"ROLE OF TCP IN LARGE DATA TRANSFERS"

J. Crichigno
Department of Integrated Information Technology
University of South Carolina

NSF Award 1829698
"CyberTraining CIP: Cyberinfrastructure Expertise on High-throughput Networks for Big Science Data Transfers"

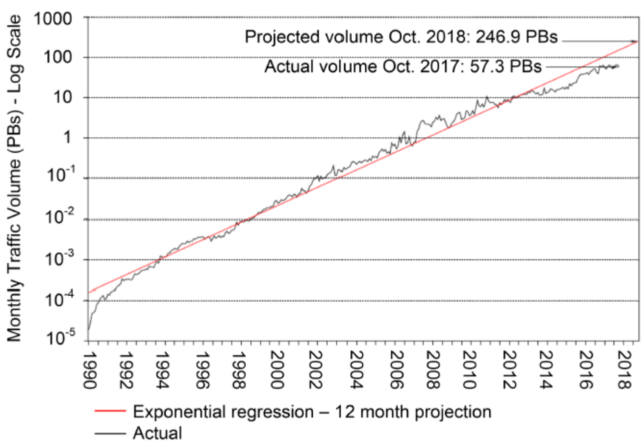
Agenda

- Motivation for a high-speed science architecture
- Enterprise network limitations
- Science DMZs
- TCP considerations
  - Congestion control algorithms
  - Parallel streams
  - Maximum Segment Size (MSS)
  - Pacing, fairness, TCP buffers, router's buffers, … (discussed in labs)

Motivation for a High-Speed Science Architecture

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

Applications

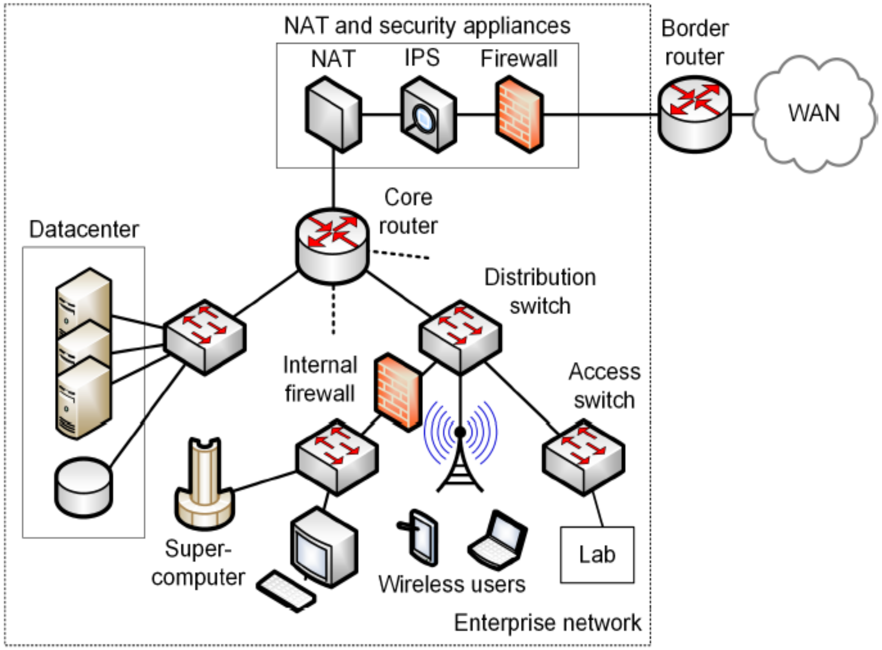


ESnet traffic

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

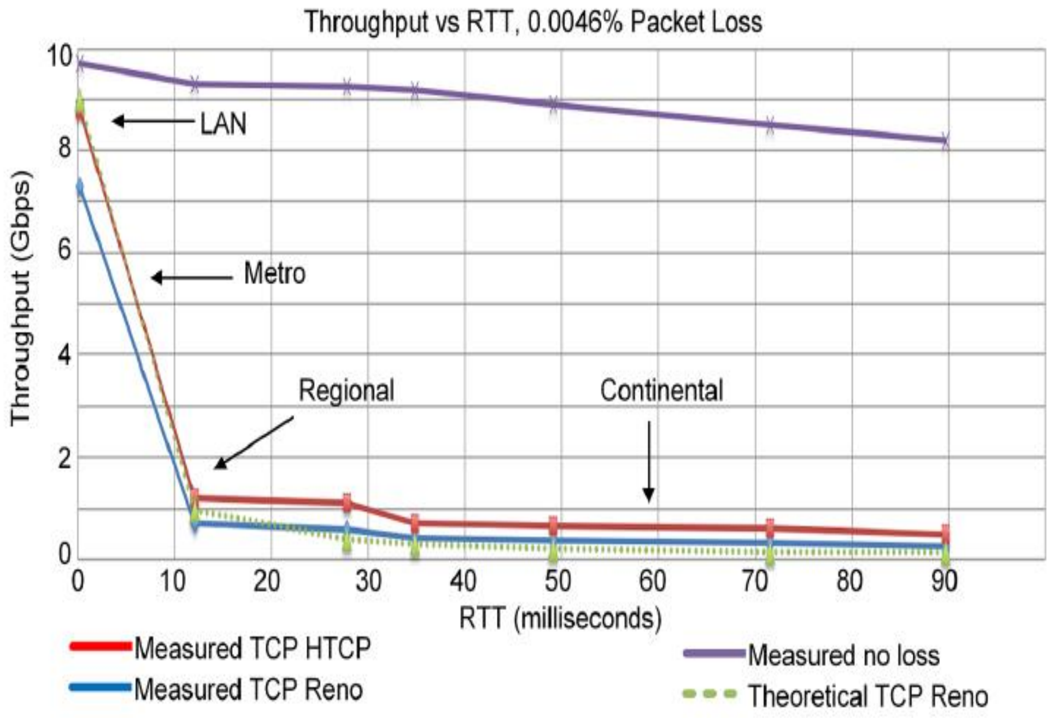
Enterprise Network Limitations

- 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

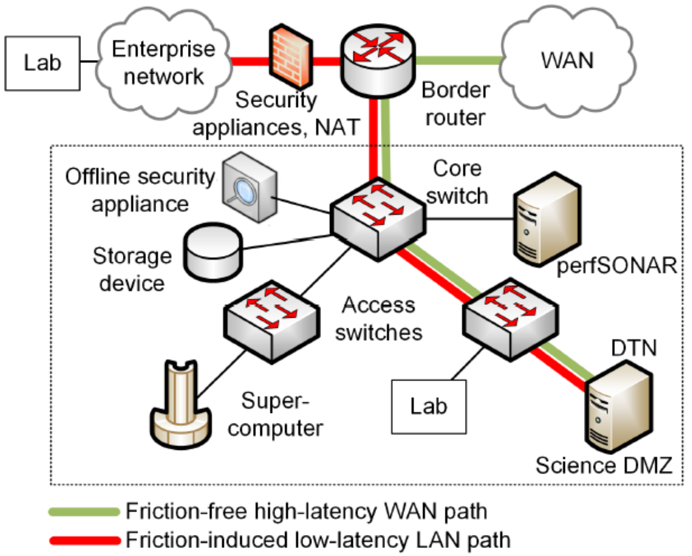
- Effect of packet loss and latency on TCP throughput



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.

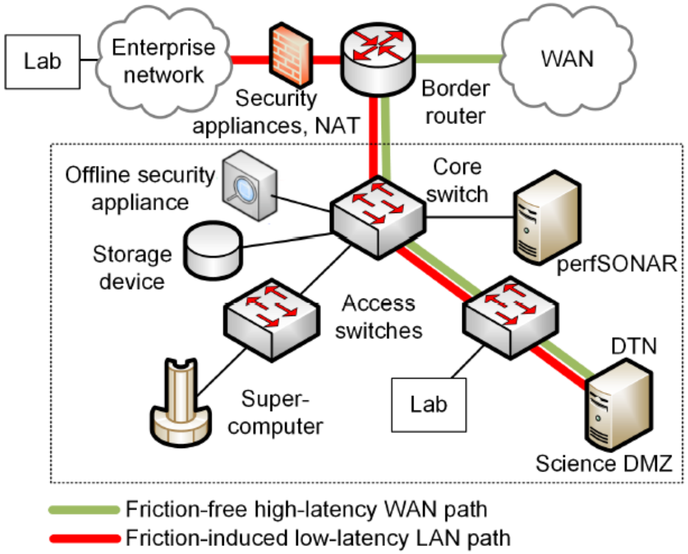
Science DMZ

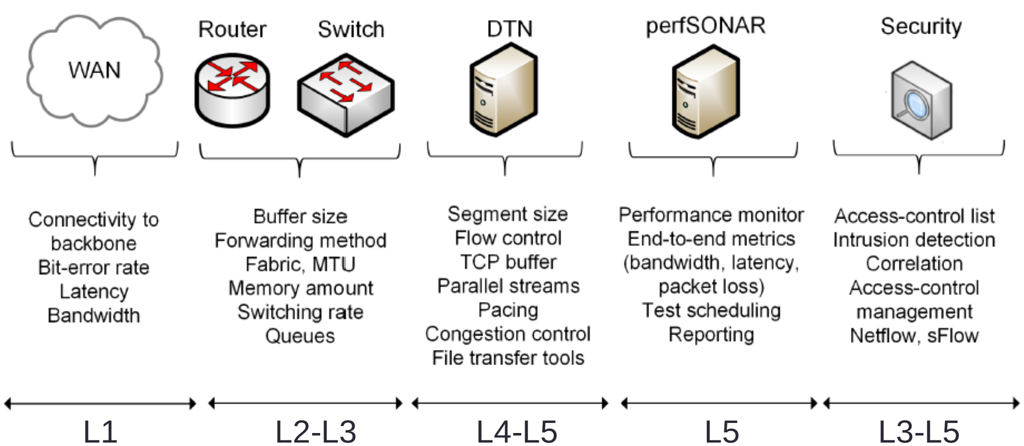
- The Science DMZ is a network designed for big science data
- Main elements
  - High throughput, friction free WAN paths
  - Data Transfer Nodes (DTNs)
  - End-to-end monitoring = perfSONAR
  - Security tailored for high speeds



Science DMZ

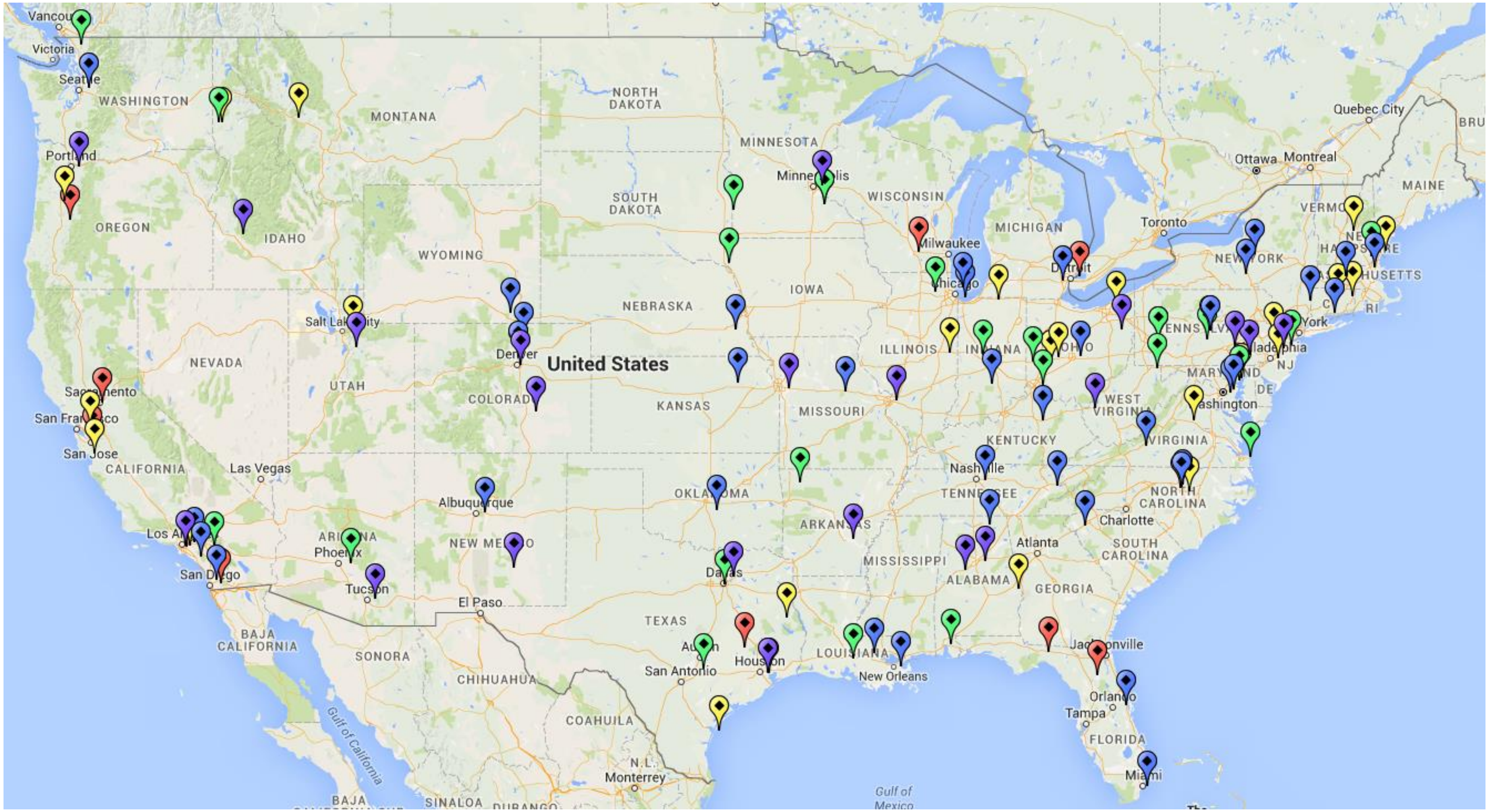
- The Science DMZ is a network designed for big science data
- Main elements
  - High throughput, friction free WAN paths
  - Data Transfer Nodes (DTNs)
  - End-to-end monitoring = perfSONAR
  - Security tailored for high speeds





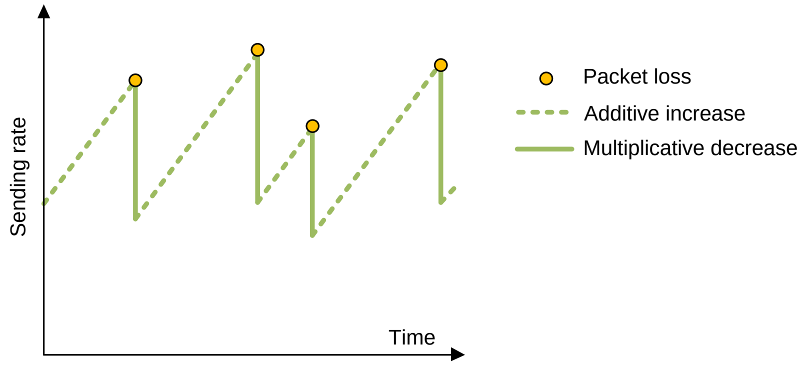
Science DMZ

- Science DMZ deployments, U.S.



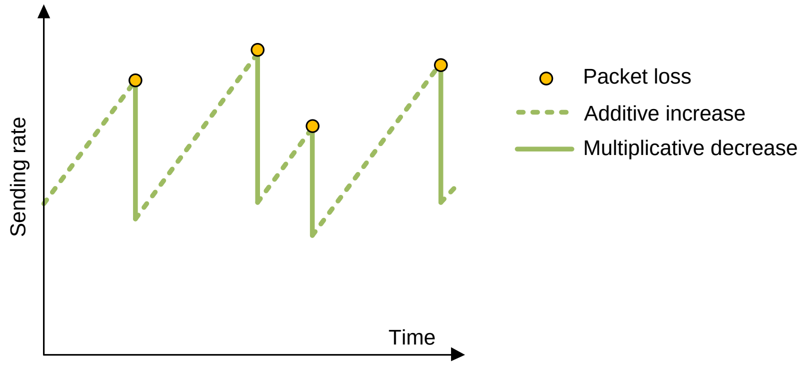
TCP Traditional Congestion Control (CC)

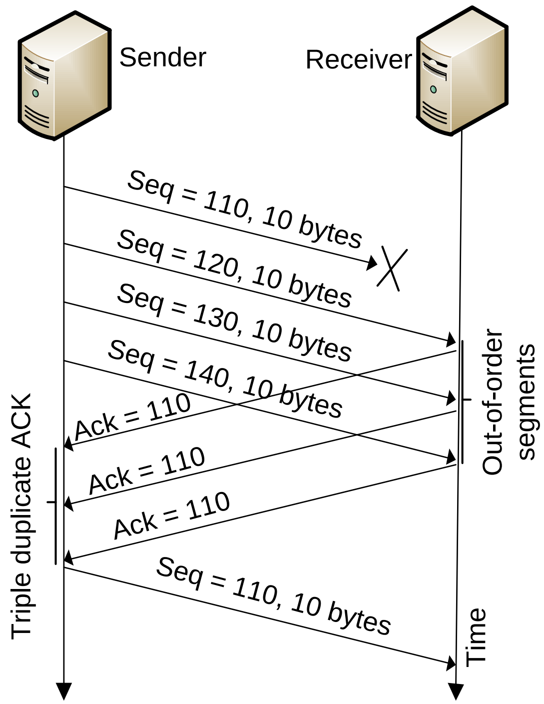
- The CC algorithm determines the sending rate
- Traditional CC algorithms follow an additive-increase multiplicative-decrease (AIMD) form of congestion control



TCP Traditional Congestion Control (CC)

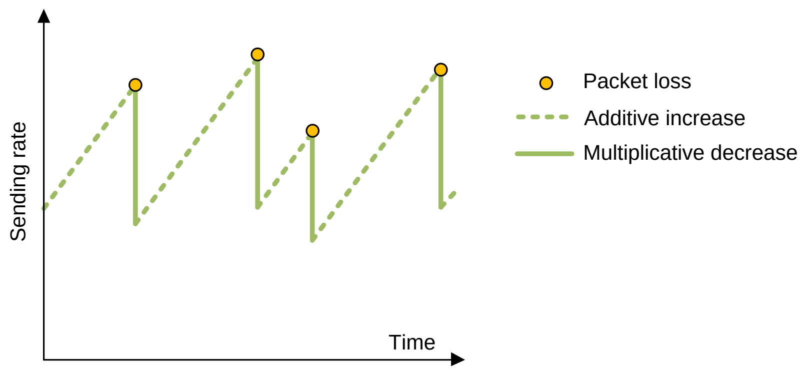
- The CC algorithm determines the sending rate
- Traditional CC algorithms follow an additive-increase multiplicative-decrease (AIMD) form of congestion control

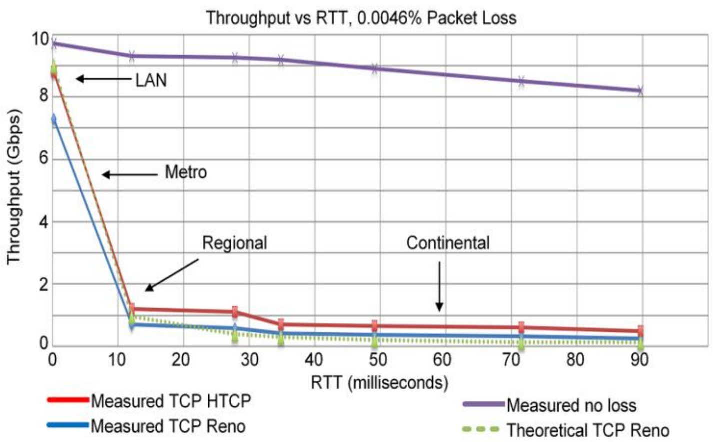


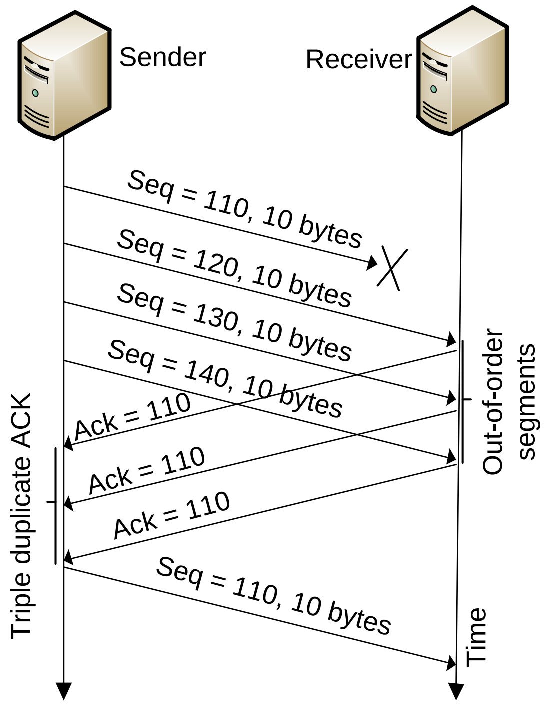


TCP Traditional Congestion Control (CC)

- The CC algorithm determines the sending rate
- Traditional CC algorithms follow an additive-increase multiplicative-decrease (AIMD) form of congestion control

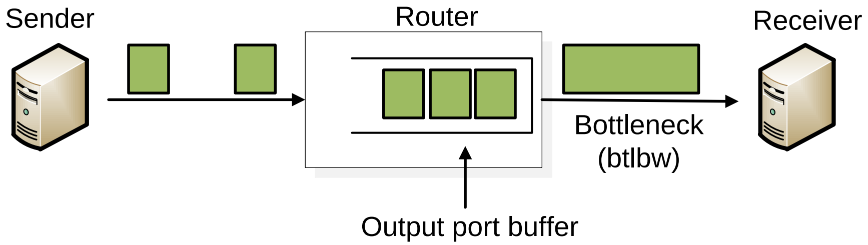






BBR: Rate-based CC

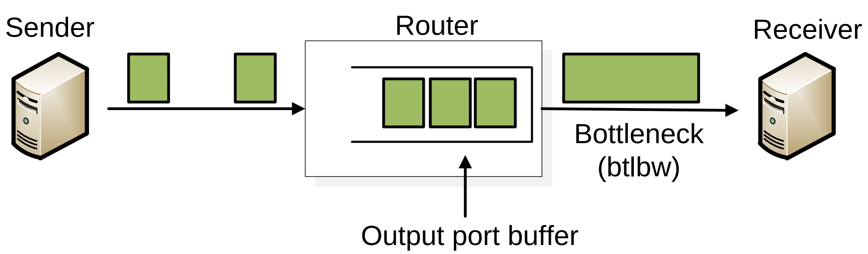
- TCP Bottleneck Bandwidth and RTT (BBR) is a rate-based congestion-control algorithm
- At any time, a TCP connection has one slowest link or bottleneck bandwidth (btlbw)

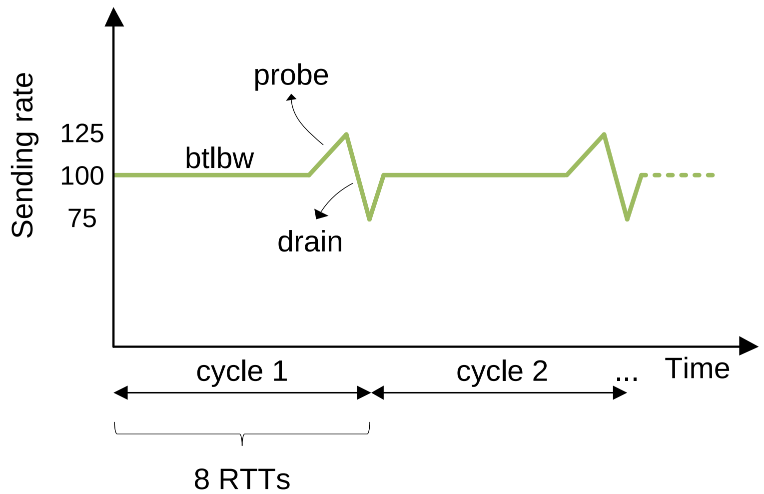


1. N. Cardwell, Y. Cheng, C. Gunn, S. Yeganeh, V. Jacobson, “BBR: congestion-based congestion control,” *Communications of the ACM*, vol 60, no. 2, pp. 58-66, Feb. 2017.
2. https://www.thequilt.net/wp-content/uploads/BBR-TCP-Opportunities.pdf

BBR: Rate-based CC

- TCP Bottleneck Bandwidth and RTT (BBR) is a rate-based congestion-control algorithm
- At any time, a TCP connection has one slowest link or bottleneck bandwidth (btlbw)
- BBR tries to find btlbw and set the sending rate to that value
  - The sending rate is independent of current packet losses; no AIMD rule

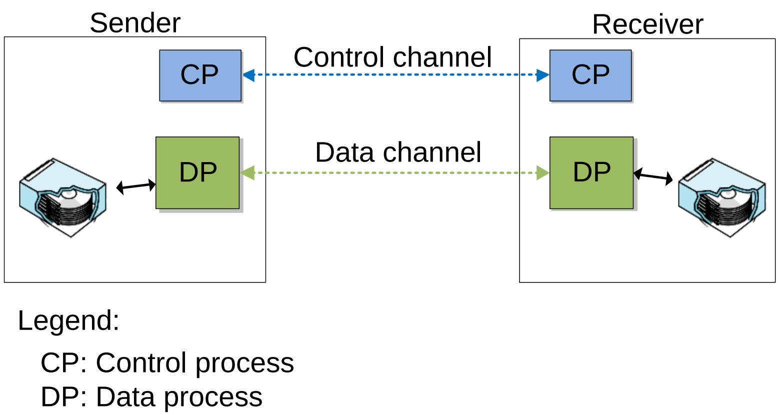




1. N. Cardwell, Y. Cheng, C. Gunn, S. Yeganeh, V. Jacobson, "BBR: congestion-based congestion control," *Communications of the ACM*, vol 60, no. 2, pp. 58-66, Feb. 2017.
2. https://www.thequilt.net/wp-content/uploads/BBR-TCP-Opportunities.pdf

Parallel Streams

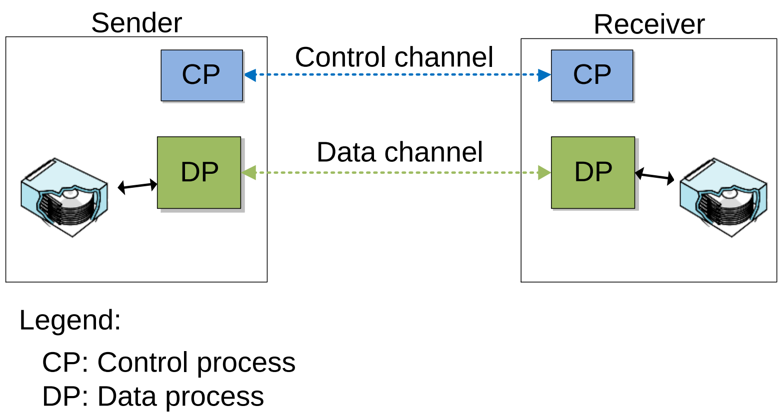
- Conventional file transfer protocols use a control channel and a (single) data channel (FTP model)



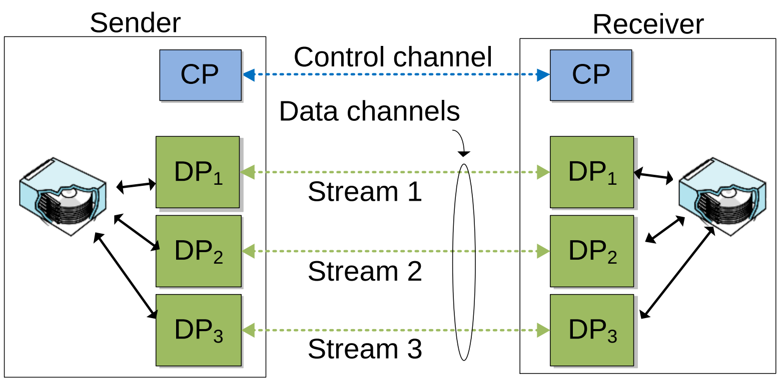
FTP model

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[1]
  - Parallel streams increase the recovery speed after the multiplicative decrease

1. T. Hacker, B. Athey, B. Noble, “The end-to-end performance effects of parallel TCP sockets on a lossy wide-area network,” in Proceedings of the Parallel and Distributed Processing Symposium, Apr. 2001.

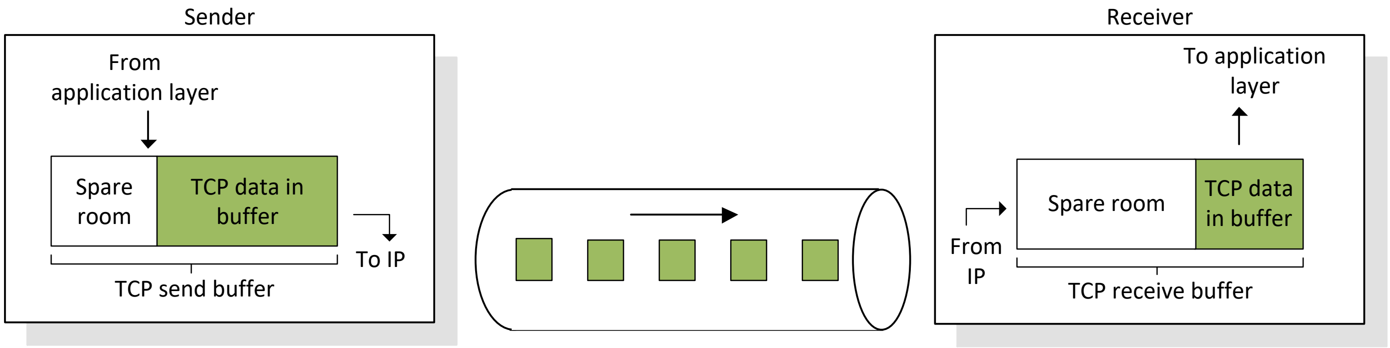
Advantages of Parallel Streams

- Combat random packet loss not due congestion[1]
  - Parallel streams increase the recovery speed after the multiplicative decrease
- Mitigate TCP round-trip time (RTT) bias[2]
  - 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

1. T. Hacker, B. Athey, B. Noble, “The end-to-end performance effects of parallel TCP sockets on a lossy wide-area network,” in Proceedings of the Parallel and Distributed Processing Symposium, Apr. 2001.
2. M. Mathis, J. Semke, J. Mahdavi, T. Ott, “The macroscopic behavior of the TCP congestion avoidance algorithm,” ACM Computer Communication Review, vol. 27, no 3, pp. 67-82, Jul. 1997.

Advantages of Parallel Streams

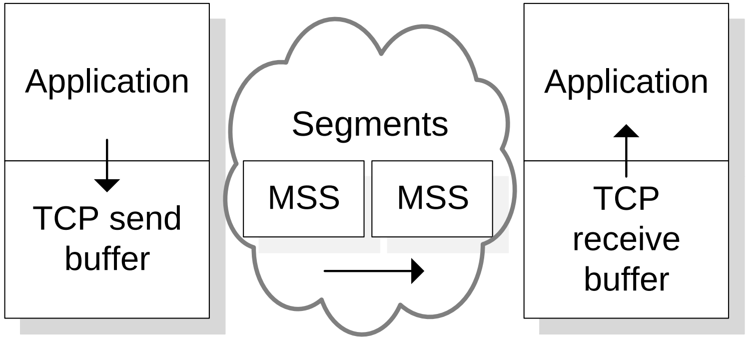
- Combat random packet loss not due congestion[1]
  - Parallel streams increase the recovery speed after the multiplicative decrease
- Mitigate TCP round-trip time (RTT) bias[2]
  - A low-RTT flow gets a higher share of the bandwidth than that of a high-RTT flow
  - Increase bandwidth allocated to big 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



1. T. Hacker, B. Athey, B. Noble, "The end-to-end performance effects of parallel TCP sockets on a lossy wide-area network," in Proceedings of the Parallel and Distributed Processing Symposium, Apr. 2001.
2. M. Mathis, J. Semke, J. Mahdavi, T. Ott, "The macroscopic behavior of the TCP congestion avoidance algorithm," ACM Computer Communication Review, vol. 27, no 3, pp. 67-82, Jul. 1997.

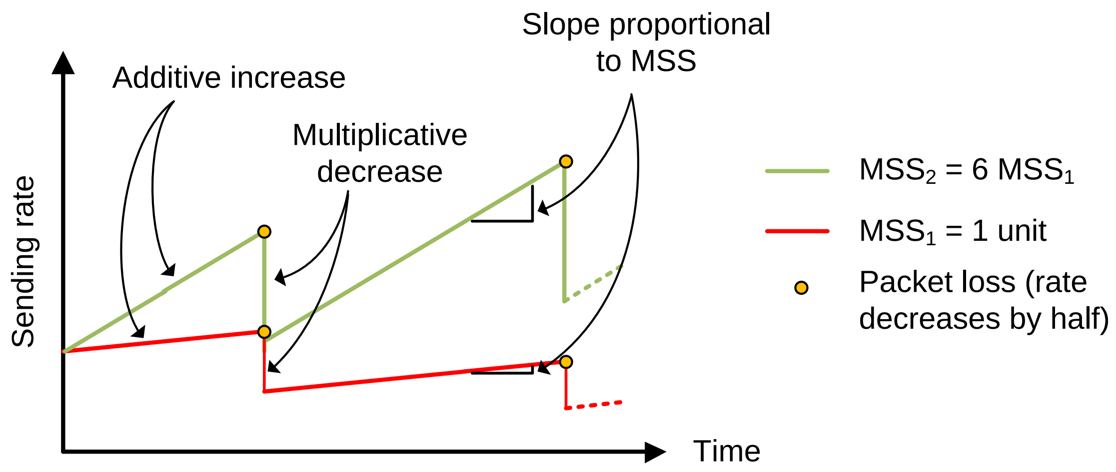
Maximum Segment Size (MSS)

- 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



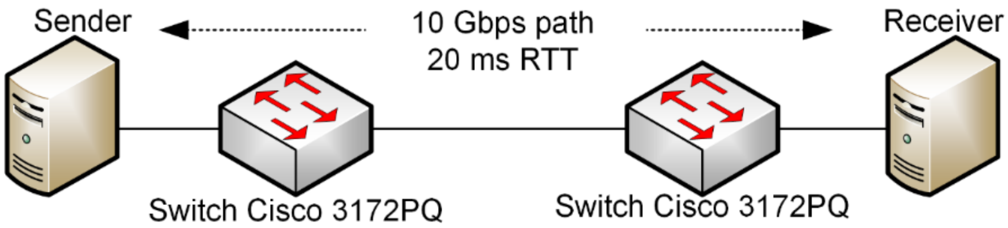
Advantages of Large MSS

- 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



Results on a 10 Gbps Network

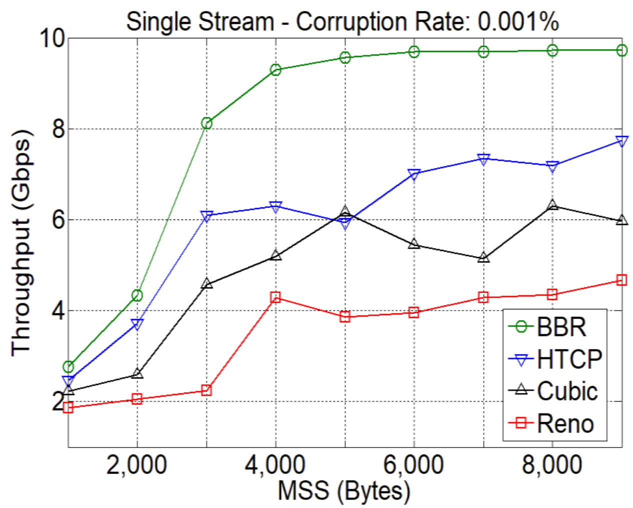
- 70-second experiments (first 10 seconds not considered)
- Ten experiments conducted and the average throughput is reported
- Impact of MSS and parallel streams on BBR, Reno, HTCP, Cubic



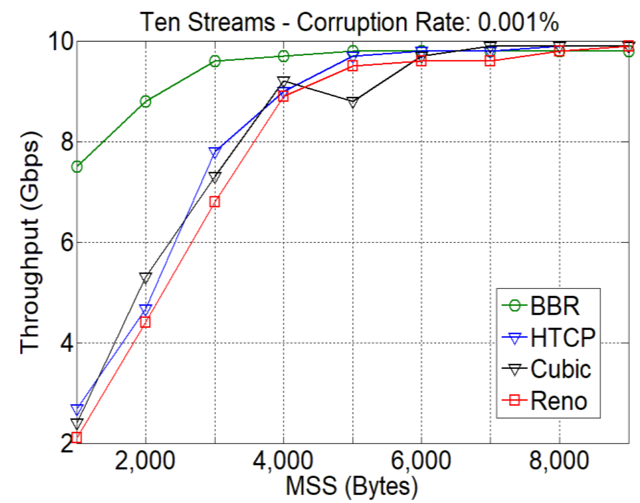
1. J. Crichigno, Z. Csibi, E. Bou-Harb, N. Ghani, "Impact of segment size and parallel streams on TCP BBR," IEEE Telecommunications and Signal Processing Conference (TSP), Athens, Greece, July 2018.

Results on a 10 Gbps Network

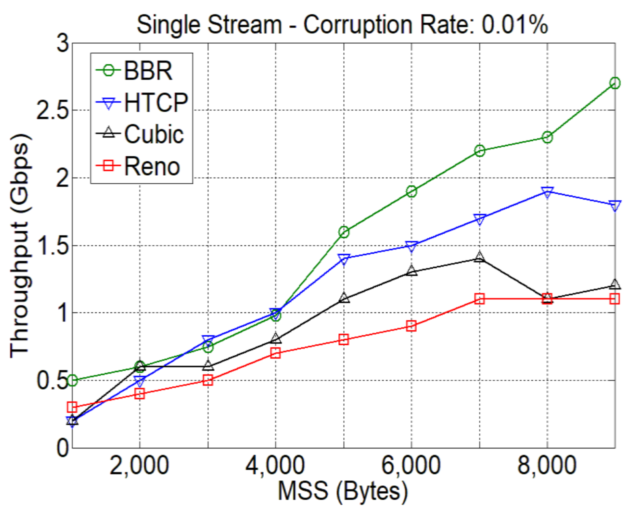
1



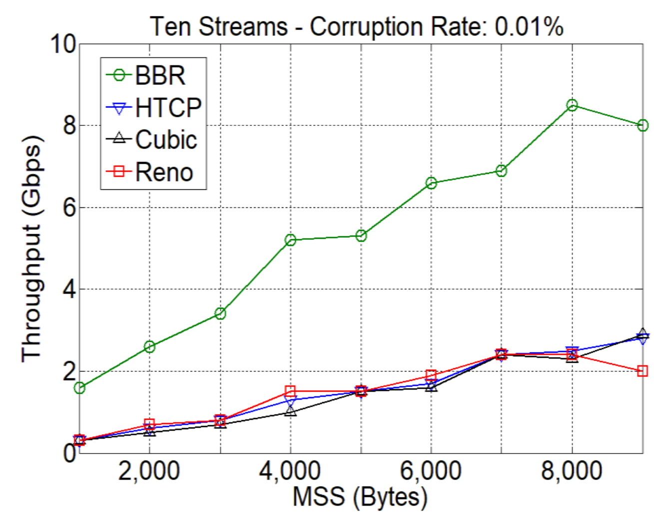
2



3



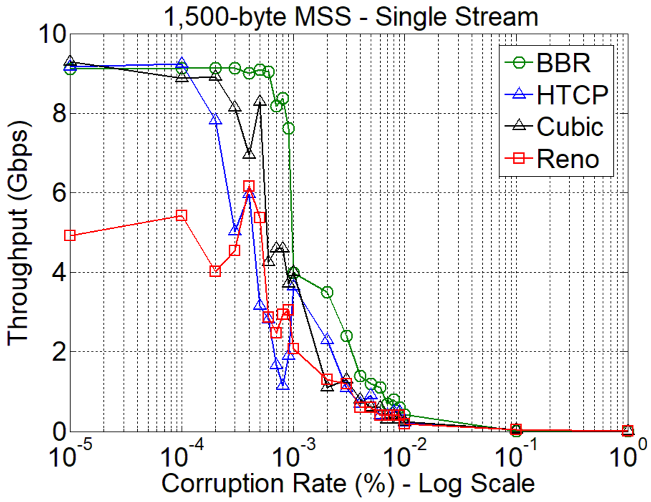
4



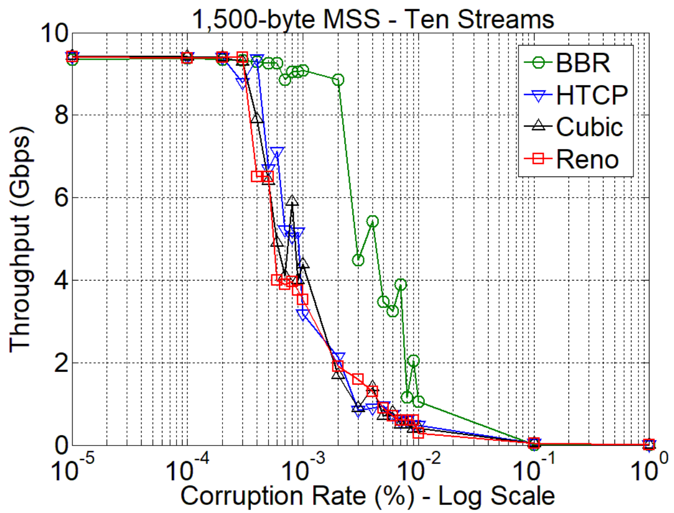
1. J. Crichigno, Z. Csibi, E. Bou-Harb, N. Ghani, "Impact of segment size and parallel streams on TCP BBR," IEEE Telecommunications and Signal Processing Conference (TSP), Athens, Greece, July 2018.

Results on a 10 Gbps Network

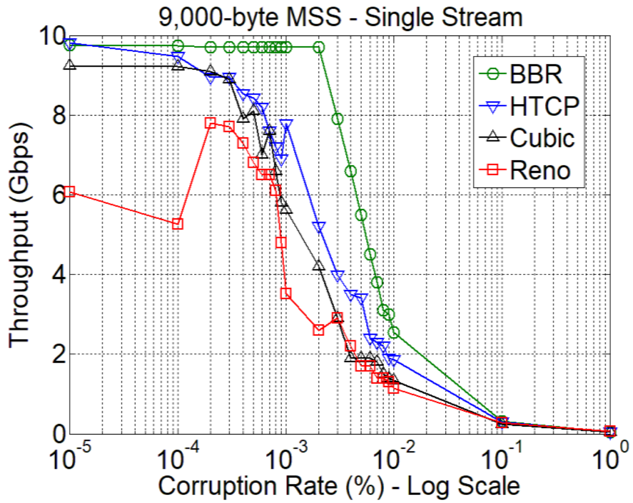
1



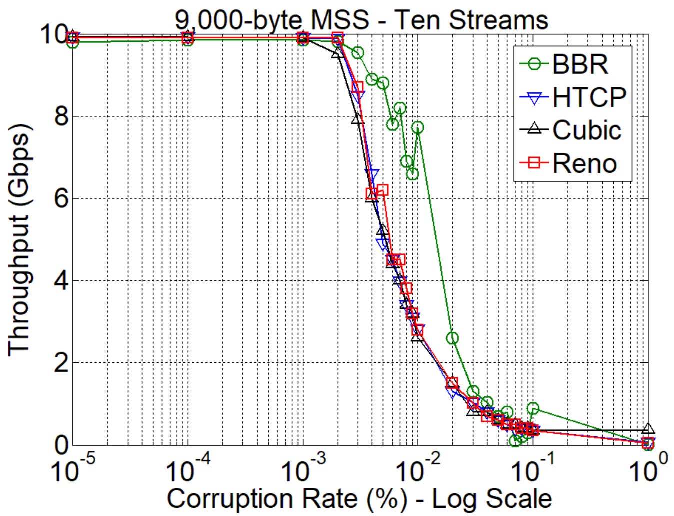
2



3



4



1. J. Crichigno, Z. Csibi, E. Bou-Harb, N. Ghani, “Impact of segment size and parallel streams on TCP BBR,” IEEE Telecommunications and Signal Processing Conference (TSP), Athens, Greece, July 2018.

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