TCP IN LARGE DATA TRANSFERS

Jorge Crichigno, Elie Kfoury
Department of Integrated Information Technology
University of South Carolina

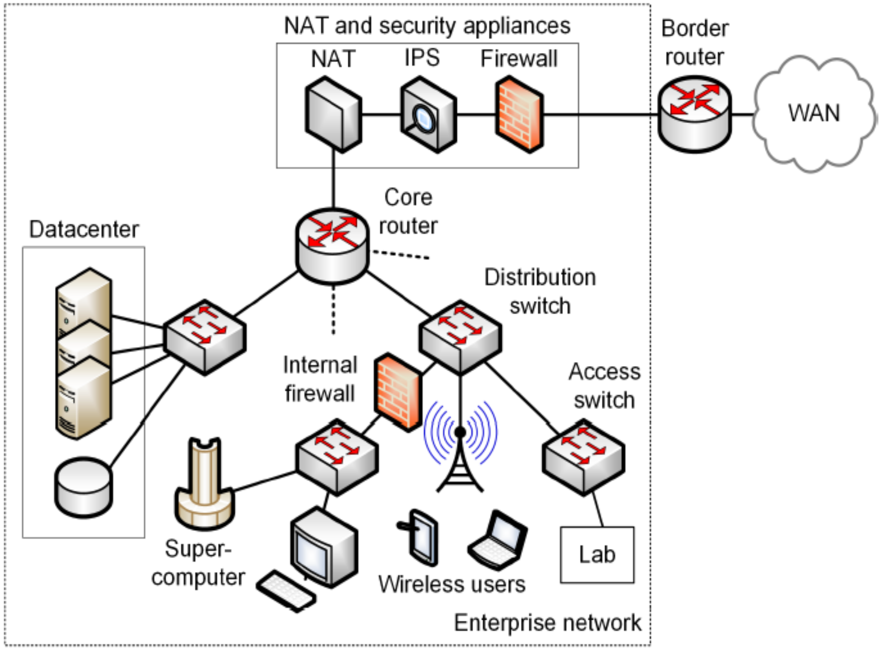
NTU Technical Workshop
July 31 – August 1, 2019

Agenda

- 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)

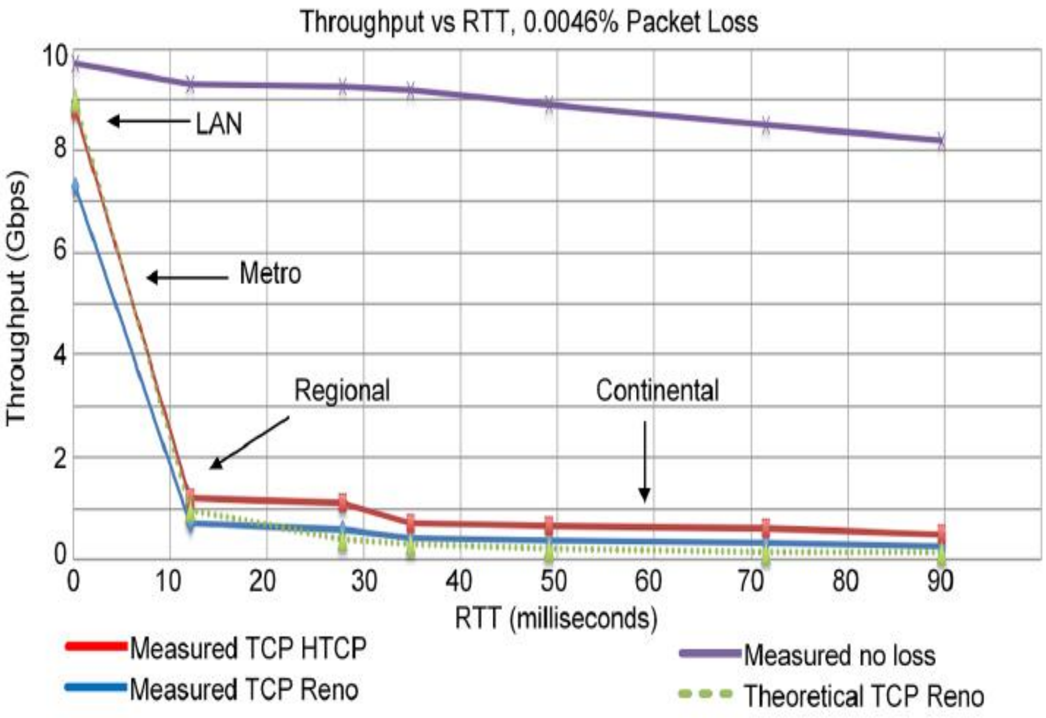
Enterprise Network Limitations

- 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
- Many of the issues above relate to TCP



Enterprise Network Limitations

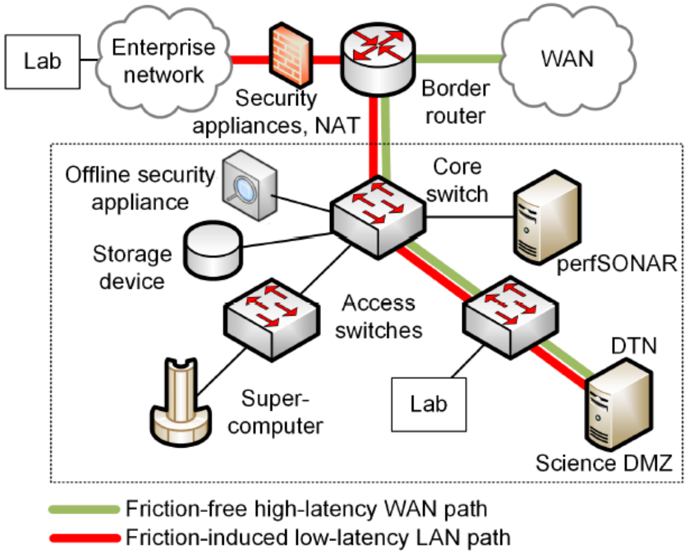
- Effect of packet loss and latency on TCP throughput



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.

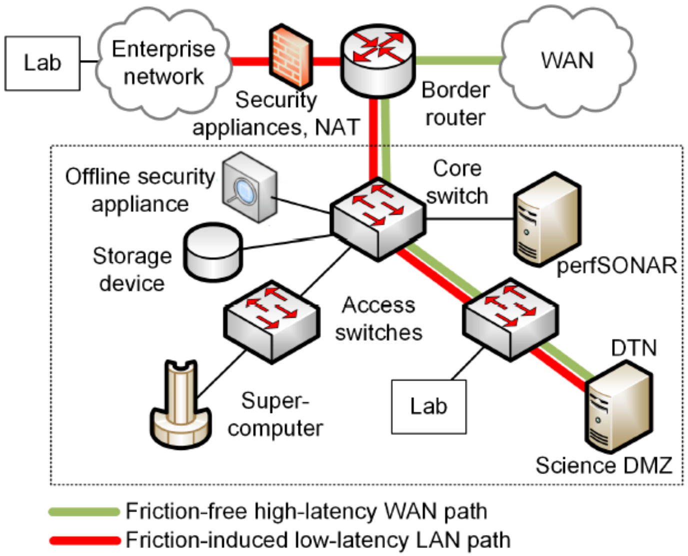
Science DMZ

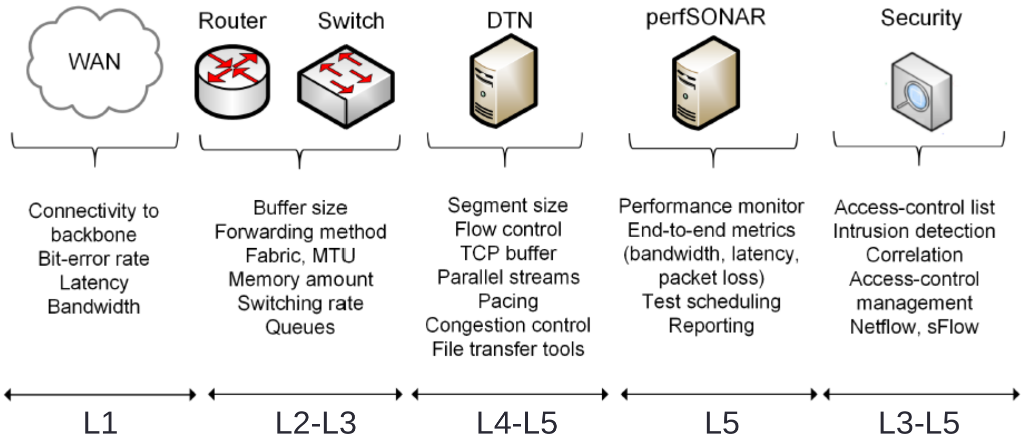
- The Science DMZ is a network designed for big science data
- Main elements
  - High throughput, friction free WAN paths
  - Data Transfer Nodes (DTNs)
  - End-to-end monitoring = perfSONAR
  - Security tailored for high speeds



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 Needs

- USC

Researchers	Topic	Current support	Requirements
Gothe Ilieva Strauch	Experimental nuclear physics (ENP)	NSF: 1505615 ($1.2M), 1614773 ($610K), 1812382 ($350K); Brookhaven National Laboratory (BNL) 218624 ($15K); Jefferson Science Associates / DOE ($11K)	100 Gbps throughput to PSI, JLab. High throughput to other collaborators (Brookhaven, Argonne)
Heyden Lauterbach	Chemical engineering	NSF: 1254352 ($400K), 1534260 ($840K), 1565964 ($300K), 1832809 ($160K), 1632824 ($3M), 1805307 ($75K)	High throughput (at least 10 Gbps) to XSEDE (SDSC, TACC), PNNL
Bayoumi	Aerospace, predictive maintenance	Siemens ($628M in-kind [44]), Boeing ($5M [45]), DOD hq017-17-c-7110 ($240K), Missile Def. Ag. HQ0147-16-C-7606 ($35K), Boeing SSOW-BRT-W0915-0001 ($275K)	High throughput with encryption (10 Gbps) to internal and external HPCs, XSEDE, SDSC, TACC
Baalousha Lead	Environment nanoscience	NSF: 1828055 ($635K), 1738340 ($286K), 1655926 (4K), 1553909 ($510K), 1437307 ($300K), 1508931 ($390K), 1834638 ($380K); DOD 450388-19545 ($380K); NIEH 1P01ES028942-01 ($6M), NIH R03ES027406-01 ($144K).	High throughput (5 Gbps) connection from TOF-ICP-MS instrument to Internet2
Sutton Xiaomin Kidane	Digital image correlation (DIC)	NASA C15-2A38-USC ($1.2M), NSF 1537776 ($165K), Boeing SSOW-BRT-W0915-0003 ($140K)	High throughput from USC's DIC laboratory to HPCs (SDSC, TACC) running ABAQUS, ANSYS
Porter	Ntl. Estuarine Research Reserve System	NOAA: NA18NOS4200120 ($760K), NA17NOS4200104 ($980K), OOS.16 (028)USC.DP.MOD.1 ($100K), U. Mich. 3003300692 ($340K), FL Env. Protection CM08P ($92K), NIEHS 1P01ES028942-01 ($6M), USDA ($43K).	High throughput from NOAA's NERRS repository (located at USC) to Internet2 (large datasets downloads worldwide)
Avignone Guiseppe	Particle astrophysics	NSF 1614611 ($900K), NSF 1307204 ($1M), NSF 1808426 ($306K)	100 Gbps connection to MAJORANA (SD), CUORE (Italy), NERSC (CA)
Chandra	Semiconductor material	NSF: 1810116 ($371K), 1711322 ($370K), 1553634 ($695K); NIBIB 1R03EB026813-01 ($136K), DOD W911NF-18-1-0029 ($585K), SRNL/DOE UC150 ($24K), DOE DE-SC0019360 ($666K), RCSA 23976 ($100K)	High throughput (at least 10 Gbps) from X-ray photoelectron spectroscopy instrument and storage to Internet2 (SRNL, INL, Sandia, other institutions)

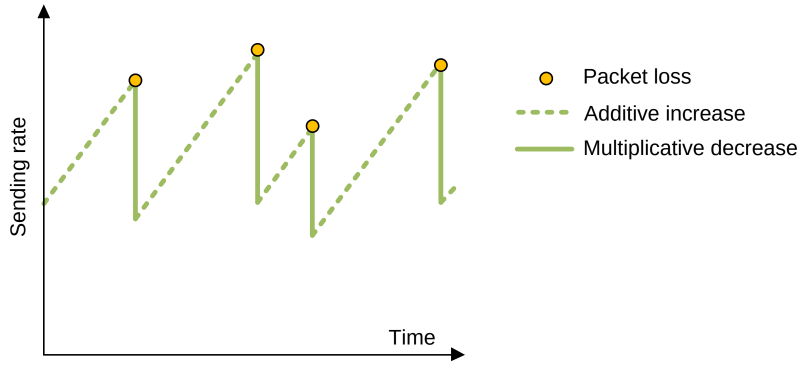
Science DMZ Needs

- USC

Chandra Shustova	Semiconductor material	NSF: 1810116 ($371K), 1711322 ($370K), 1553634 ($695K); NIBIB 1R03EB026813-01 ($136K), DOD W911NF-18-1-0029 ($585K), SRNL/DOE UC150 ($24K), DOE DE-SC0019360 ($666K), RCSA 23976 ($100K)	High throughput (at least 10 Gbps) from X-ray photoelectron spectroscopy instrument and storage to Internet2 (SRNL, INL, Sandia, other institutions)
Richardson Myrick	Phytoplankton spectroscopy	NSF 1542555 ($2M) and DXP Supply Chain Services ($40K)	High throughput (10 Gbps) from image photometer, storage to internal and external HPC
Norman	Genomics data mining	NSF 1149447 ($850K), NIEH 1P01ES028942-01 ($6M), NSF SC EPSCoR 2031-231-2022570 ($100K)	100 Gbps throughput from genomics seq. instrument/storage to USC's HPC; 10+ Gbps connection to Frederick, Argonne, Oak Ridge Ntl. Laboratories, XSEDE resources
Pinckney Benitez	Estuarine ecology	NSF 1736557 ($1M), NOAA R/ER-49 ($130K), NSF 1829519 ($265K), NSF 1458416 ($593K), NSF 1433313 ($362K), NASA 23175500 ($167K)	High throughput from USC's estuarine database to HPCs and Internet2 (datasets downloads)
Dudycha	Genomics, aquatic biology	NSF 1556645 ($1.2M), SC Sea Grant Consortium/NOAA/DOC N250 ($40K), DOD W81XWH1810088 ($287K)	100 Gbps connection to USC's HPC; 10+ Gbps connection to transport DNA / RNA-seq. datasets to XSEDE
Vasquez	Math, genome dynamics	NSF: 1751339 ($290K), 1410047 ($210K)	100 Gbps connection from genomics laboratory to USC's HPC, XSEDE
Brooks Hikmet Schooley	Mathematical models for patient treatment	SC Department of Commerce ($300K), Duke Endowment Child Care Division 1971-SP ($646K), American Cancer Society IRG-17-179-04 ($30K), Patient-Centered Outcomes Research Institute ME-1303-6011 ($960K)	100 Gbps connection from engineering storage to USC's HPC
Ramstad Shervette Ghoshroy	Other USC campuses, genomics	NOAA/DOC NA18NMF4330239 ($503K), NOAA/DOC NA18NMF4270203 ($230K), NOAA NA17NMF4540137 ($153K), NOAA 719583-712683 ($189K), NOAA NA15NMF4330157 ($466K).	10 Gbps connection to move datasets between USC Aiken - Internet2
Crichigno	Cyberinfrast.	NSF 1822567 ($420K), NSF 1829698 ($500K)	100 Gbps programmable network

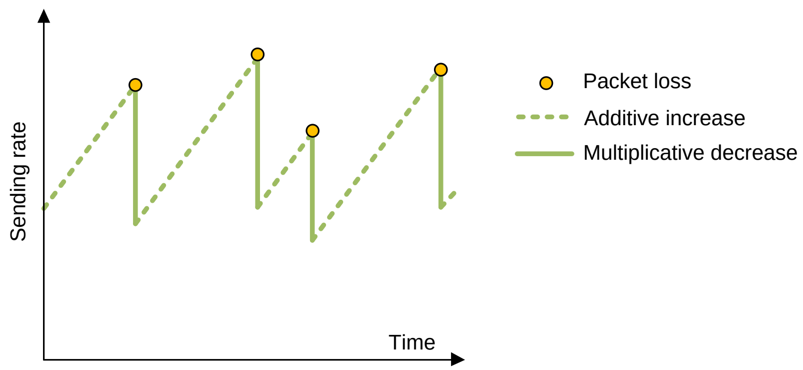
TCP Traditional Congestion Control (CC)

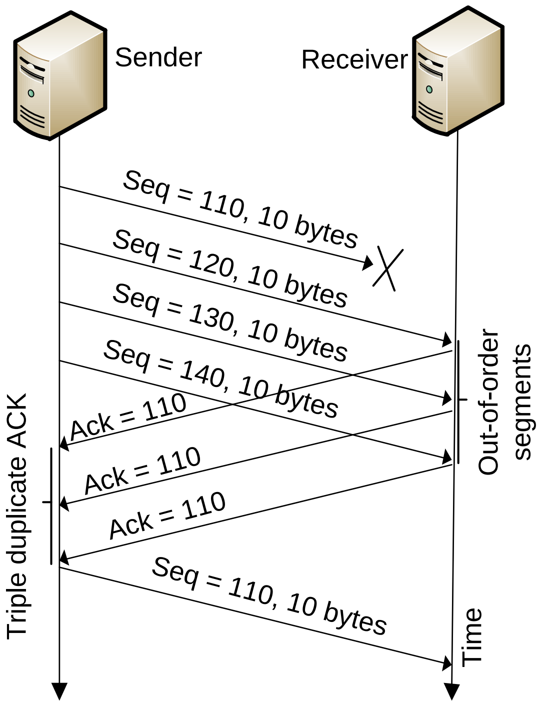
- The CC algorithm determines the sending rate
- Traditional CC algorithms follow an additive-increase multiplicative-decrease (AIMD) form of congestion control



TCP Traditional Congestion Control (CC)

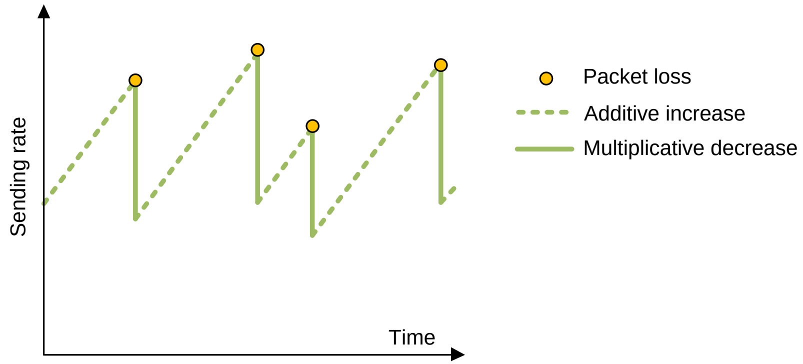
- The CC algorithm determines the sending rate
- Traditional CC algorithms follow an additive-increase multiplicative-decrease (AIMD) form of congestion control

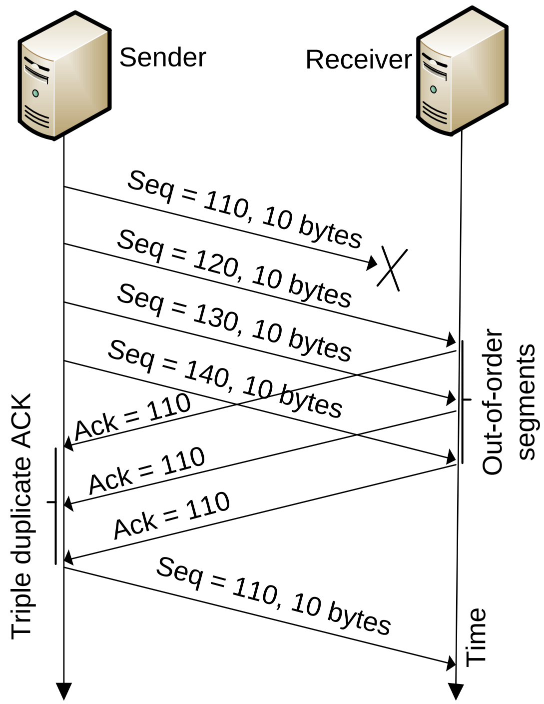


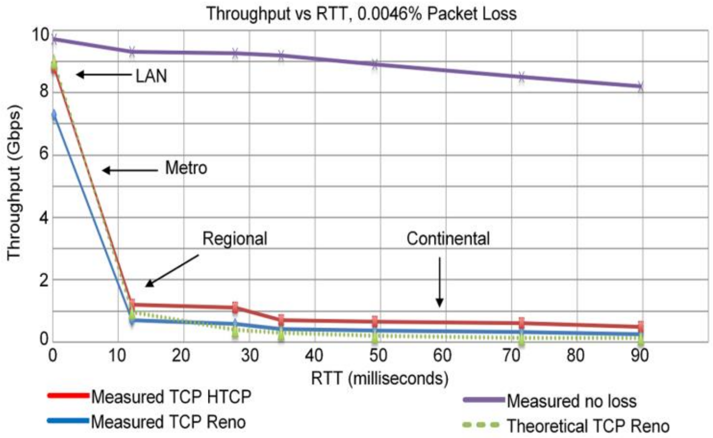


TCP Traditional Congestion Control (CC)

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

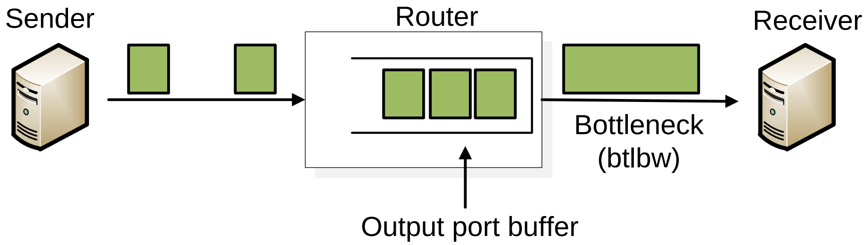






BBR: Rate-based CC

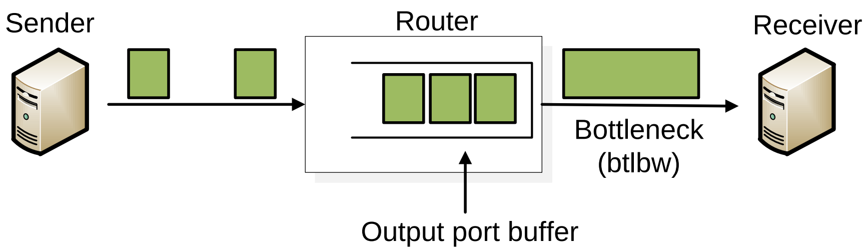
- TCP Bottleneck Bandwidth and RTT (BBR) is a rate-based congestion-control algorithm
- At any time, a TCP connection has one slowest link or bottleneck bandwidth (btlbw)

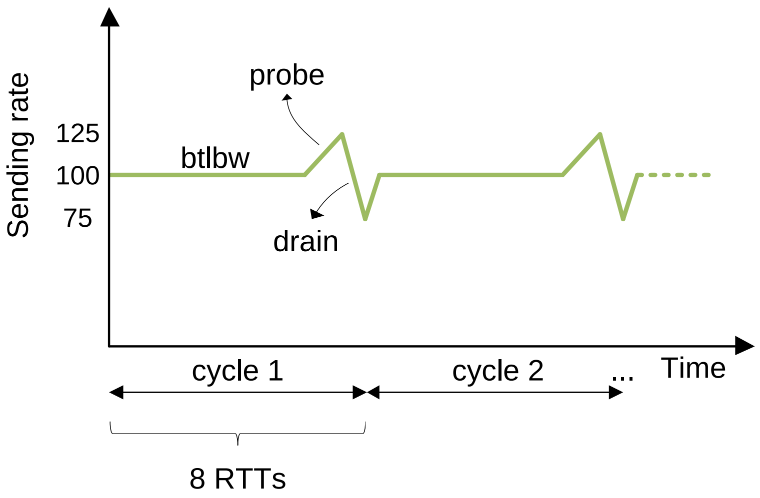


1. N. Cardwell, Y. Cheng, C. Gunn, S. Yeganeh, V. Jacobson, “BBR: congestion-based congestion control,” *Communications of the ACM*, vol 60, no. 2, pp. 58-66, Feb. 2017.
2. https://www.thequilt.net/wp-content/uploads/BBR-TCP-Opportunities.pdf

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

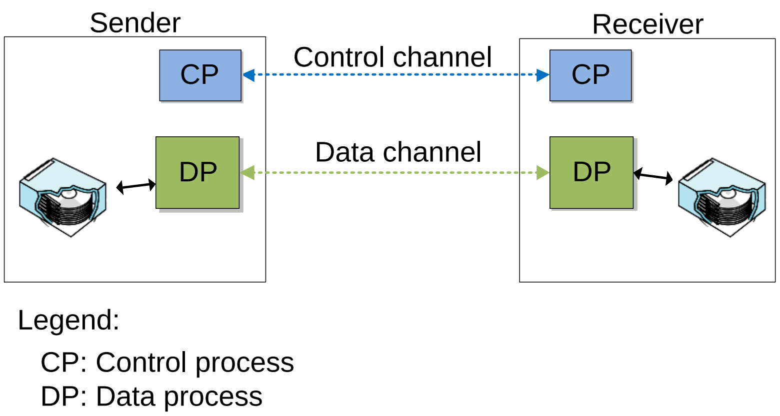




1. N. Cardwell, Y. Cheng, C. Gunn, S. Yeganeh, V. Jacobson, "BBR: congestion-based congestion control," *Communications of the ACM*, vol 60, no. 2, pp. 58-66, Feb. 2017.
2. https://www.thequilt.net/wp-content/uploads/BBR-TCP-Opportunities.pdf

Parallel Streams

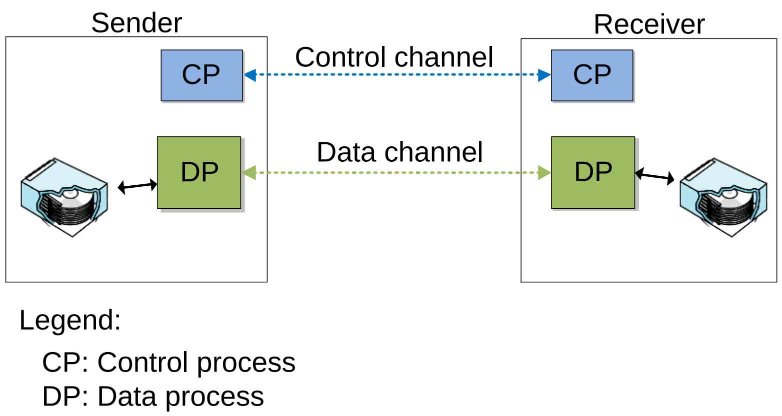
- Conventional file transfer protocols use a control channel and a (single) data channel (FTP model)



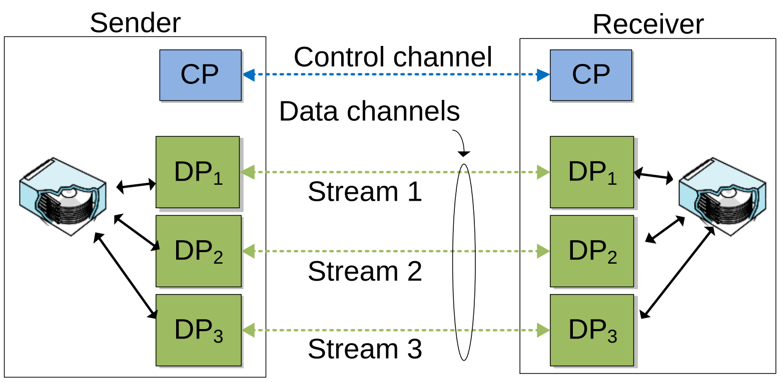
FTP model

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.

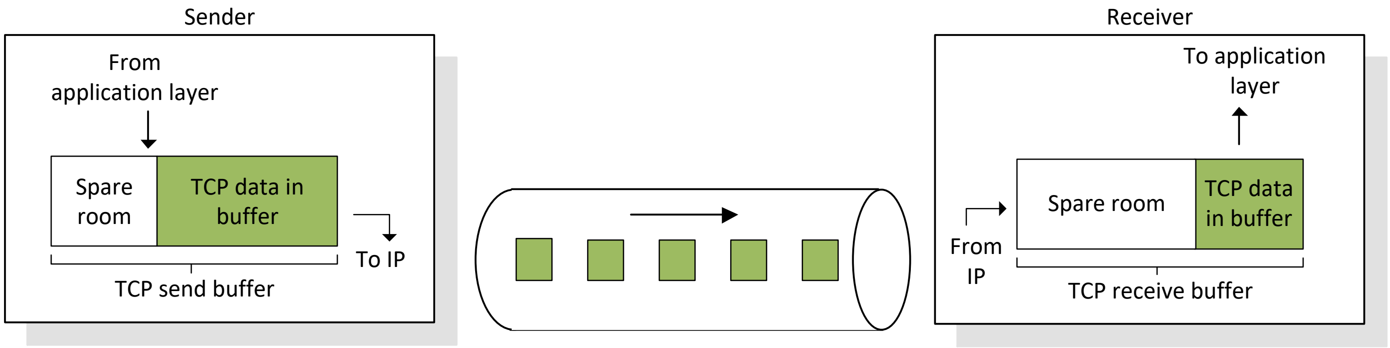
Advantages of Parallel Streams

- 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

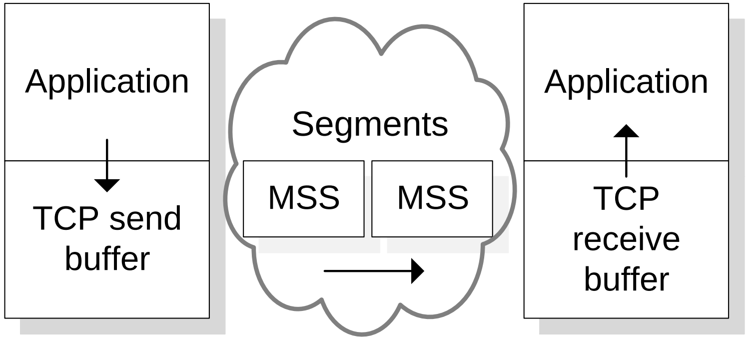
- Combat random packet loss not due congestion[1]
  - Parallel streams increase the recovery speed after the multiplicative decrease
- Mitigate TCP round-trip time (RTT) bias[2]
  - A low-RTT flow gets a higher share of the bandwidth than that of a high-RTT flow
  - Increase bandwidth allocated to big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

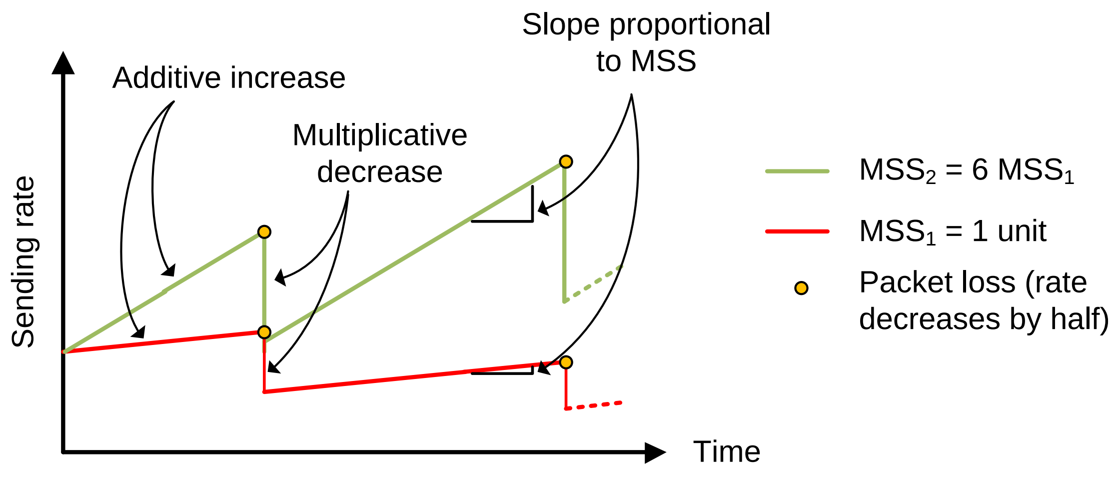
Maximum Segment Size (MSS)

- 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



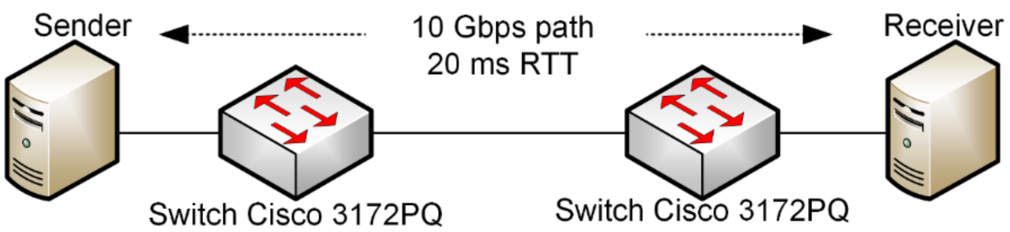
Advantages of Large MSS

- 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

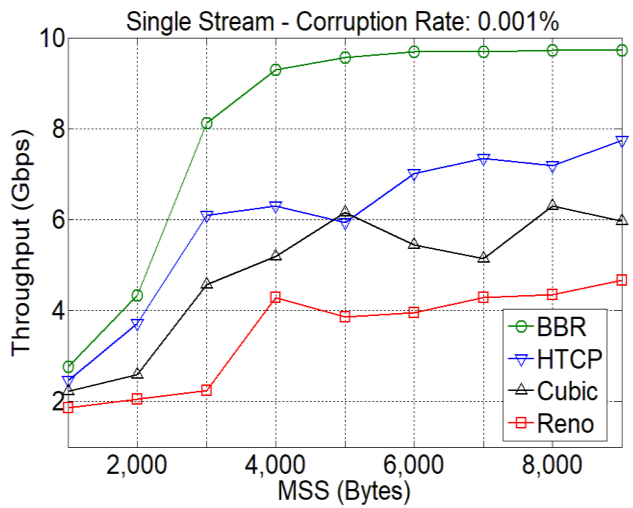
- 70-second experiments (first 10 seconds not considered)
- Ten experiments conducted and the average throughput is reported
- Impact of MSS and parallel streams on BBR, Reno, HTCP, Cubic



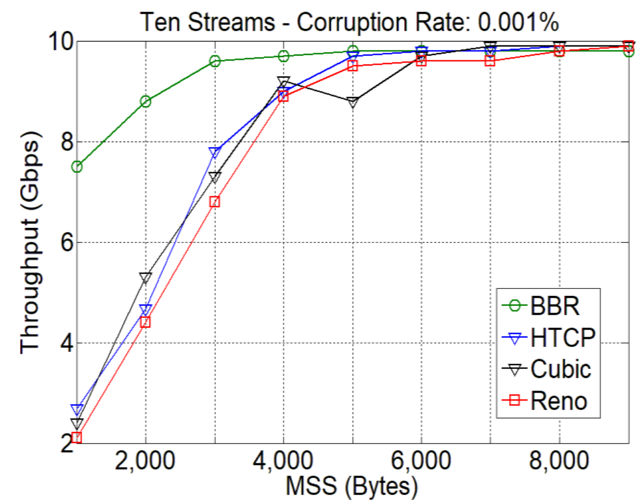
1. J. Crichigno, Z. Csibi, E. Bou-Harb, N. Ghani, "Impact of segment size and parallel streams on TCP BBR," IEEE Telecommunications and Signal Processing Conference (TSP), Athens, Greece, July 2018.

Results on a 10 Gbps Network

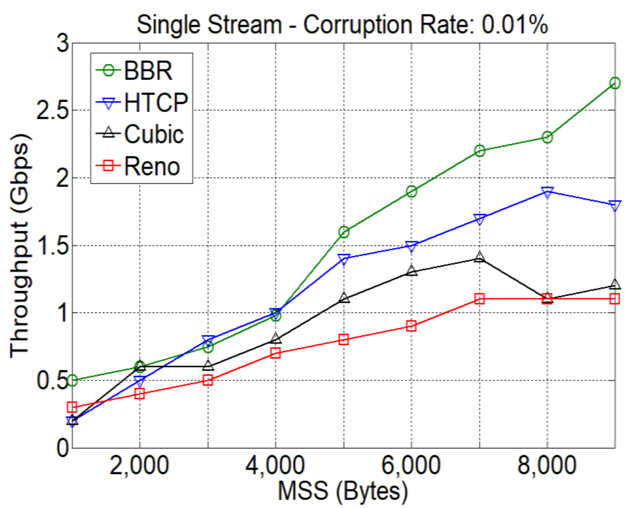
1



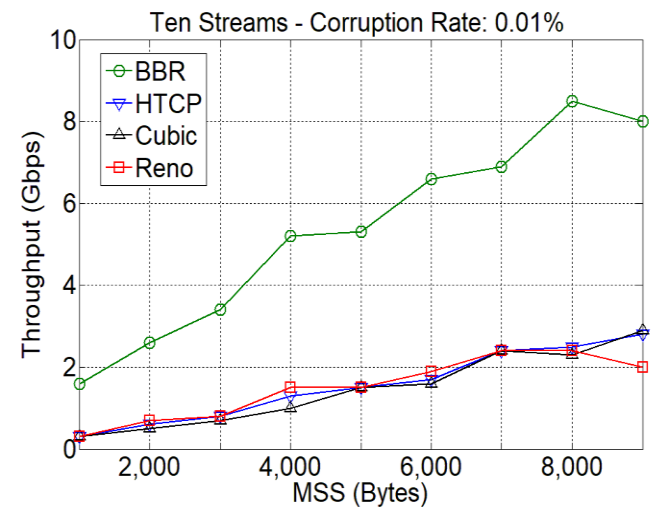
2



3

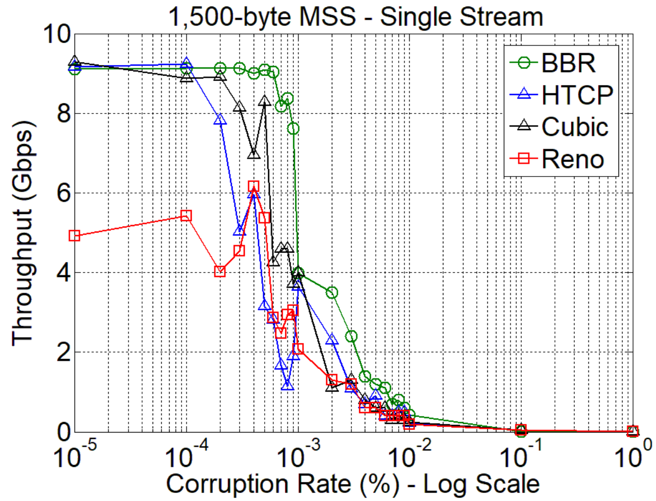


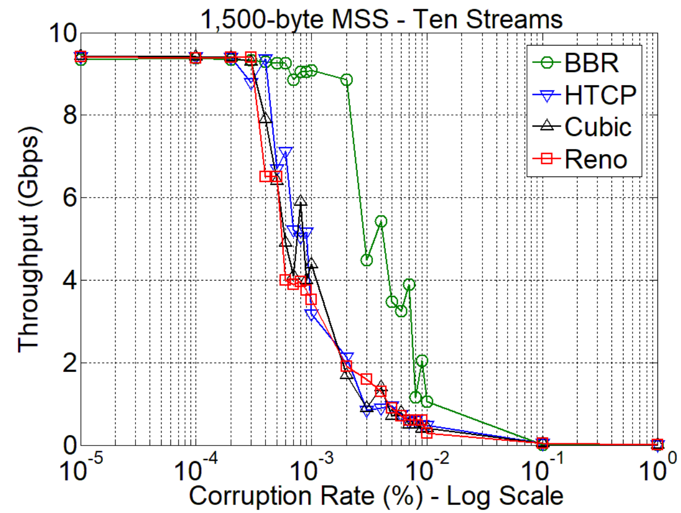
4

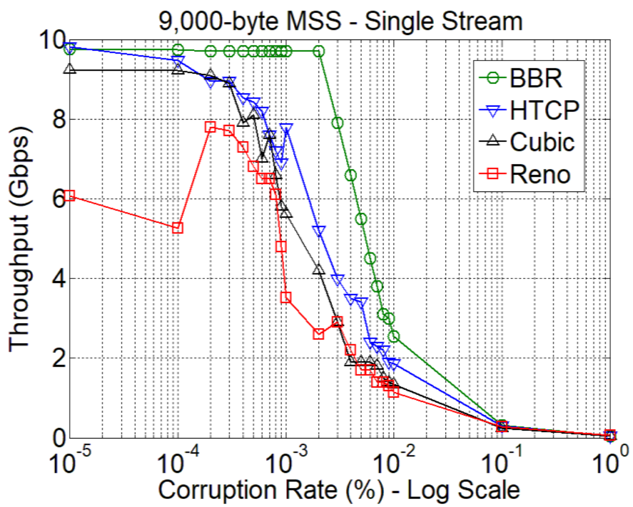


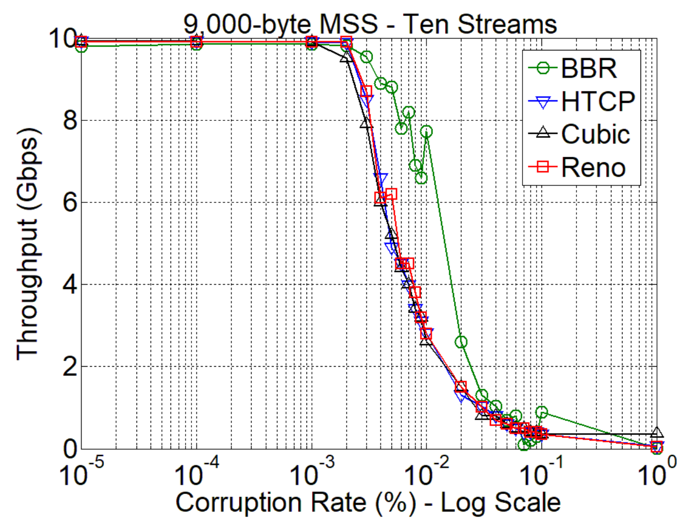
1. J. Crichigno, Z. Csibi, E. Bou-Harb, N. Ghani, "Impact of segment size and parallel streams on TCP BBR," IEEE Telecommunications and Signal Processing Conference (TSP), Athens, Greece, July 2018.

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.

DEMO
END-HOSTS TUNING IN HIGH SPEED NETWORKS

Demo activities are described in Lab 6, 8, 13 (“Network Tools and Protocols”)

Lab Information

https://netlab.cec.sc.edu/	**URL of the virtual lab platform** **Username: lastname (lowercase letters)** **Password: nsf2019**

Labs Series: Networks Tools and Protocols

- Lab 1: Introduction to Mininet
- Lab 2: Introduction to iPerf
- Lab 3: Emulating WAN with NETEM I Latency, Jitter
- Lab 4: Emulating WAN with NETEM II Packet Loss, Duplication, Reordering, and Corruption
- Lab 5: Setting WAN Bandwidth with Token Bucket Filter (TBF)
- Lab 6: Understanding Traditional TCP Congestion Control (HTCP, Cubic, Reno)
- Lab 7: Understanding Rate-based TCP Congestion Control (BBR)
- Lab 8: Bandwidth-delay Product and TCP Buffer Size
- Lab 9: Enhancing TCP Throughput with Parallel Streams
- Lab 10: Measuring TCP Fairness
- Lab 11: Router's Buffer Size
- Lab 12: TCP Rate Control with Pacing
- Lab 13: Impact of Maximum Segment Size on Throughput
- Lab 14: Router's Bufferbloat

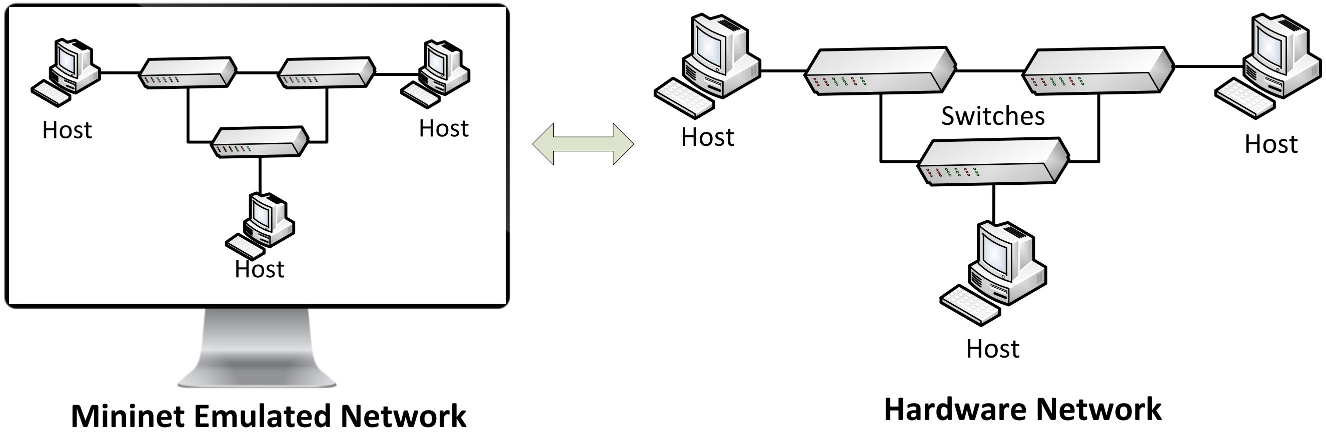
Organization of Lab Manuals

- Each lab starts with a section *Overview*
  - Objectives
  - Lab settings: passwords, device names
  - Roadmap: organization of the lab
- *Section 1*
  - Background information of the topic being covered (e.g., fundamentals of TCP congestion control)
  - Section 1 is optional (i.e., the reader can skip this section and move to lab directions)
- *Section 2… n*
  - Step-by-step directions

LAB 1: INTRODUCTION TO MININET

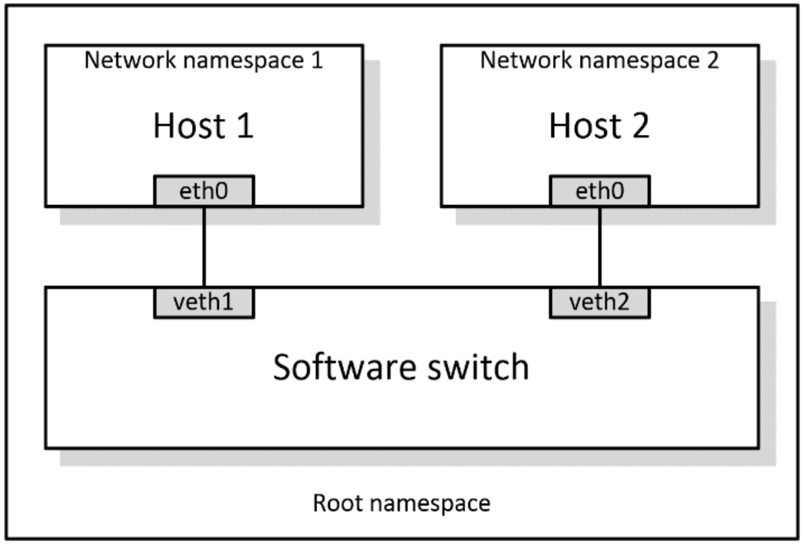
What is Mininet?

- A virtual testbed capable of recreating realistic scenarios
- It enables the development, testing of network protocols
- Inexpensive solution, real protocol stack, reasonably accurate

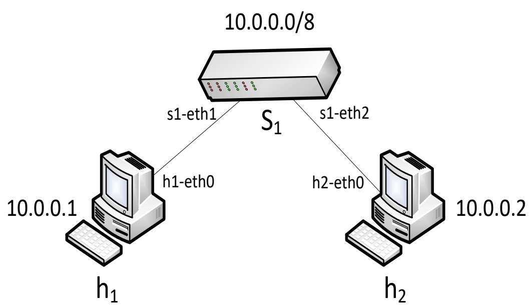


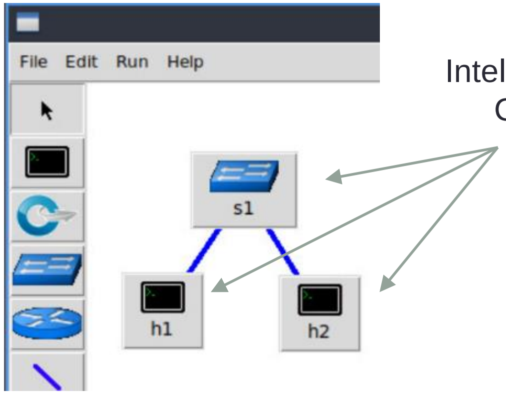
What is Mininet?

- Mininet nodes are network namespaces
  - Each node has different / separate virtual interface, routing tables
- Nodes use the underlying protocol stack of the host device
- Nodes are connected via virtual Ethernet (veth) links, which behave as Ethernet links

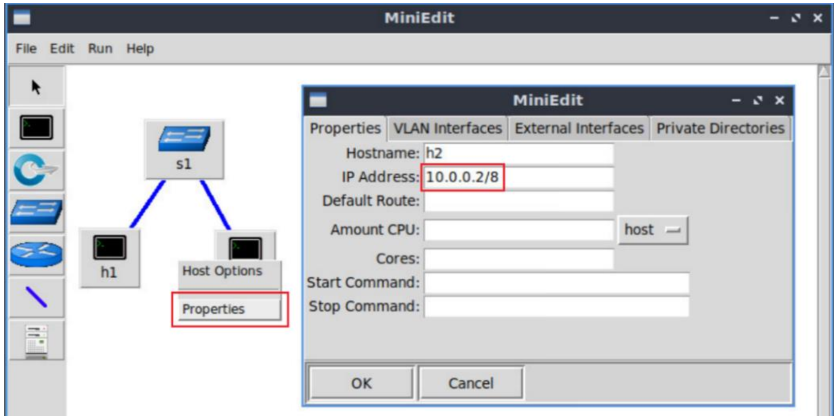


What is Mininet?





Intel Xeon Gold 6130 CPU, 2.1 GHz



sysctl

- *sysctl* is a tool for reading and modifying attributes of the system kernel
  - TCP buffer size (send and receive buffers)
  - Congestion control algorithm
  - IP forwarding and others

- Modify TCP read and write buffers

```
sysctl -w net.ipv4.tcp_rmem='10240 87380 52428800'
```

```
sysctl -w net.ipv4.tcp_wmem='10240 87380 52428800'
```

- Modify TCP congestion control algorithm

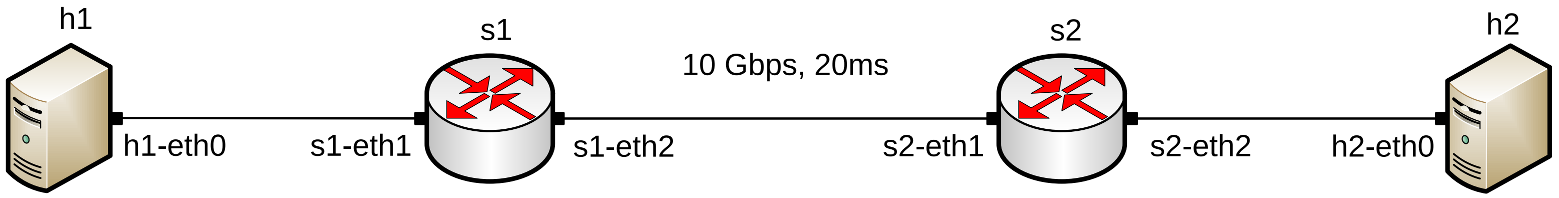
```
sysctl -w net.ipv4.tcp_congestion_control=bbr
```

- Check current values

```
sysctl net.ipv4.tcp_congestion_control
```

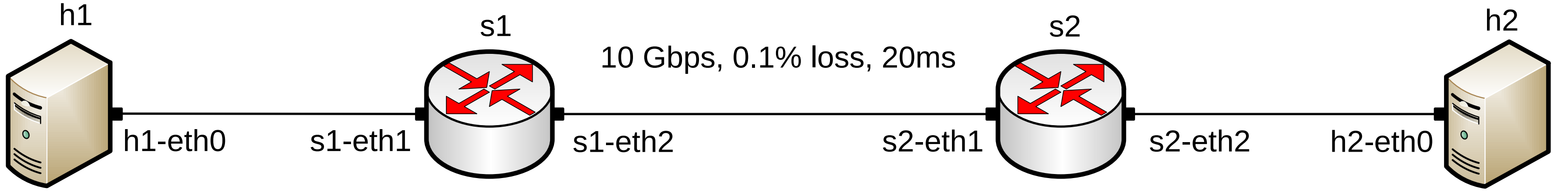
Experiment 1: TCP Buffer Size

- Lab 5 topology
- 10 Gbps, 20ms link s1-s2
- Measure throughput h1 > h2
- Modify TCP buffers at h1 and h2
  - Case 1: Small buffer size = 16,777,216 [bytes] (default in Linux)
  - Case 2: $2 \cdot BDP = 2 \cdot (10 \cdot 10^{9}) \cdot (20 \cdot 10^{-3})$ [bits] = 50,000,000 [bytes]



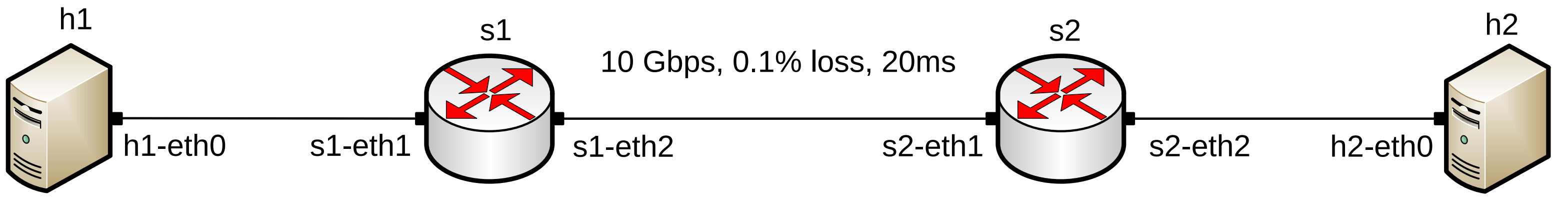
Experiment 2: TCP Congestion Control

- Lab 5 topology
- 10 Gbps, 0.1% loss, 20ms link s1-s2
- Measure throughput h1 > h2
  - Case 1: CUBIC as congestion control algorithm
  - Case 2: BBR as congestion control algorithm



Experiment 3: TCP Congestion Control

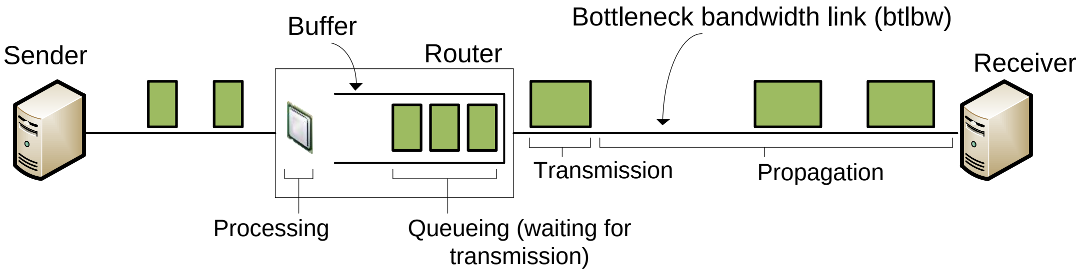
- Lab 5 topology
- 10 Gbps, 0.1% loss, 20ms. link s1-s2
- Increase buffer size to several BDPs (8 BDPs)
  - Buffer size = 200,000,000 [bytes]
- Measure throughput h1 > h2
  - Case 1: CUBIC as congestion control algorithm
  - Case 2: BBR as congestion control algorithm



LAB 14: ROUTER'S BUFFERBLOAT

Bufferbloat

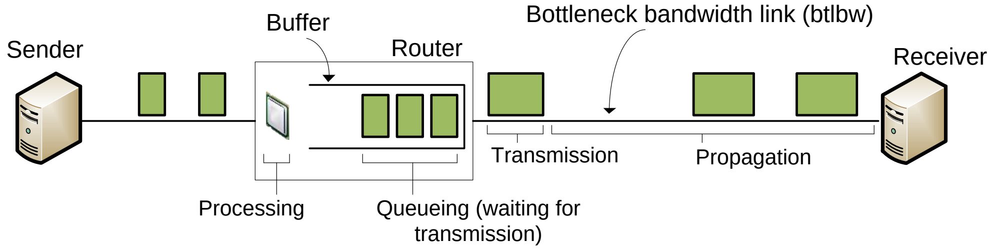
- Routers and switches must have enough memory allocated to hold packets momentarily (buffering)
- Rule of thumb:
  - Buffer size = RTT · bottleneck bandwidth[1, 2]



1. C. Villamizar, C. Song, "High performance TCP in ansnet," ACM Computer Communications Review, vol. 24, no. 5, pp. 45-60, Oct. 1994.
2. R. Bush, D. Meyer, "Some internet architectural guidelines and philosophy," Internet Request for Comments, RFC Editor, RFC 3439, Dec. 2003. [Online]. Available: https://www.ietf.org/rfc/rfc3439.txt.

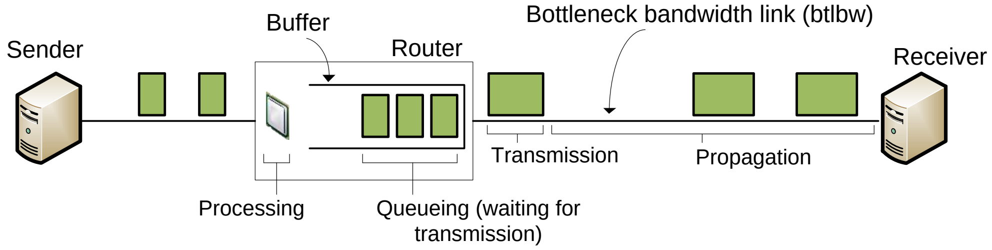
Bufferbloat

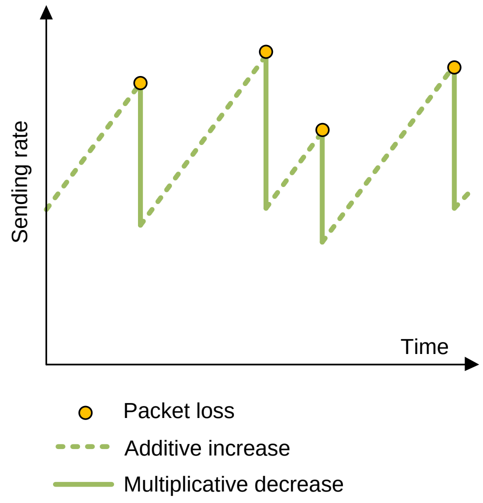
- Bufferbloat is a condition that occurs when the router buffers too much data, leading to excessive delays



Bufferbloat

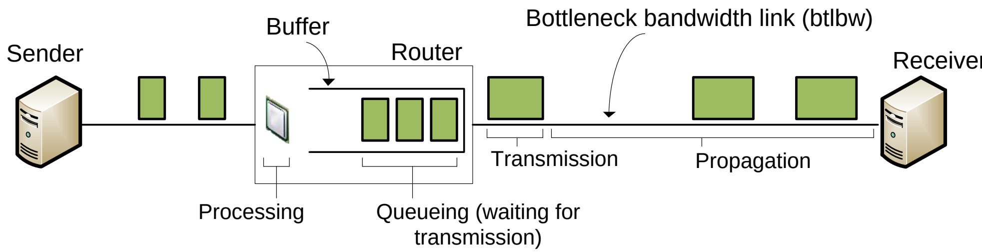
- Bufferbloat is a condition that occurs when the router buffers too much data, leading to excessive delays

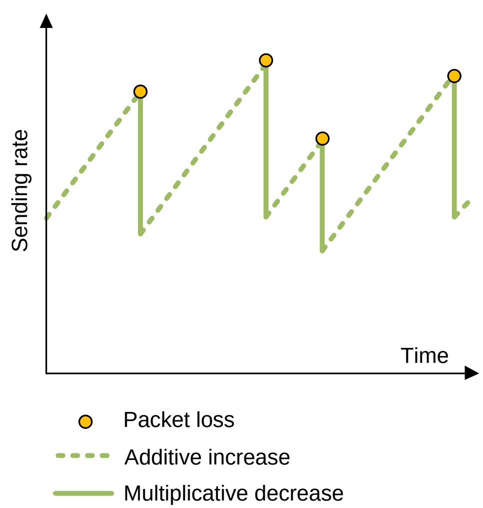


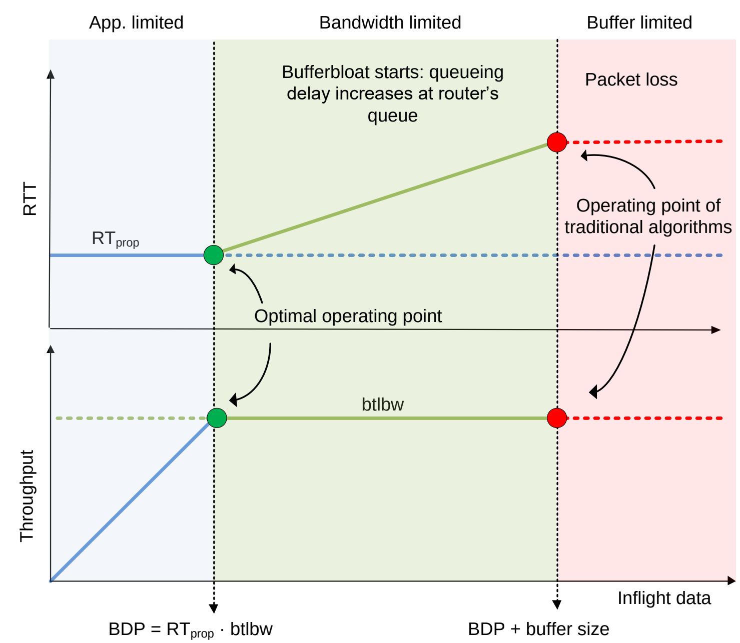


Bufferbloat

- Bufferbloat is a condition that occurs when the router buffers too much data, leading to excessive delays



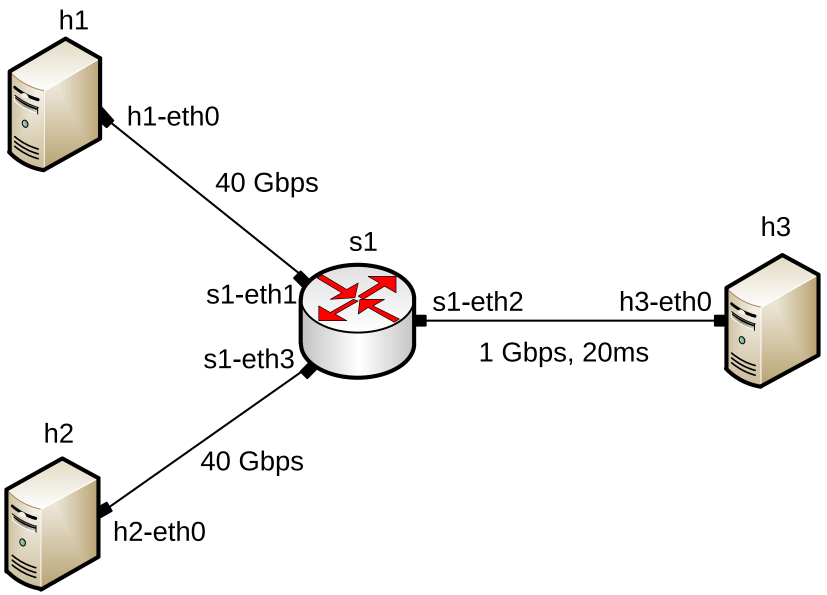




1. N. Cardwell, Y. Cheng, C. Gunn, S. Yeganeh, V. Jacobson, "BBR: congestion-based congestion control," *Communications of the ACM*, vol 60, no. 2, pp. 58-66, Feb. 2017.

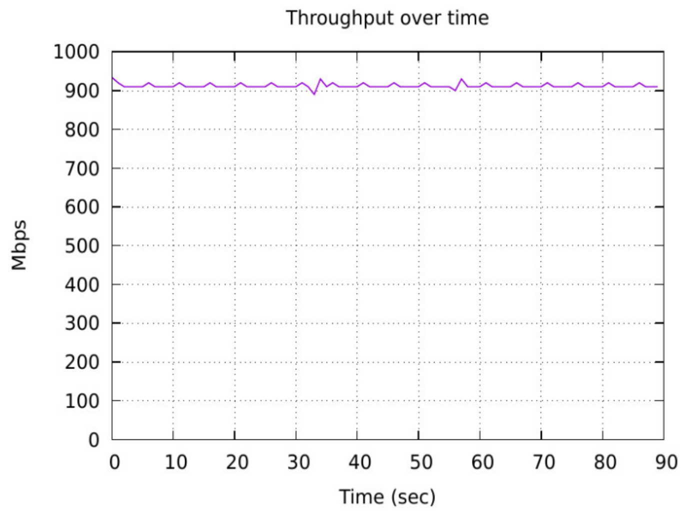
Bufferbloat

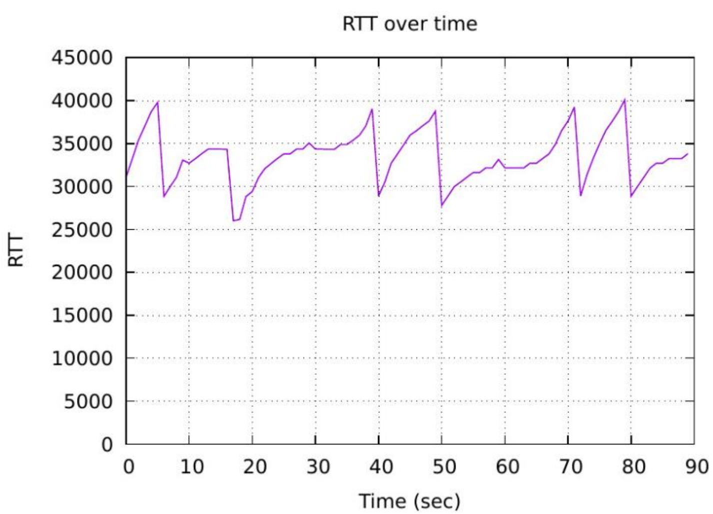
- Topology Lab 14
- 1 Gbps, 20ms link s1-h3
  - Measure RTT and throughput h1 > h3
  - Modify buffer size at s1 (interface s1-eth2)
    - Case 1: buffer size = $(1 \cdot 10^9) \cdot (20 \cdot 10^{-3})$ [bits] = 2,500,000 [bytes]
    - Case 2: buffer size = 25,000,000 [bytes]



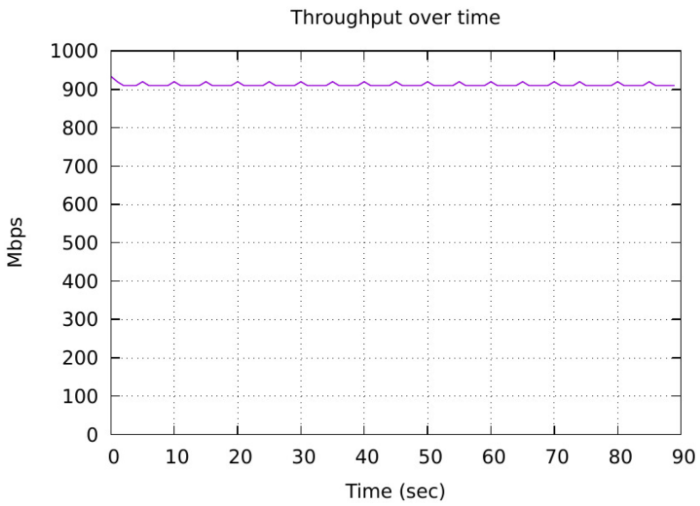
Bufferbloat

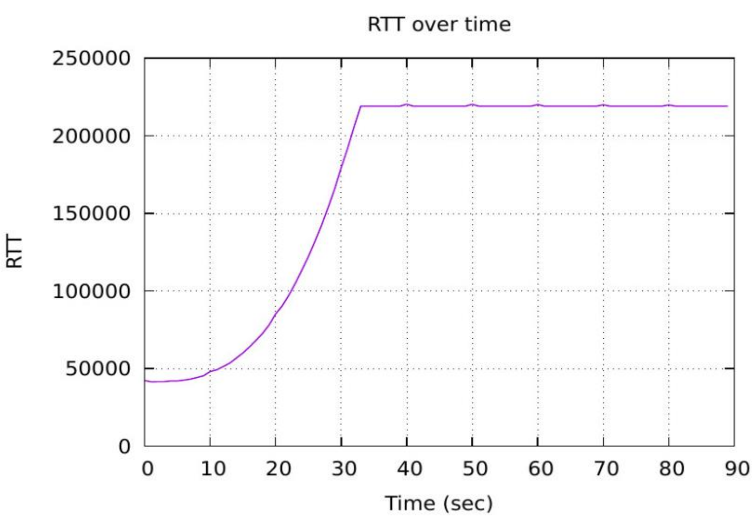
Buffer size = 1 BDP





Buffer size = 10 BDP





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