CC* Cyberinfrastructure Topics
Introduction to Science DMZ

Jorge Crichigno
University of South Carolina
http://ce.sc.edu/cyberinfra

Minority Serving - Cyberinfrastructure Consortium (MS-CC)
University of South Carolina (USC)

Claflin University
Orangeburg, SC
March 22nd, 2023

Workshop on Networking Topics

- Webpage with PowerPoint presentations:

  http://ce.sc.edu/cyberinfra/workshop_2023_claflin.html

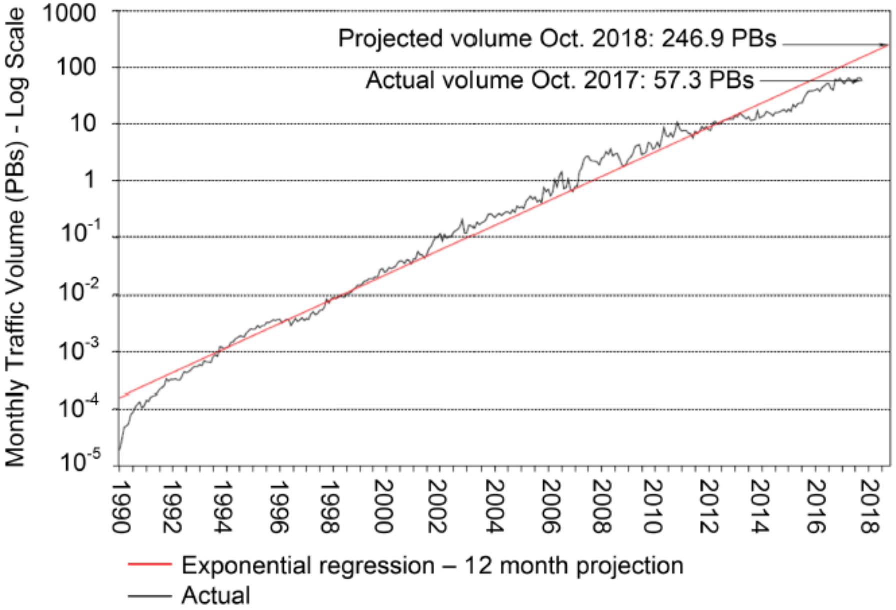
- Hands-on sessions: to access labs for the hands-on sessions, use the following link:

  https://netlab.cec.sc.edu/

- Credentials are provided on site

Motivation for a High-Speed Science Architecture

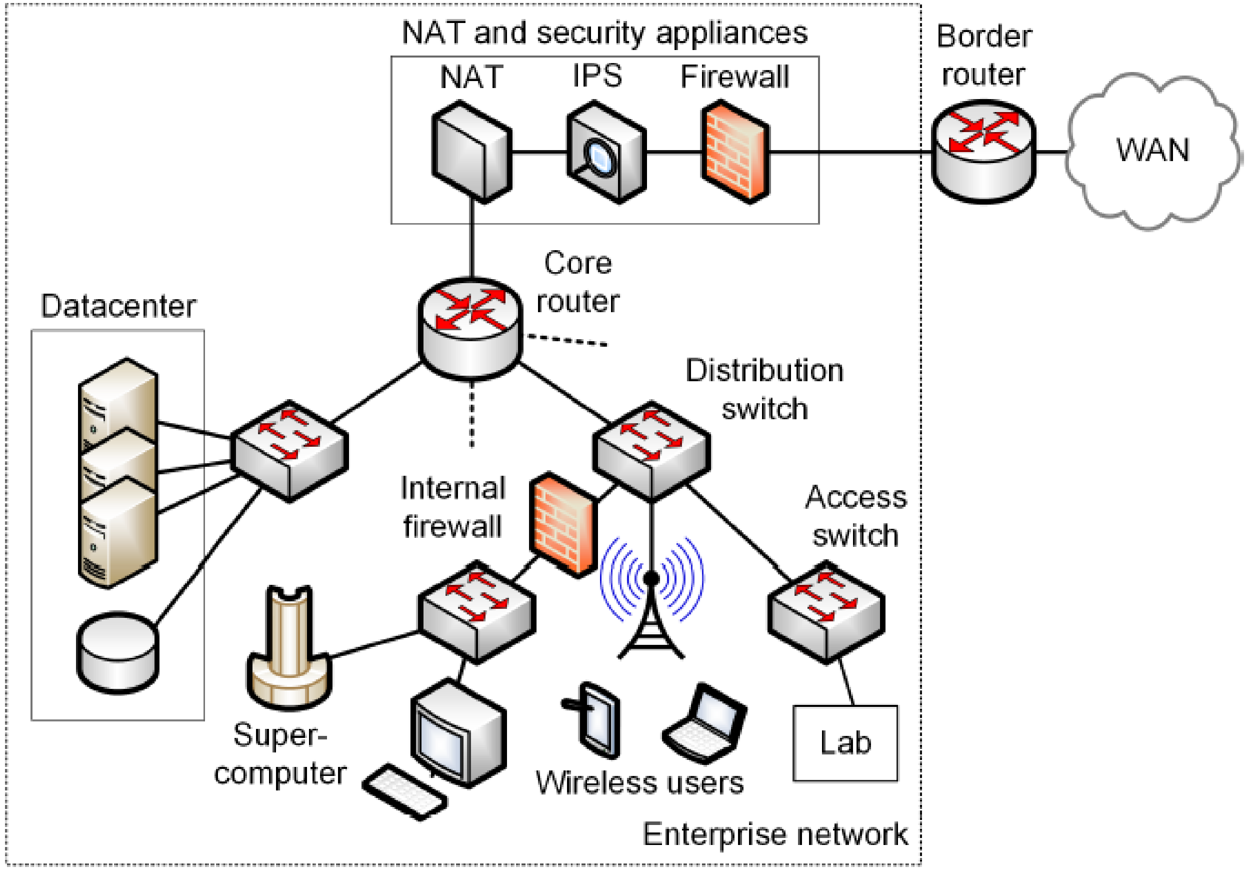
- Science and engineering applications are generating data at an unprecedented rate
- Instruments produce hundreds of terabytes in short time periods ("big science data")
- Data must be typically transferred across high-bandwidth high-latency Wide Area Networks (WANs)



The Energy Science Network (ESnet) is the backbone connecting U.S. national laboratories and research centers

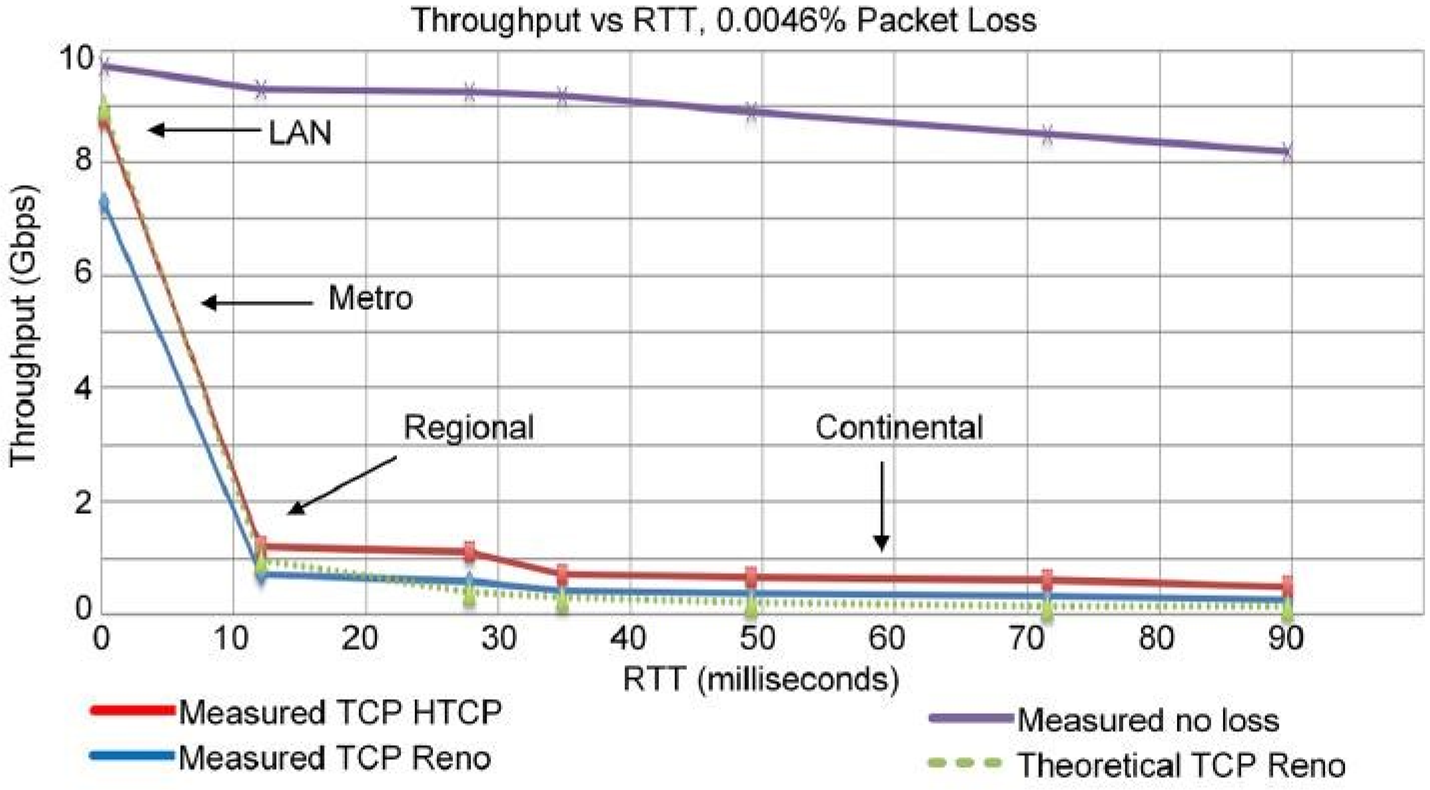
Enterprise Network Limitations

- Security appliances (IPS, firewalls, etc.) are CPU-intensive
- Inability of small-buffer routers/switches to absorb traffic bursts
- End devices incapable of sending/receiving data at high rates
- Lack of data transfer applications to exploit available bandwidth
- Many of the issues above relate to TCP



Enterprise Network Limitations

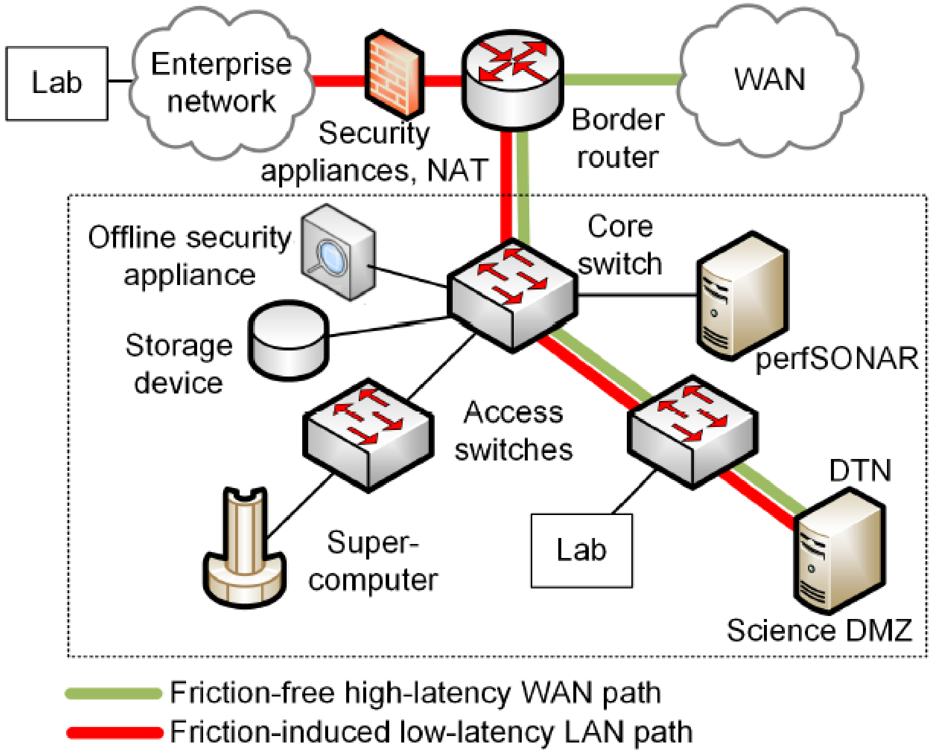
- Effect of packet loss and latency on TCP throughput



E. Dart, L. Rotman, B. Tierney, M. Hester, J. Zurawski, “The science dmz: a network design pattern for data-intensive science,” *International Conference on High Performance Computing, Networking, Storage and Analysis*, Nov. 2013.

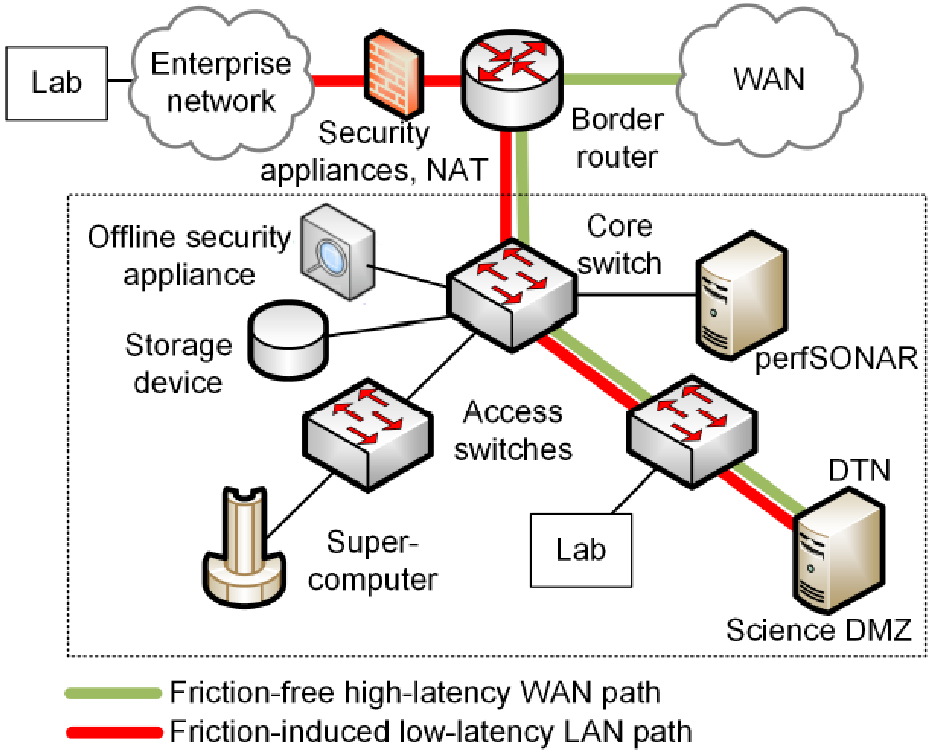
Science DMZ

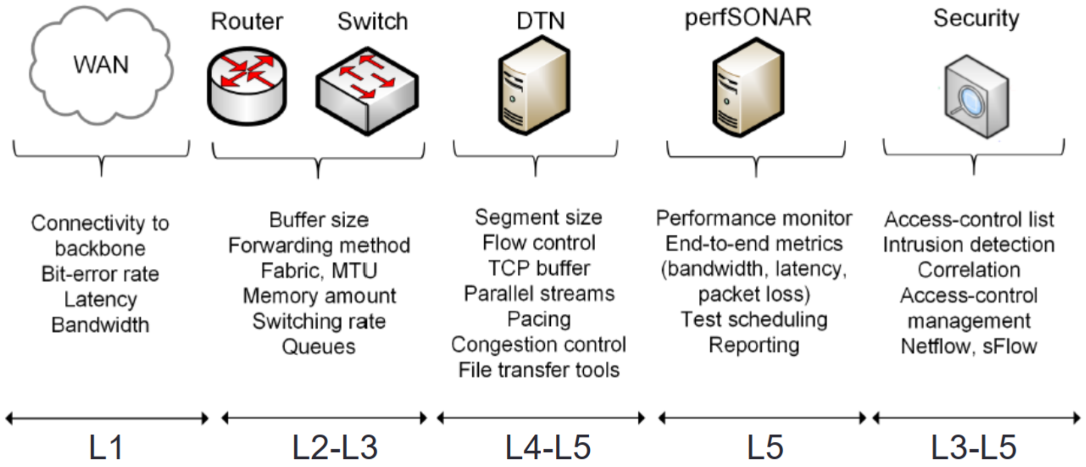
- The Science DMZ is a network designed for big science data
- Main elements
  - High throughput, friction free WAN paths
  - Data Transfer Nodes (DTNs)
  - End-to-end monitoring = perfSONAR
  - Security tailored for high speeds



Science DMZ

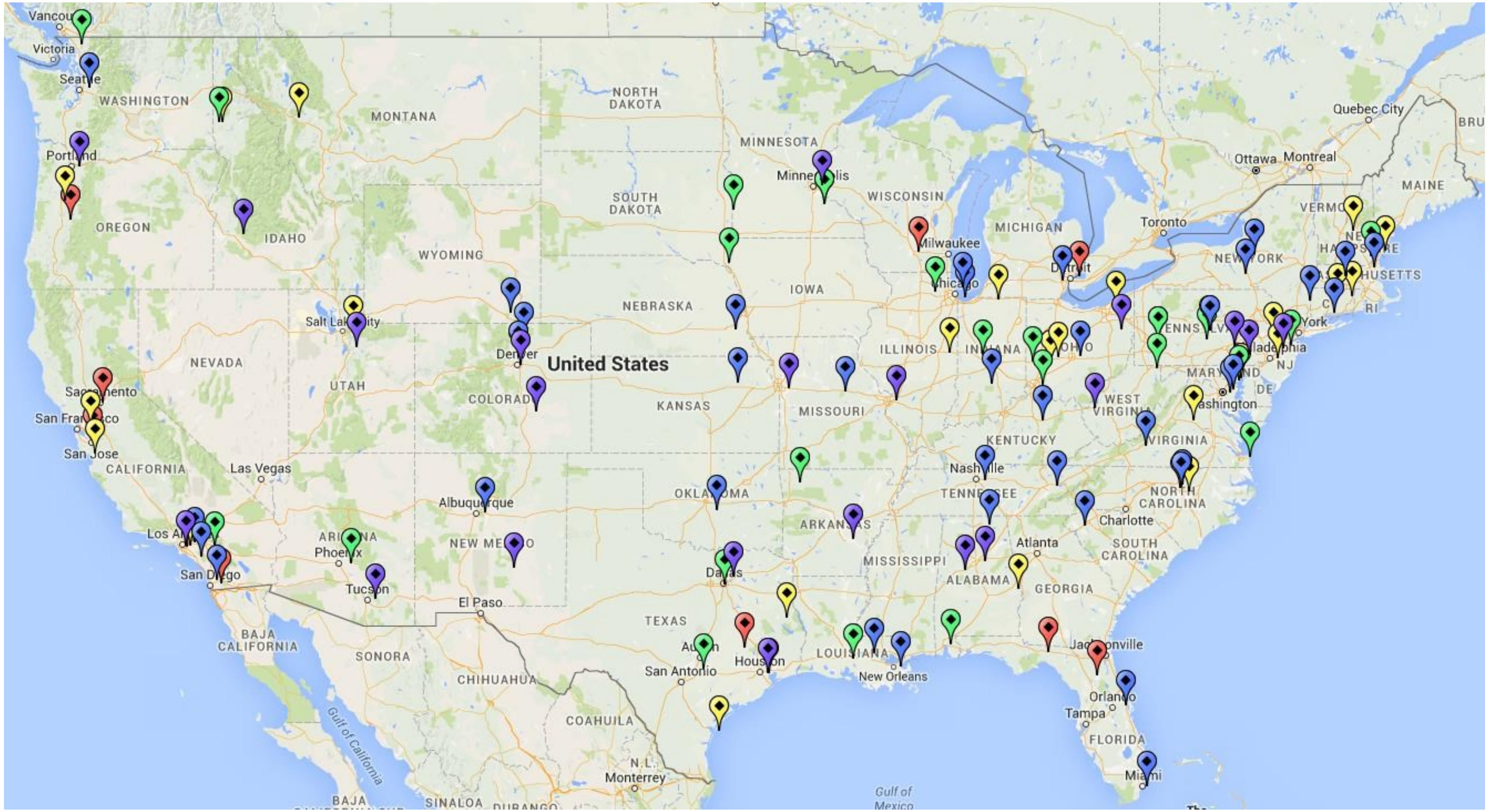
- The Science DMZ is a network designed for big science data
- Main elements
  - ➢ High throughput, friction free WAN paths
  - ➢ Data Transfer Nodes (DTNs)
  - ➢ End-to-end monitoring = perfSONAR
  - ➢ Security tailored for high speeds





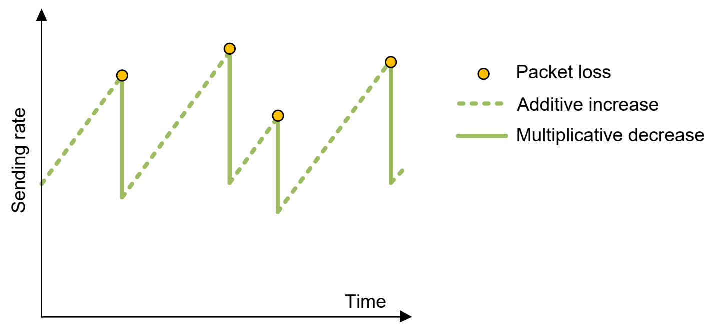
Science DMZ

- Science DMZ deployments, U.S.



TCP Traditional Congestion Control

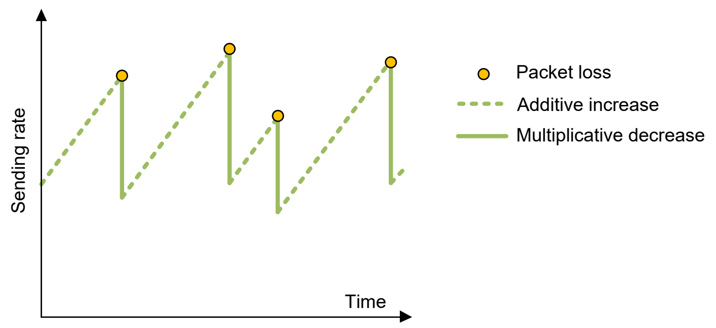
- The principles of window-based CC were described in the 1980s[1]
- Traditional CC algorithms follow the additive-increase multiplicative-decrease (AIMD) form of congestion control

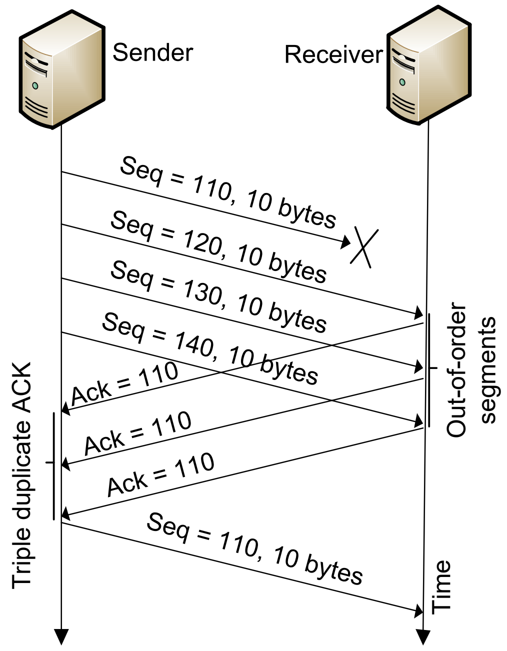


1. V. Jacobson, M. Karels, Congestion avoidance and control, ACM SIGCOMM Computer Communication Review 18 (4) (1988).

TCP Traditional Congestion Control

- The principles of window-based CC were described in the 1980s[1]
- Traditional CC algorithms follow the additive-increase multiplicative-decrease (AIMD) form of congestion control

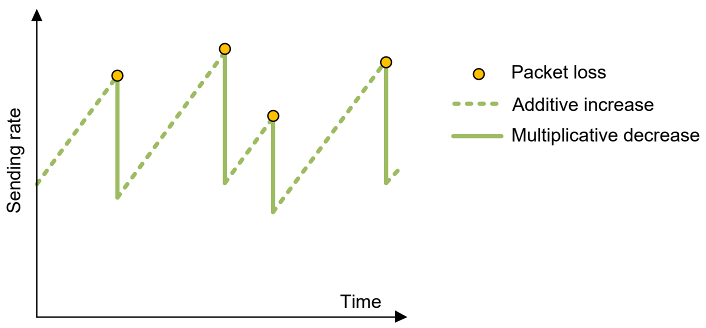


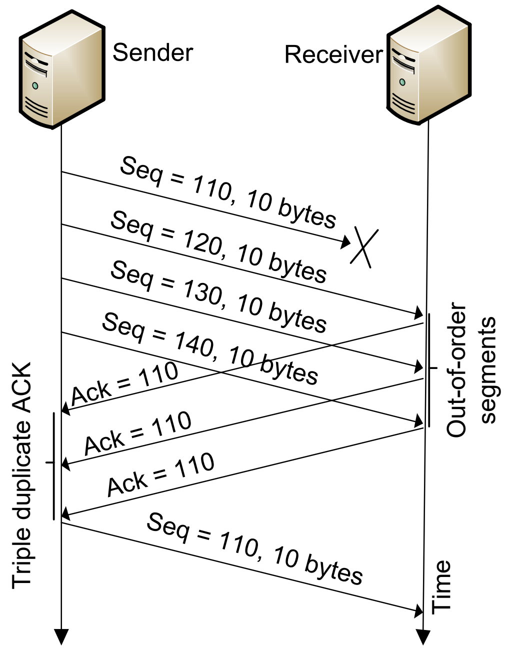


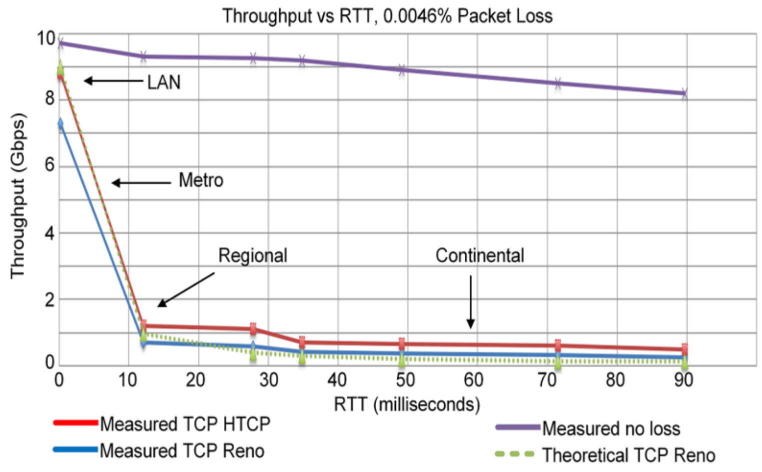
1. V. Jacobson, M. Karels, Congestion avoidance and control, ACM SIGCOMM Computer Communication Review 18 (4) (1988).

TCP Traditional Congestion Control

- The principles of window-based CC were described in the 1980s[1]
- Traditional CC algorithms follow the additive-increase multiplicative-decrease (AIMD) form of congestion control



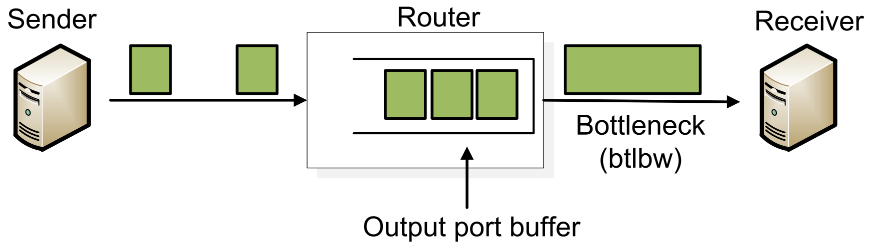




1. V. Jacobson, M. Karels, Congestion avoidance and control, ACM SIGCOMM Computer Communication Review 18 (4) (1988).

BBR: Model-based CC

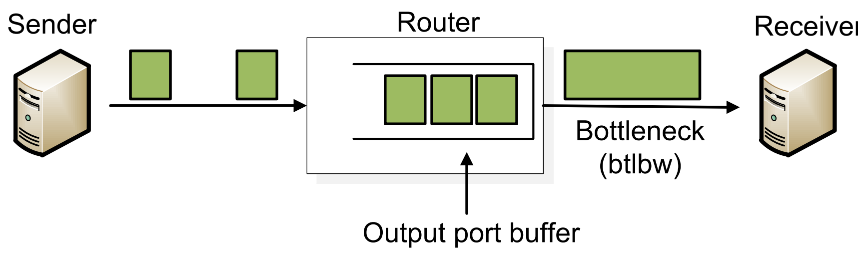
- TCP Bottleneck Bandwidth and RTT (BBR) is a rate-based congestion-control algorithm[1]
- BBR represented a disruption to the traditional CC algorithms:
  - is not governed by AIMD control law
  - does not use packet loss as a signal of congestion
- At any time, a TCP connection has one slowest link bottleneck bandwidth (btlbw)

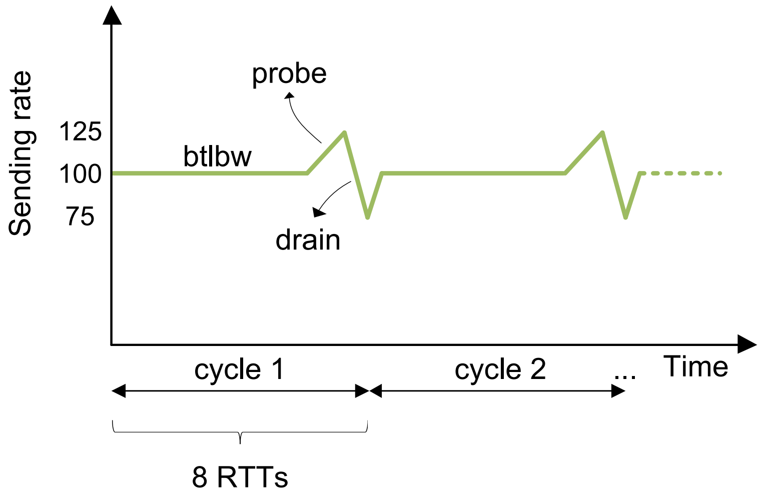


1. N. Cardwell et al. "BBR v2, A Model-based Congestion Control." IETF 104, March 2019.

BBR: Model-based CC

- TCP Bottleneck Bandwidth and RTT (BBR) is a rate-based congestion-control algorithm[1]
- BBR represented a disruption to the traditional CC algorithms:
  - is not governed by AIMD control law
  - does not use packet loss as a signal of congestion
- At any time, a TCP connection has one slowest link bottleneck bandwidth (btlbw)

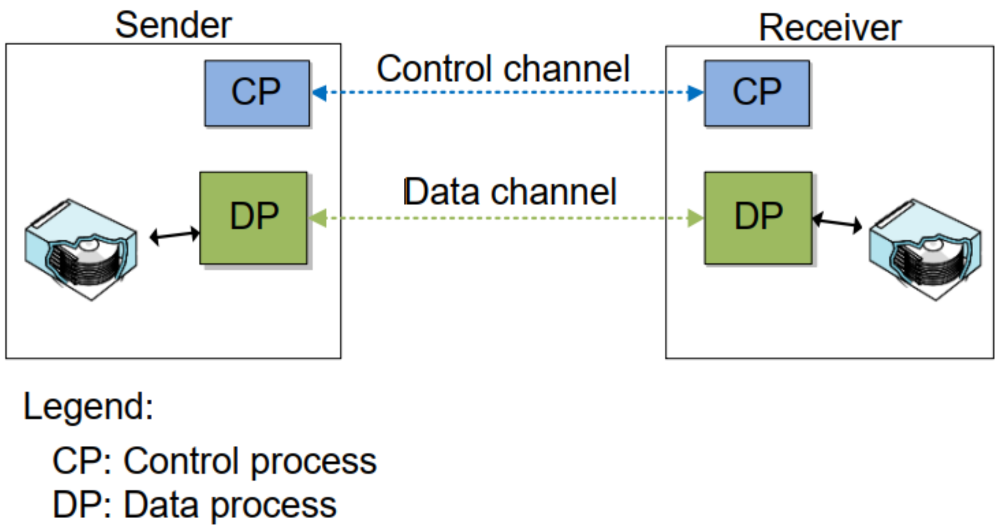




1. N. Cardwell et al. "BBR v2, A Model-based Congestion Control." IETF 104, March 2019.

Parallel Streams

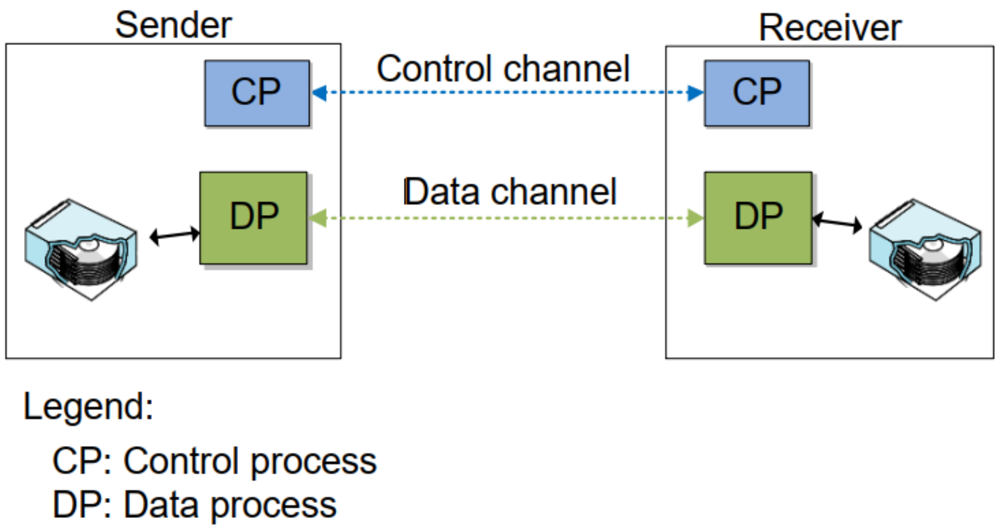
- Conventional file transfer protocols use a control channel and a (single) data channel (FTP model)



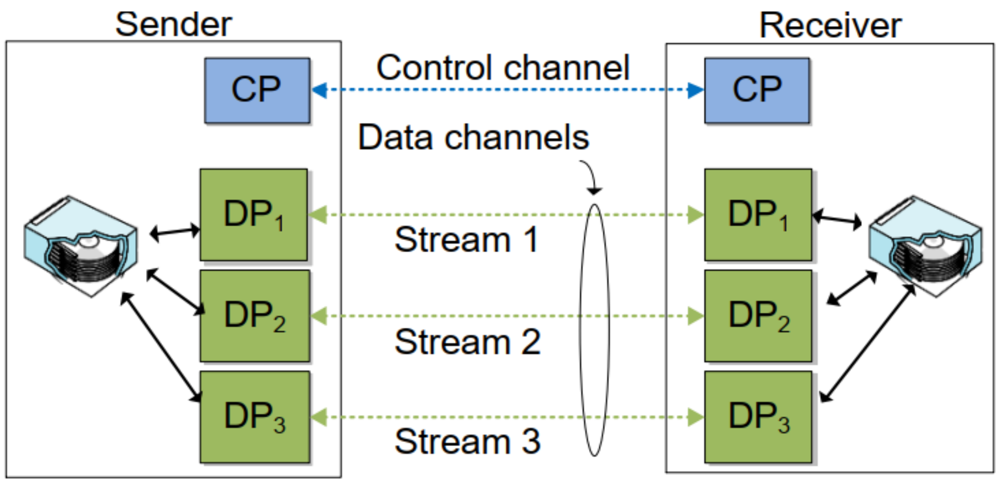
FTP model

Parallel Streams

- Conventional file transfer protocols use a control channel and a (single) data channel (FTP model)
- gridFTP is an extension of the FTP protocol
- A feature of gridFTP is the use of parallel streams



FTP model



gridFTP model

Advantages of Parallel Streams

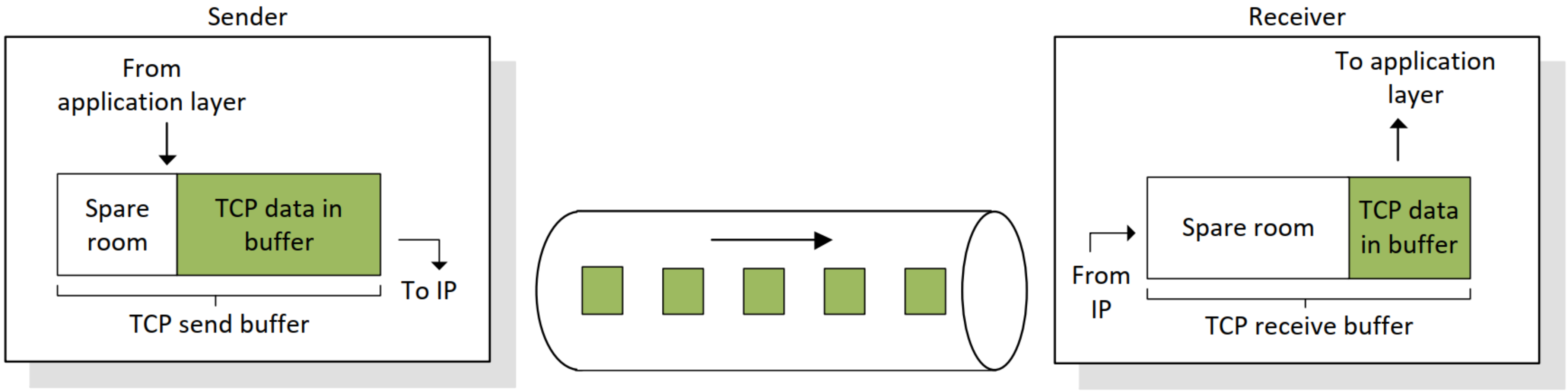
- Combat random packet loss not due congestion
  - Parallel streams increase the recovery speed after the multiplicative decrease

Advantages of Parallel Streams

- Combat random packet loss not due congestion
  - Parallel streams increase the recovery speed after the multiplicative decrease
- Mitigate TCP round-trip time (RTT) bias
  - A low-RTT flow gets a higher share of the bandwidth than that of a high-RTT flow
  - Increase bandwidth allocated to big science flows

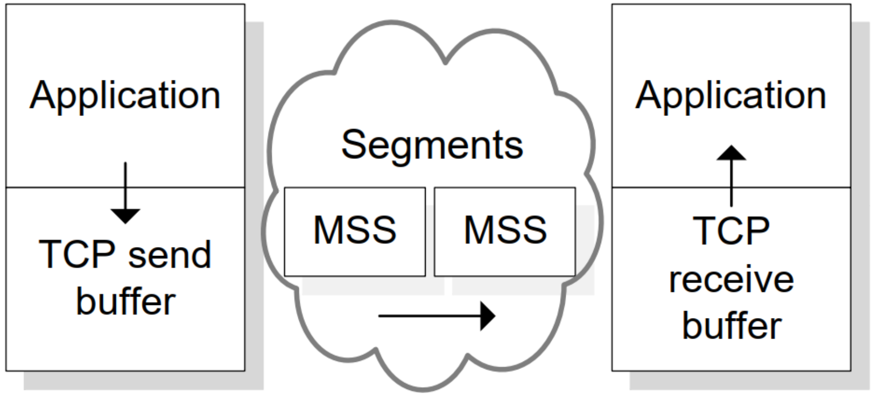
Advantages of Parallel Streams

- Combat random packet loss not due congestion
  - Parallel streams increase the recovery speed after the multiplicative decrease
- Mitigate TCP round-trip time (RTT) bias
  - A low-RTT flow gets a higher share of the bandwidth than that of a high-RTT flow
  - Increase bandwidth allocated to big science flows
- Overcome TCP buffer limitations
  - An application opening K parallel connections creates a virtual large buffer size on the aggregate connection that is K times the buffer size of a single connection



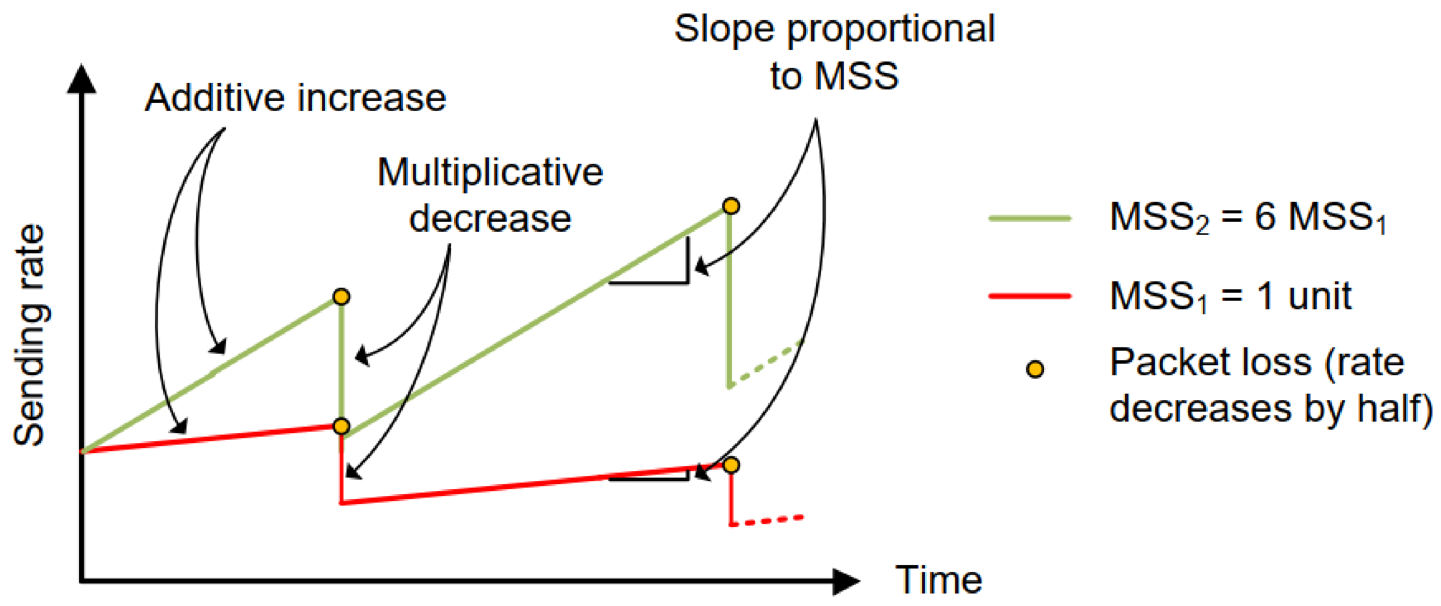
Maximum Segment Size (MSS)

- TCP receives data from application layer and places it in send buffer
- Data is typically broken into MSS units
- A typical MSS is 1,500 bytes, but it can be as large as 9,000 bytes



Advantages of Large MSS

- Less overhead
- The recovery after a packet loss is proportional to the MSS
  - ➢ During the additive increase phase, TCP increases the congestion window by approximately one MSS every RTT
  - ➢ By using a 9,000-byte MSS instead of a 1,500-byte MSS, the throughput increases six times faster

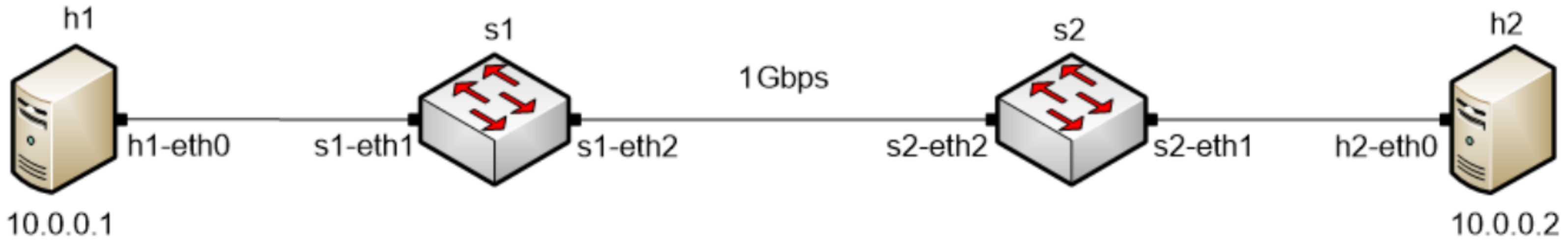


TCP Buffer Size

- In many WANs, the round-trip time (RTT) is dominated by the propagation delay
- To keep the sender busy while ACKs are received, the TCP buffer must be:

Traditional congestion controls: **TCP buffer size ≥ 2BDP**

BBRv1 and BBRv2: **TCP buffer size must be considerable larger than 2BDP**

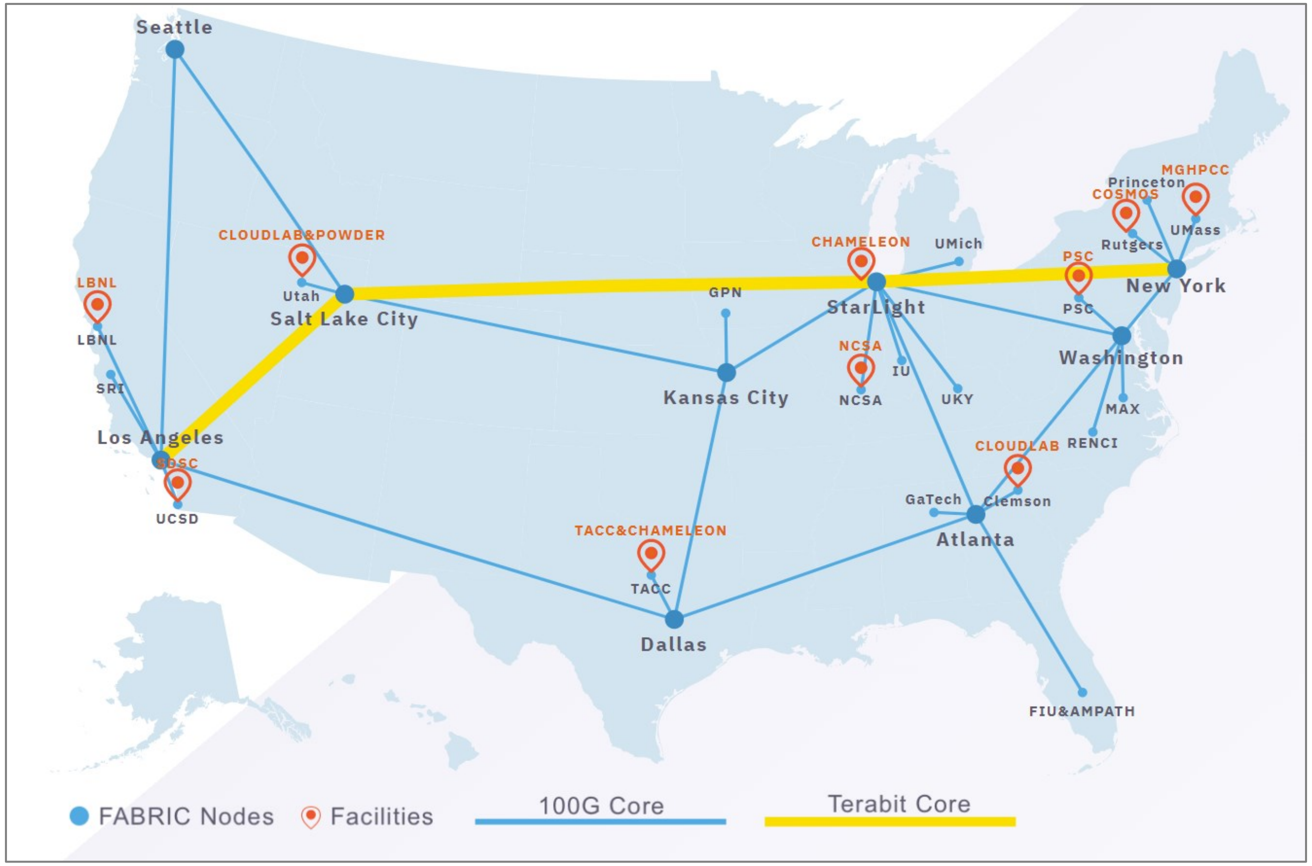


Summary

- There are many aspects of TCP / transport protocol that are essential to consider for high-performance networks
  - Parallel streams
  - MSS
  - TCP buffers
  - Router’s buffers, and others
- Still there is a need for applied research; e.g.,
  - Performance studies of new congestion control algorithms
  - TCP pacing
  - Application of programmable switches

Additional Slides

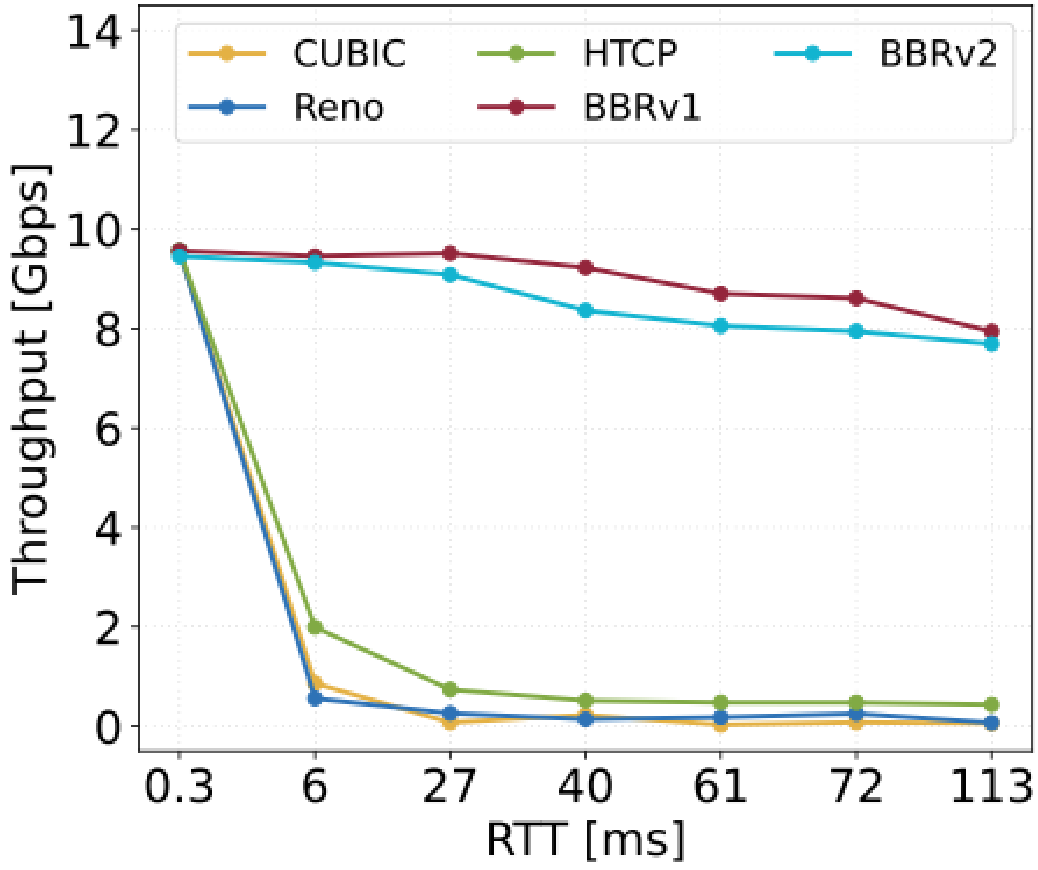
- BBR performance on FABRIC
- Performance measurements for a single flow, 0.0046% packet loss rate

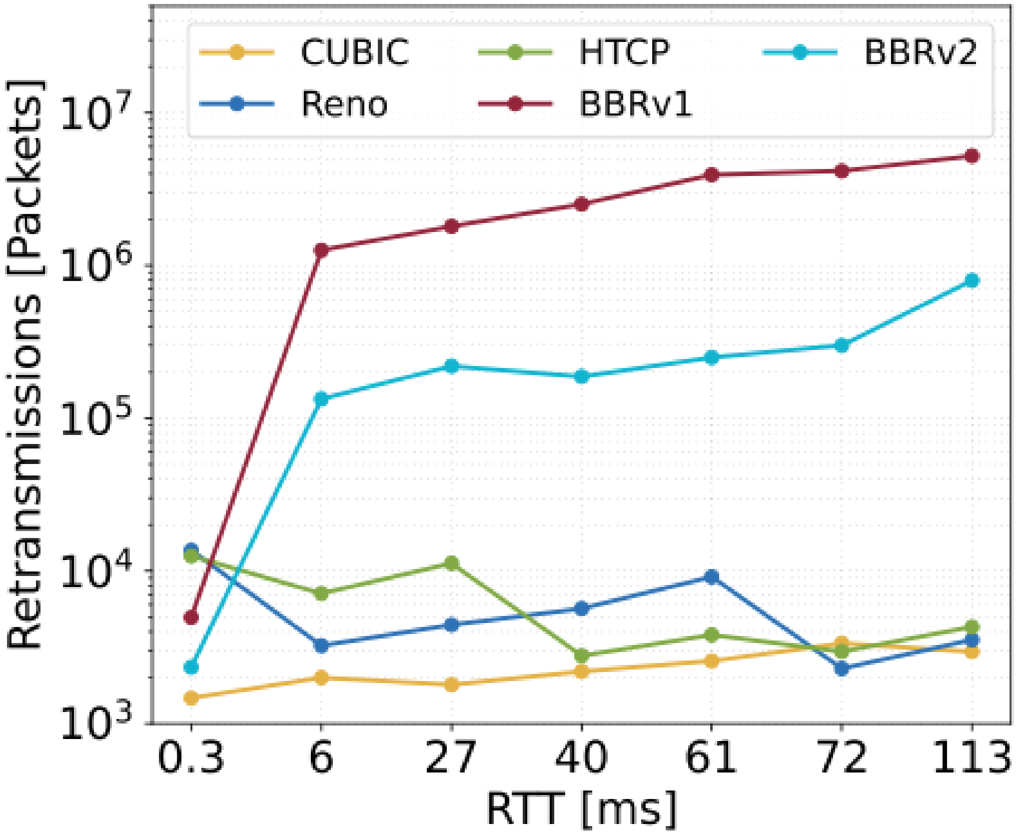


Site 1	Site 2	RTT
TACC (TX)	TACC (TX)	0.3ms
DALL (TX)	TACC (TX)	6ms
DALL (TX)	WASH (DC)	27ms
SALT (UT)	FIU (FL)	44ms
GPN (MO)	DALL (TX)	61ms
UTAH (UT)	WASH (DC)	72ms
GPN (MO)	FIU (FL)	113ms

Additional Slides

- BBR performance on FABRIC
- Performance measurements for a single flow, 0.0046% packet loss rate





BDP

- Bandwidth = 1Gbps
- RTT = 30ms
- BDP (bytes) = 3,750,000 bytes
- BDP (MB) = 3.57MB