2023 Internet2 Technology Exchange

Science DMZs and Networking for All

Importance of TCP Congestion Control for Research and Education Data Transfers

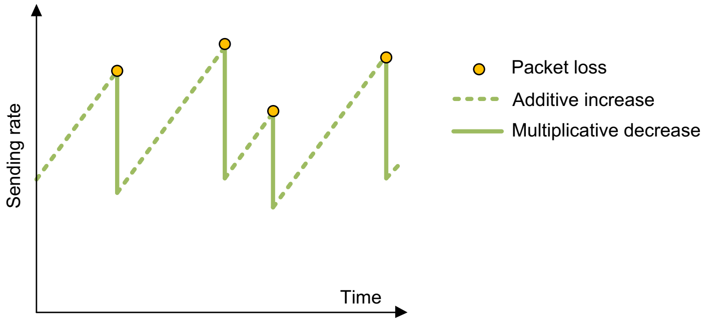
Jorge Crichigno, Elie Kfoury
University of South Carolina
https://research.cec.sc.edu/cyberinfra/

University of South Carolina (USC)
Energy Sciences Network (ESnet)

September 18, 2023

TCP Traditional Congestion Control

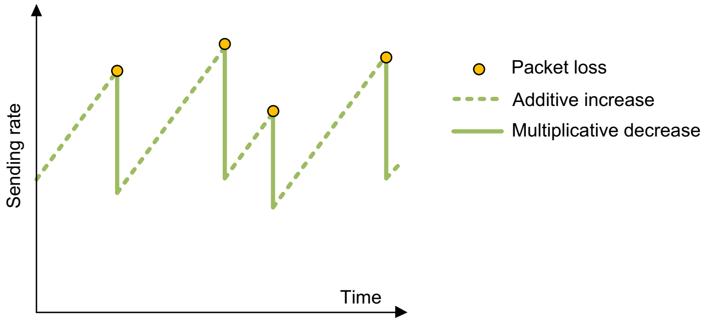
- The principles of window-based CC were described in the 1980s[1]
- Traditional CC algorithms follow the additive-increase multiplicative-decrease (AIMD) form of congestion control

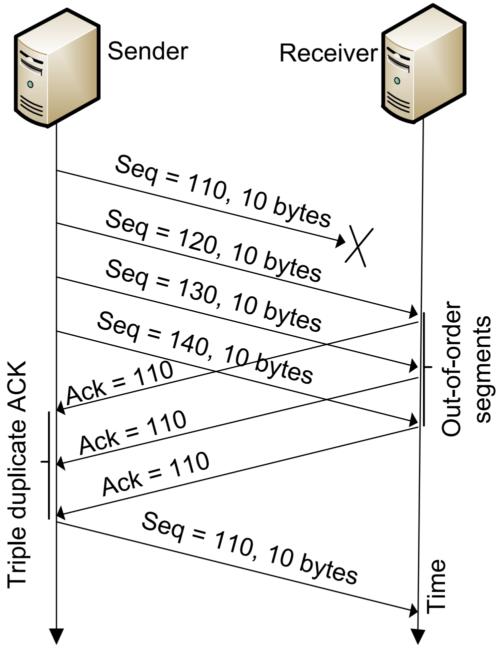


[1] V. Jacobson, M. Karels, Congestion avoidance and control, ACM SIGCOMM Computer Communication Review 18 (4) (1988).

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

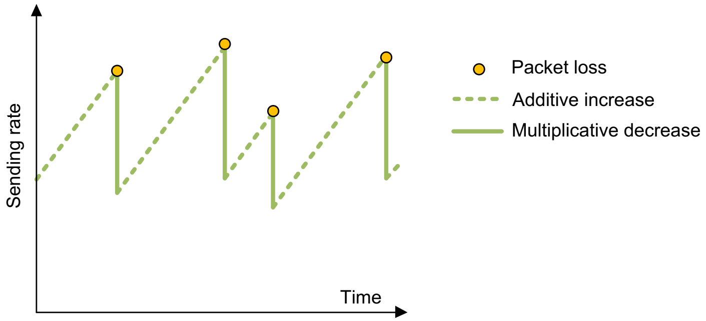


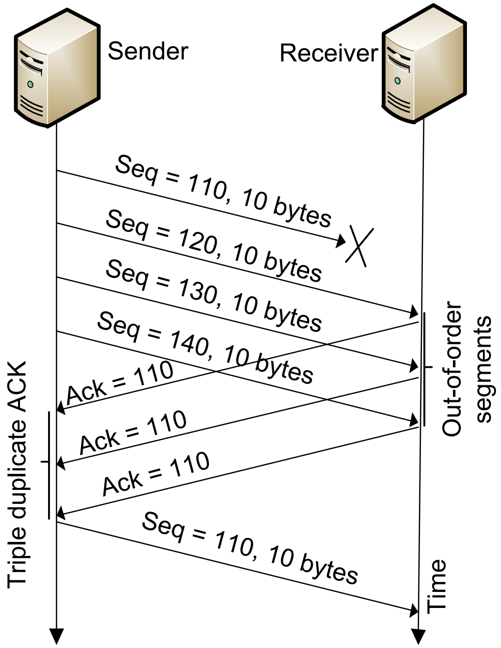


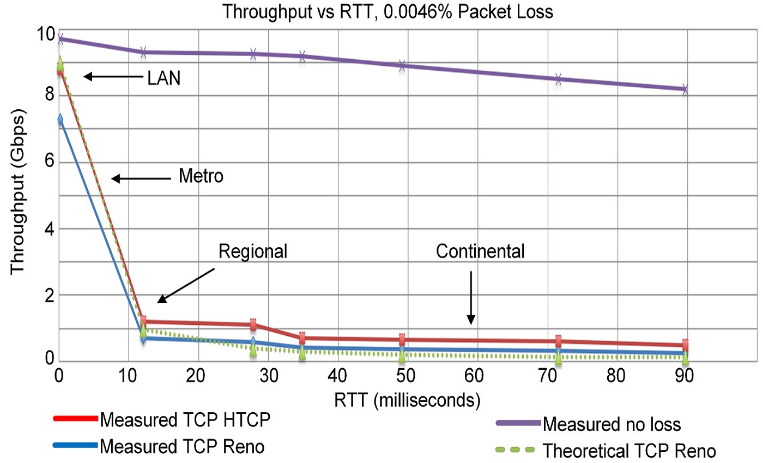
[1] V. Jacobson, M. Karels, Congestion avoidance and control, ACM SIGCOMM Computer Communication Review 18 (4) (1988).

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



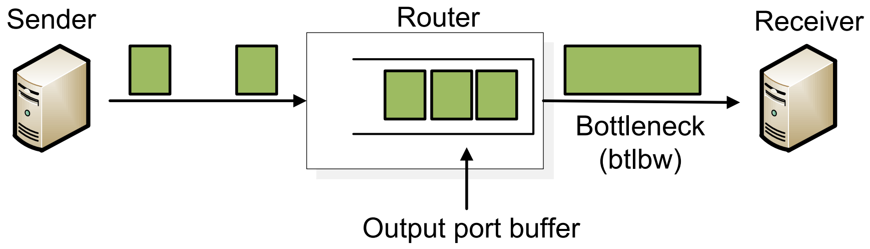




[1] V. Jacobson, M. Karels, Congestion avoidance and control, ACM SIGCOMM Computer Communication Review 18 (4) (1988).

BBR: Model-based CC

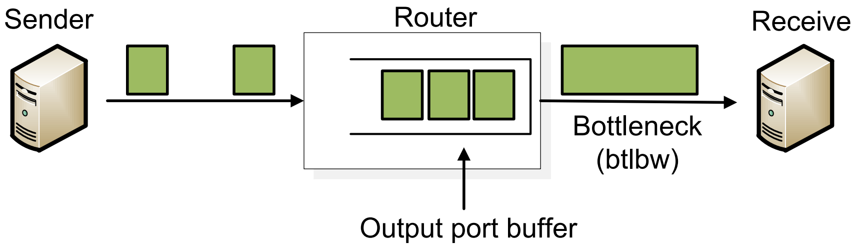
- TCP Bottleneck Bandwidth and RTT (BBR) is a rate-based congestion-control algorithm[1]
- BBR represented a disruption to the traditional CC algorithms:
  - is not governed by AIMD control law
  - does not use packet loss as a signal of congestion
- At any time, a TCP connection has one slowest link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

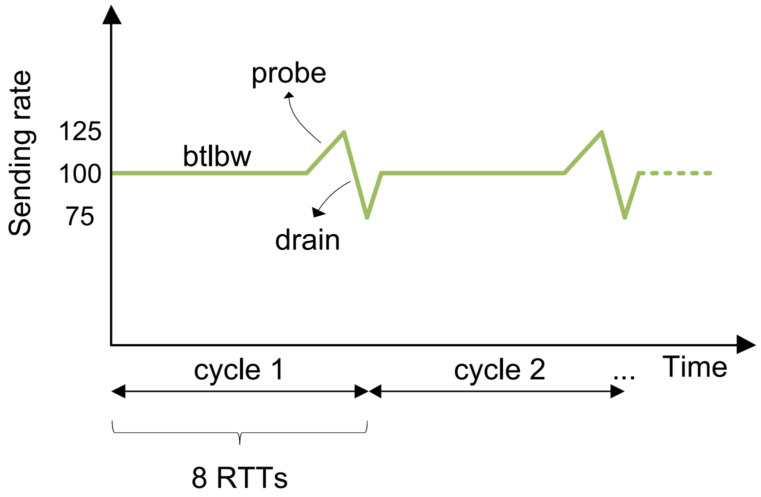


[1] N. Cardwell et al. "BBR v2, A Model-based Congestion Control." IETF 104, March 2019.

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)

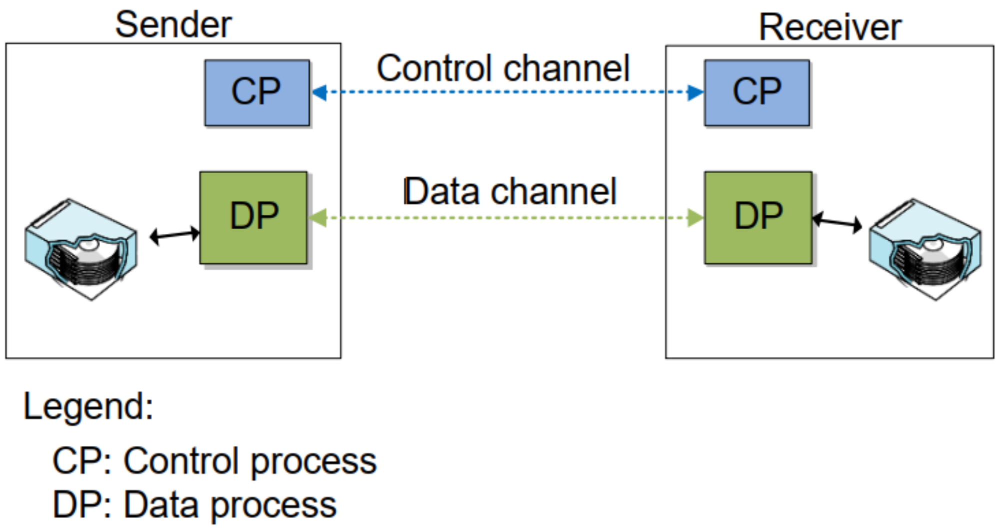




[1] N. Cardwell et al. "BBR v2, A Model-based Congestion Control." IETF 104, March 2019.

Parallel Streams

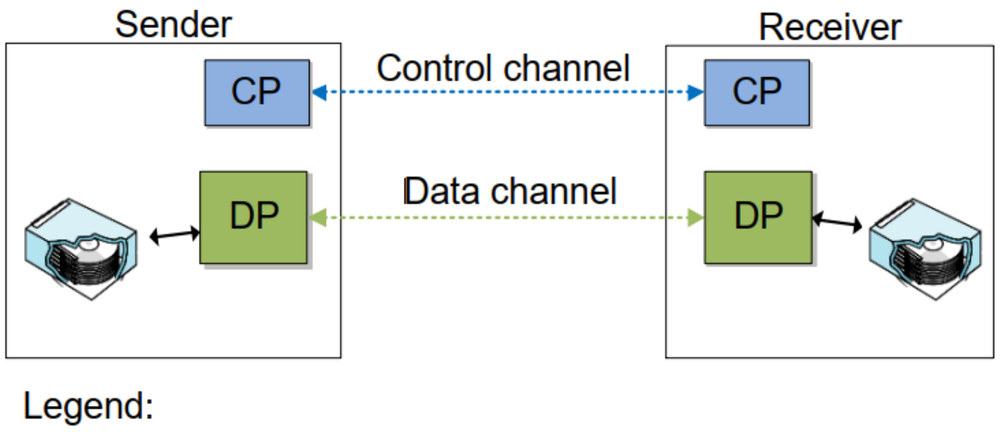
- Conventional file transfer protocols use a control channel and a (single) data channel (FTP model)



FTP model

Parallel Streams

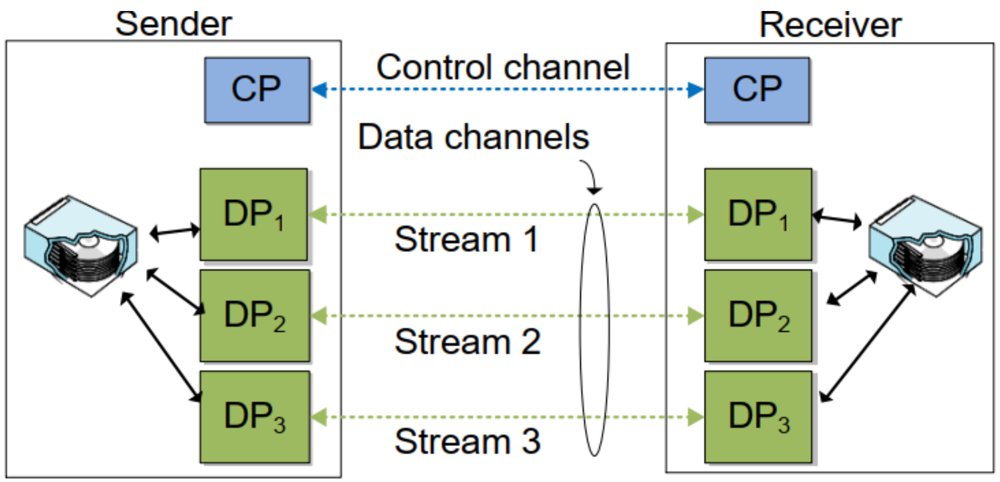
- Conventional file transfer protocols use a control channel and a (single) data channel (FTP model)
- gridFTP is an extension of the FTP protocol
- A feature of gridFTP is the use of parallel streams



Legend:

CP: Control process
DP: Data process

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

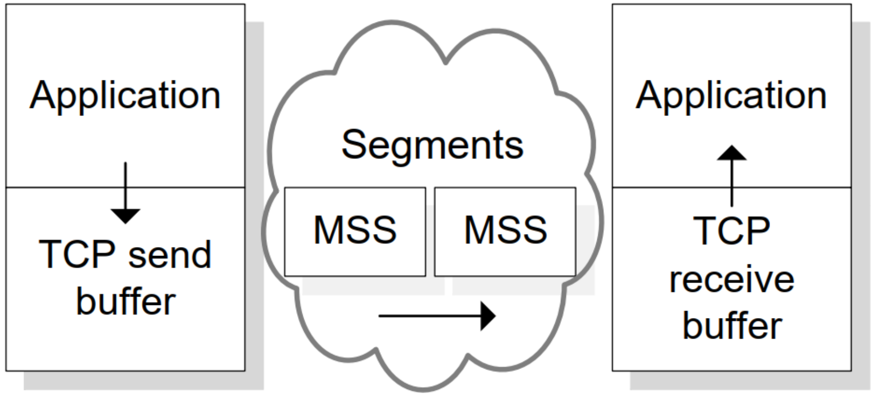
- Combat random packet loss not due congestion
  - Parallel streams increase the recovery speed after the multiplicative decrease
- Mitigate TCP round-trip time (RTT) bias
  - A low-RTT flow gets a higher share of the bandwidth than that of a high-RTT flow
  - Increase bandwidth allocated to big science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

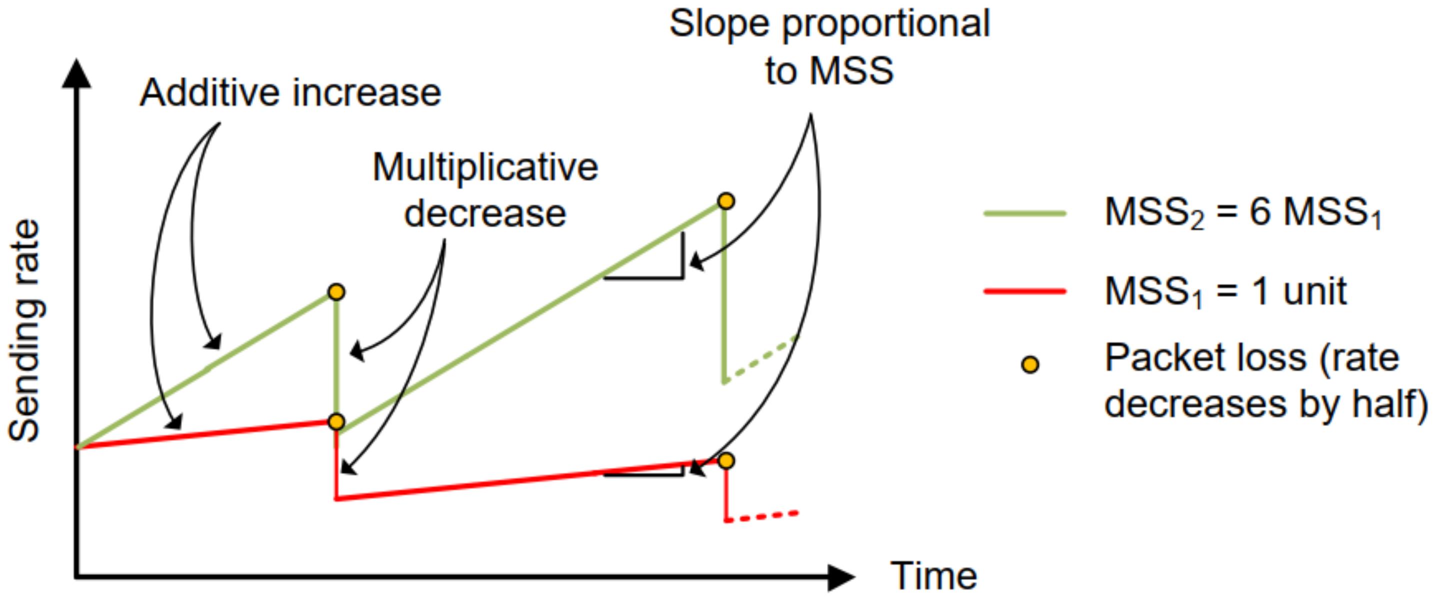
Maximum Segment Size (MSS)

- 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

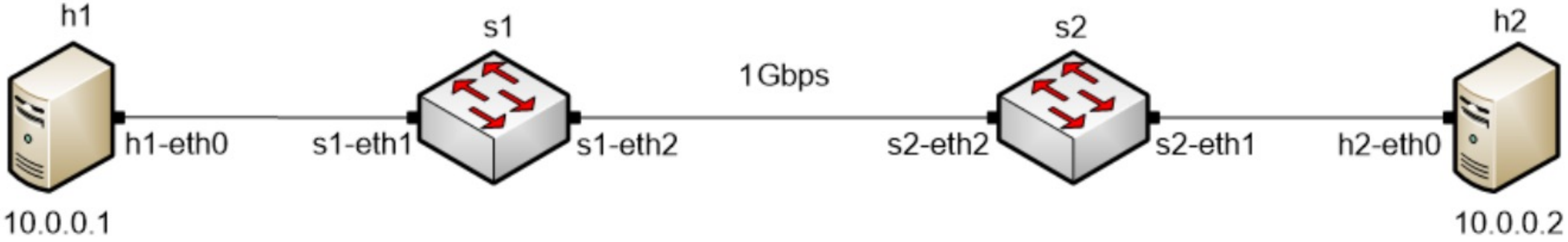


TCP Buffer Size

- 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:

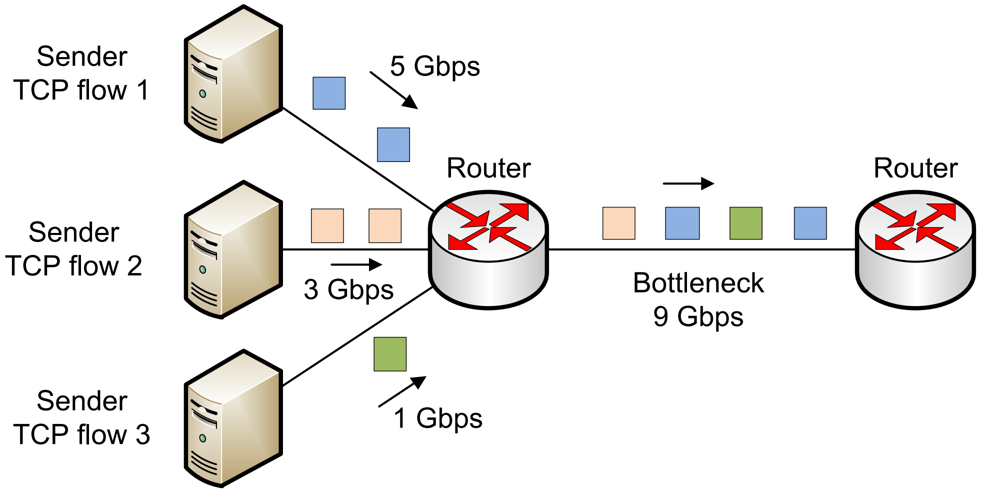
Traditional congestion controls: TCP buffer size ≥ 2BDP

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



Fairness

- Networks do not use bandwidth reservation mechanism for TCP flows
- Routers simply forward packets based on destination IP address
- The TCP congestion control algorithm ‘allocates’ bandwidth

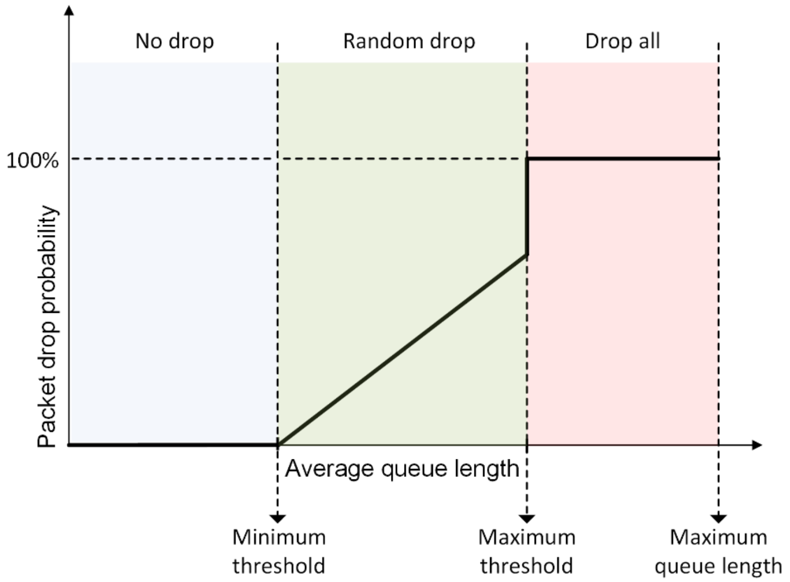


Active Queue Management (AQM)

- AQM encompasses a set of algorithms to reduce network congestion
- AQM algorithms try to prevent buffers from remaining full
- If the buffer is full, a packet must be dropped
  - A simple policy is Tail Drop: newly arriving packets are dropped until the queue has enough room to accept incoming traffic

Active Queue Management (AQM)

- AQM encompasses a set of algorithms to reduce network congestion
- AQM algorithms try to prevent buffers from remaining full
- If the buffer is full, a packet must be dropped
  - A simple policy is Tail Drop: newly arriving packets are dropped until the queue has enough room to accept incoming traffic
  - Random Early Detection: when the queue size is between min. and max. thresholds, drop with certain probability



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