this?

ESnet

A small amount of packet loss makes a huge difference in TCP performance



ESnet

TCP is “bursty” – what does that mean?

- TCP sends data when two conditions are true
  - TCP has data to send, e.g. the application wrote to the socket
  - The receiver has advertised available window space
  - TCP sends data until one of these conditions is not true
- Most host interfaces can send at wire speed
  - This means that if TCP has data to send, and it hands the data off to the host NIC, the NIC will send packets at wire speed until done
  - 10G NICs send data at 10G, 100G NICs send data at 100G
- On average the rate may be lower, but the instantaneous rate is wire speed
  - A host that runs at 50% of wire speed on average might actually send at wire speed 50% of the time and sit idle 50% of the time

ESnet

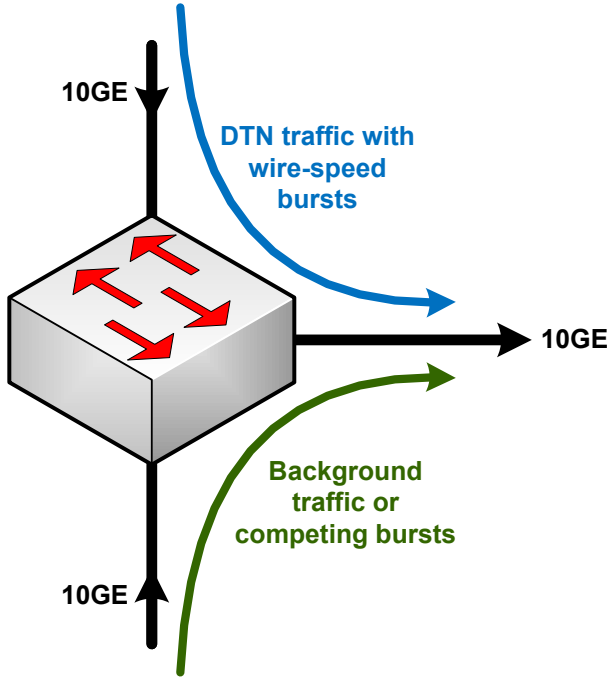
Consequences of TCP Burstiness

- A wire-speed burst has no spare room in it (by definition)
  - If a device can't process the burst at wire speed, it must buffer or drop
  - If the device can handle the burst, then it passes the burst on to the next device which must also handle it, buffer, or drop
- This goes bad in several common ways
  - Shallow switch buffers
  - Slower receivers
  - Speed mismatch on the path

ESnet

Shallow Switch Buffers

- The figure shows a switch with more egress traffic than it can instantaneously support
- On average, the traffic from the two ingress interfaces will fit into the egress interface
- In a given instant, the load may be greater than the speed of the egress interface
  - The switch must buffer or drop
  - If the switch cannot buffer the extra, the result is packet loss → poor performance



ESnet

Slow Receive Host

- If the host cannot receive packets at wire speed, it will drop them unless flow control is enabled
  - Ethernet flow control (“pause frames”) allows the host to ask its first-hop switch to buffer on its behalf
  - This immediately returns us to the shallow buffer switch problem
    - If the switch can’t buffer the rest of the burst, it drops packets
    - Deep buffer switches are better able to support receive hosts
- Some ways to help
  - NIC tuning (e.g. LRO)
  - High CPU clock rate on the host
  - Flow control + deep switch buffers
  - But: we must engineer for this, and it can be costly

ESnet

Speed Mismatch

- This can be thought of as a variant of the fan-in problem
  - Instead of background traffic reducing capacity, slower link speed reduces capacity
  - The other semantics are very similar – buffer or drop
- This can happen in many devices
  - Routers, switches
  - Hosts (slow receiver looks a lot like a speed mismatch)

ESnet

Common Theme: Bursts Cause Problems

- The fundamental issue is that, because of the way TCP behaves, we have to engineer for the instantaneous rate instead of the average rate
- All of this is just so we can provide TCP with a loss-free IP layer to maintain performance
- If we don't engineer for peaks, TCP will underperform the average it could have achieved – because of packet loss caused by bursts
- Is there a better way?

ESnet

Smooth Out The Bursts: Pacing

- Goal of pacing is to limit the burst rate of a TCP flow
- Reduce the impact of bursts on buffers, receivers, etc.
- It can sound backwards, but it works
  - Reduce the available bandwidth to achieve performance
  - This works because the performance impact from packet loss is so large

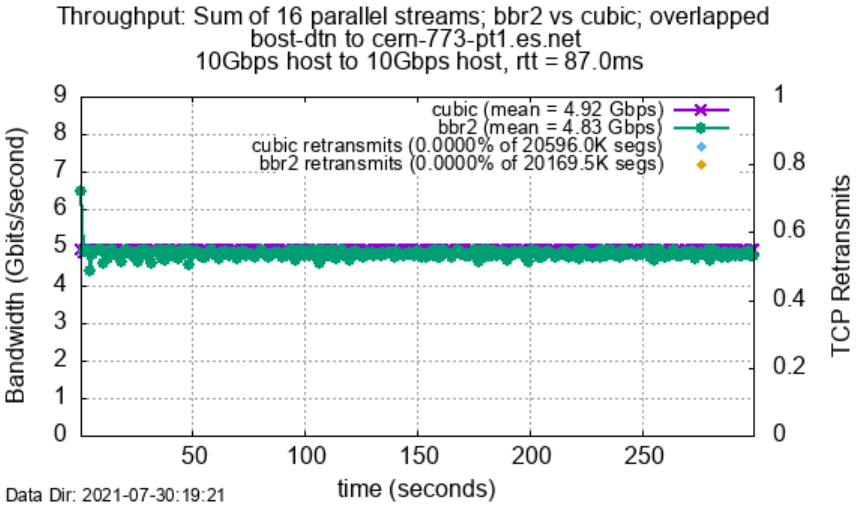
ESnet

BBR TCP has built-in pacing (slide from Matt Mathis presentation, March 2020)

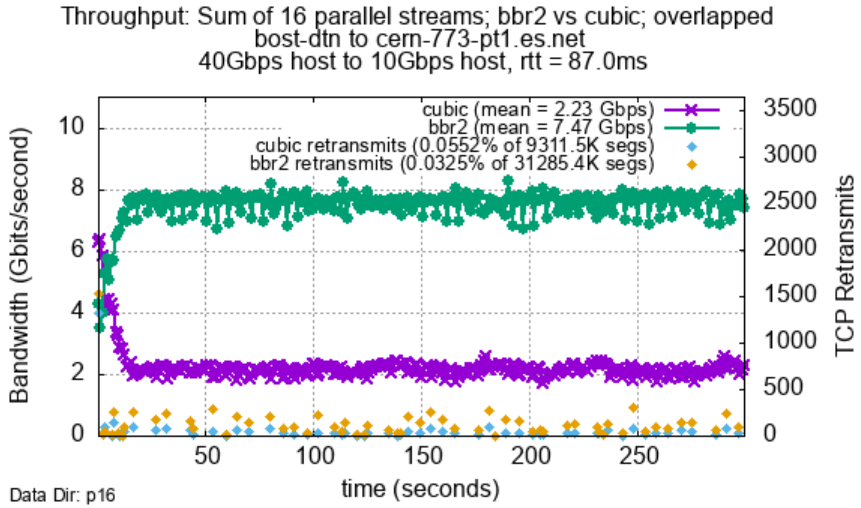
- BBR: new first principles for Congestion Control
  - BBR builds an explicit model of the network
    - Estimate max_BW and min_RTT
- The BBR core algorithm:
  - By default pace at a previously measured Max_BW
    - **Transmit based on a clock**, not ACKs
  - Vary the pacing rate to measure model parameters
    - increase to observe new max rates
    - decrease to observe the min RTT
    - gather other signals such as ECN (bbr2)
- BBR's "personality" is determined by the heuristics used to vary the rates and perform the measurements
  - These heuristics are completely unspecified by the core algorithm
  - Relatively easy to extend or adapt
  - Many different heuristics algorithms can work together

ESnet

Pacing Used In BBRv2 Testing



10G sender (620 Mbps pacing/flow, 9.9G total)

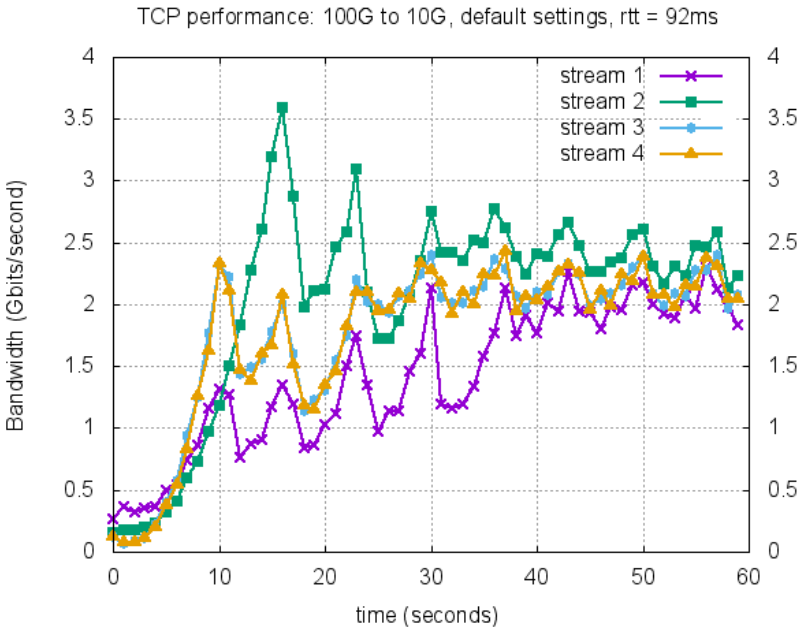


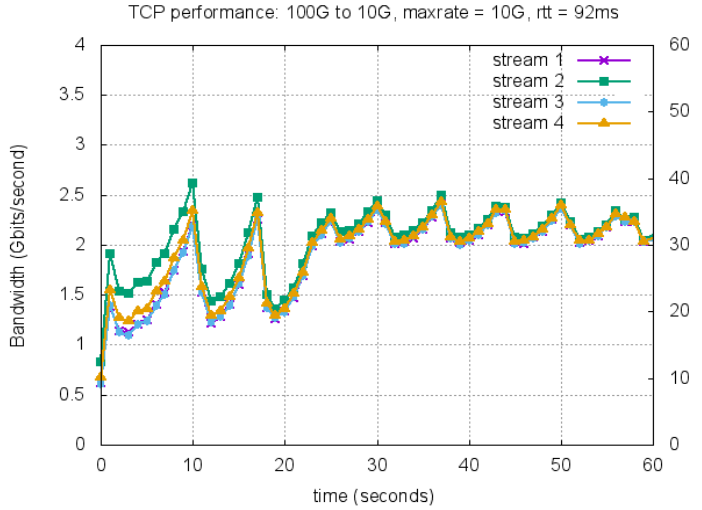
40G sender (2.4 Gbps pacing/flow, 38.4G total)

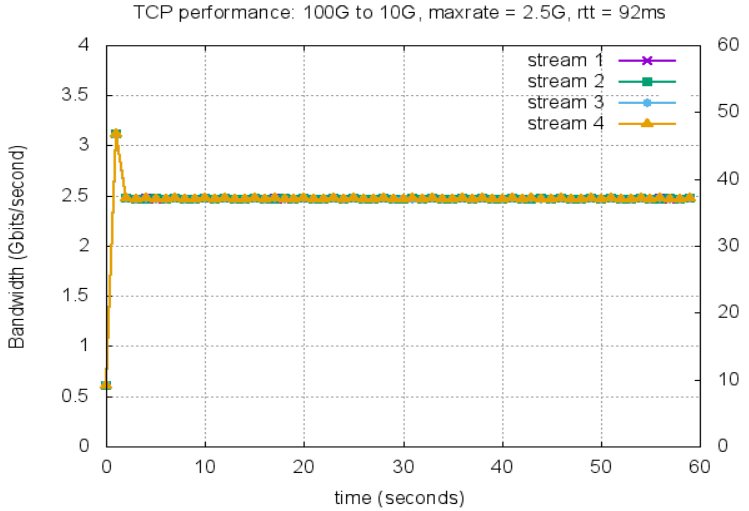
- Brian Tierney et al. doi: 10.1109/INDIS54524.2021.00008.
- No speed mismatch = No packet loss = CUBIC and BBRv2 are equivalent
- But BBRv2 does much better when sender is faster than receiver

ESnet

100G Host to 10G Host







Ideas to Explore

- Can we come up with a simple pacing configuration for WLCG DTNs that improves performance?
- What is the mix of host and network interface speeds in WLCG, and how might that affect a global pacing configuration?
- What per-flow data rate do we need in WLCG? Is it different for different workflows? What does this mean for pacing config?
- BBRv2 will probably make manual pacing configurations obsolete, but BBRv2 is years away (Google has not yet merged it upstream, so it hasn't even begun the path to production distro kernels)

ESnet

ESnet
ENERGY SCIENCES NETWORK

Thanks!

Eli Dart (dart@es.net)
Energy Sciences Network (ESnet)
Lawrence Berkeley National Laboratory

http://fasterdata.es.net/
http://my.es.net/
http://www.es.net/