Writing Fine-grained Measurements App with P4 Programmable Switches

Buffer and Queues

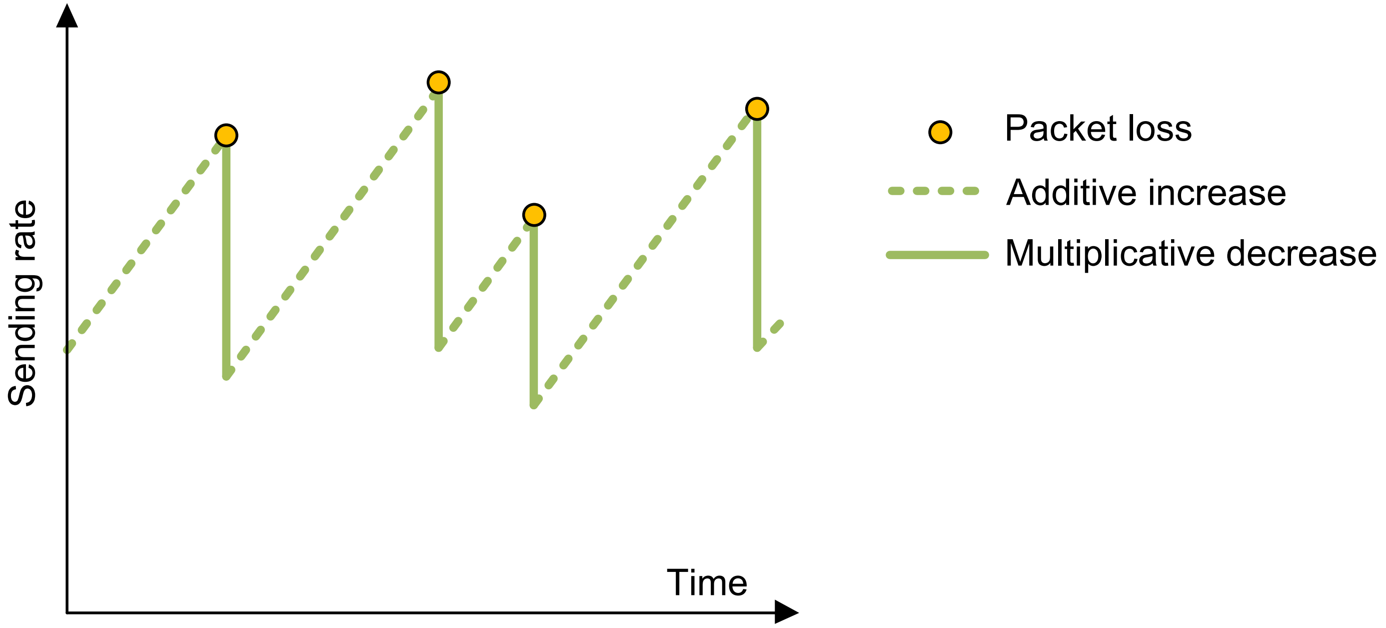
Jorge Crichigno
University of South Carolina
http://ce.sc.edu/cyberinfra

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

September 18, 2023

TCP Traditional Congestion Control

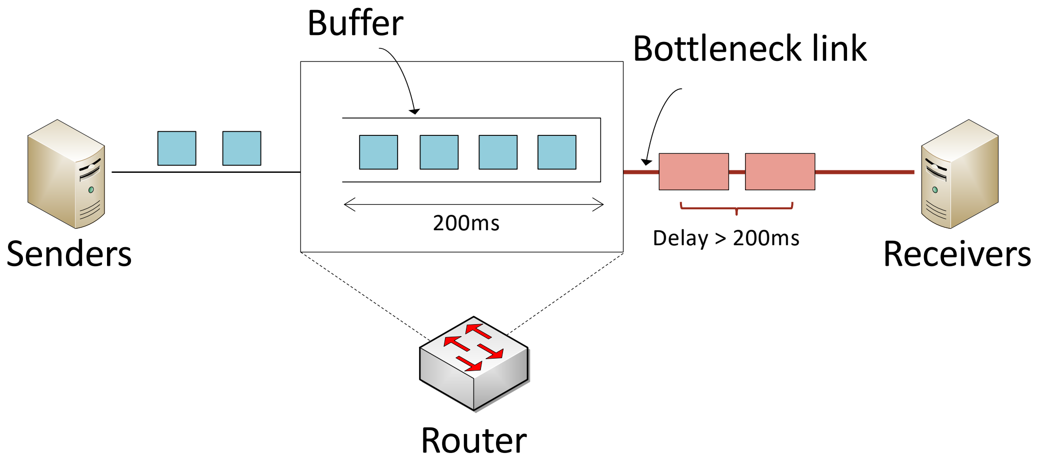
- The principles of window-based CC were described in the 1980s[1]
- Traditional CC algorithms follow the additive-increase multiplicative-decrease (AIMD) form of congestion control



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

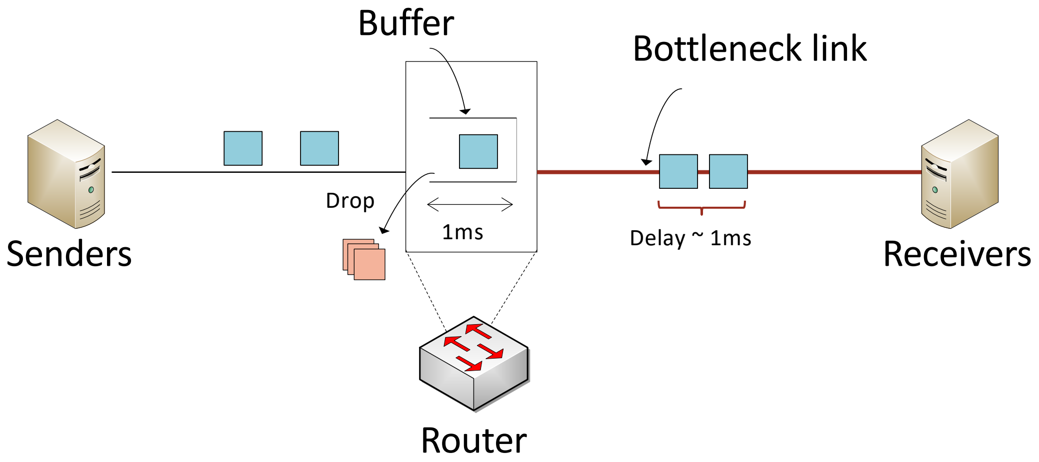
Buffer Size Problem

- Routers and switches have a memory referred to as packet buffer
- The size of the buffer impacts the network performance
  - Large buffers → TCP keeps the buffer full → excessive delays, Bufferbloat



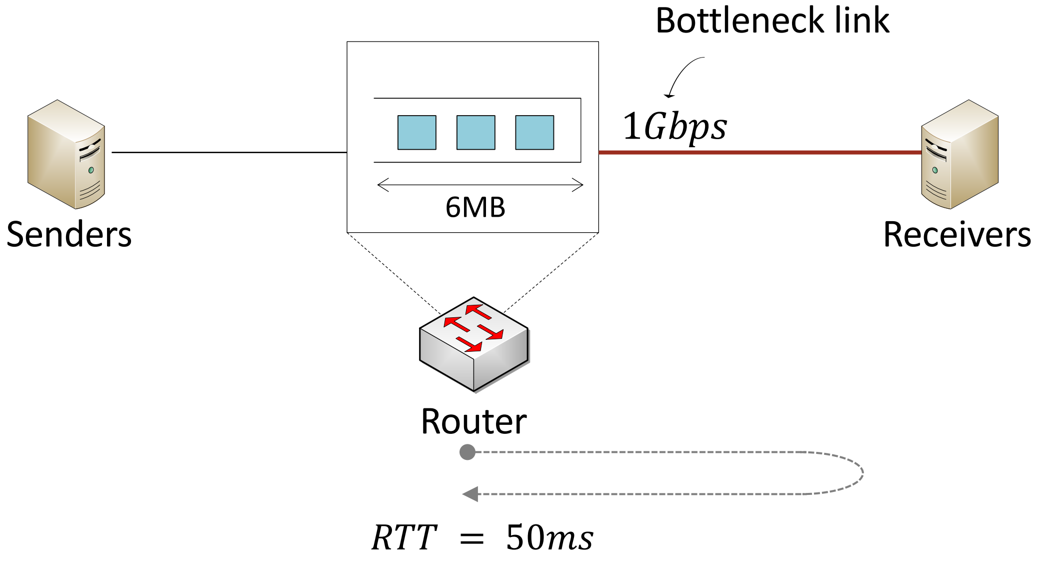
Buffer Size Problem

- Routers and switches have a memory referred to as packet buffer
- The size of the buffer impacts the network performance
  - Large buffers → TCP keeps the buffer full → excessive delays, Bufferbloat
  - Small buffers → packet drops → sender slows down → low link utilization



Buffer Sizing Rules: BDP

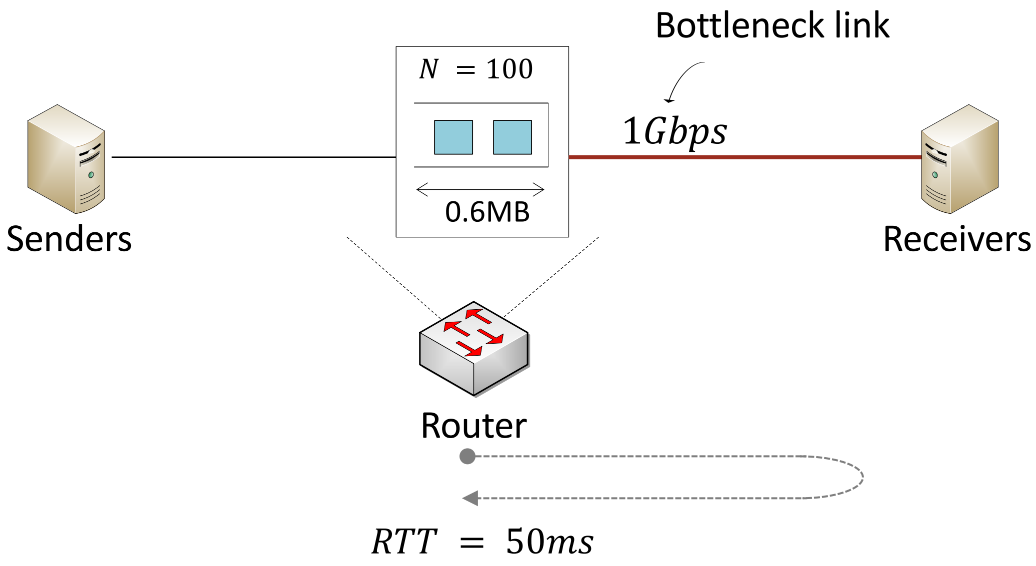
- General rule-of-thumb[1]: bandwidth-delay product (older rule)
  - $B = C \cdot RTT$
  - C is the capacity of the link and RTT is the average round-trip time
- Example: $C = 1Gbps$ with $RTT = 50ms$ → $B = 6\text{MB}$



1. C. Villamizar and C. Song. High performance TCP in ansnet. ACM Computer Communications Review, 24(5):45–60, 1994 199

Buffer Sizing Rules: Stanford

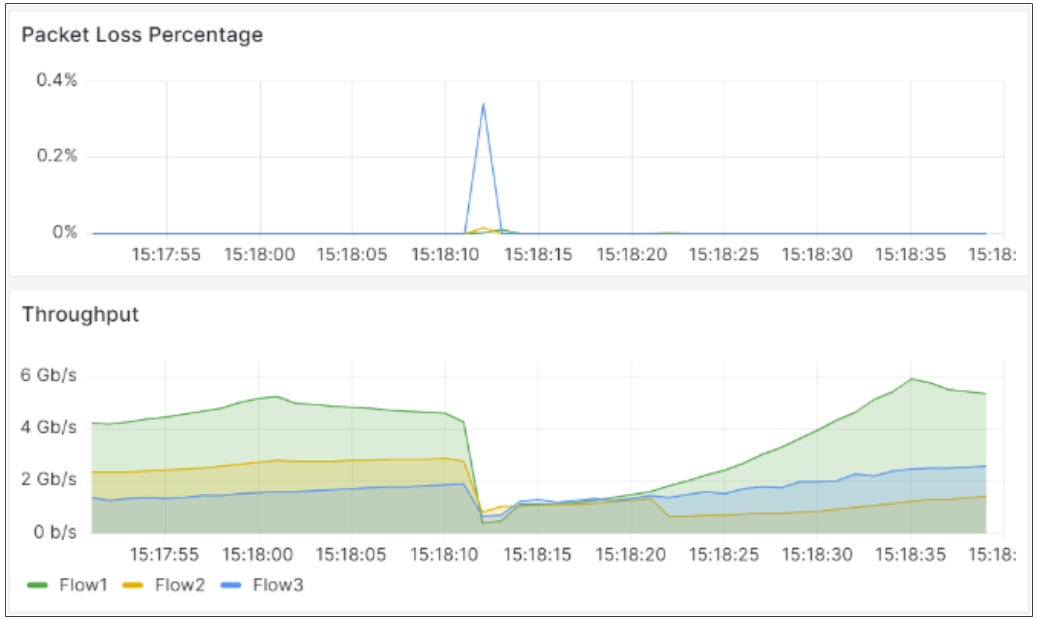
- Stanford rule[1]: smaller buffers are enough to get full link utilization
  - $B = \frac{C * RTT}{\sqrt{N}}$
  - N is the number of long (persistent over time) flows traversing the link
- Example: $C = 1Gbps$ with $RTT = 50ms$ and 100 flows → $B = 0.6\text{MB}$



1. Appenzeller, Guido, Isaac Keslassy, and Nick McKeown. "Sizing router buffers." *ACM SIGCOMM Computer Communication Review* 34.4 (2004)

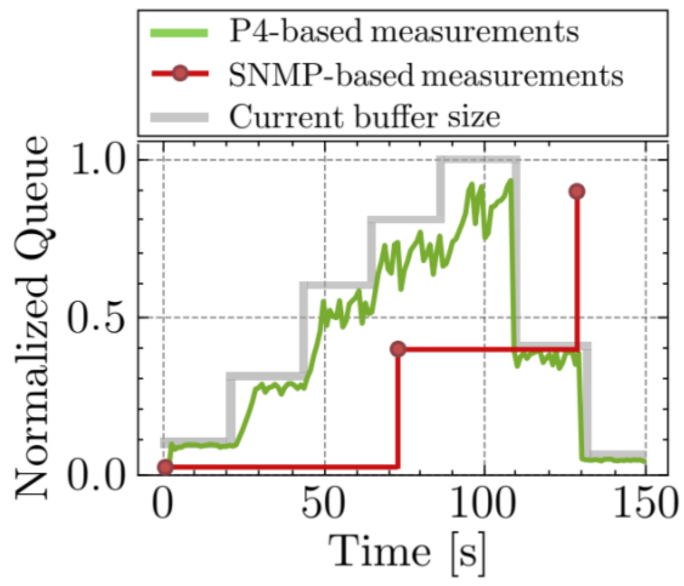
Why Queue Measurement?

- Queue measurement is important for troubleshooting purposes
- For example, it helps detecting *microbursts*
- Microbursts are rapid bursts sent in quick succession, leading to buffer overflow
- Microbursts are detrimental for the performance of TCP in a high latency high bandwidth setting



Queue Measurement in P4

- Traditionally, protocols like SNMP are used to measure the queue
- SNMP produce inaccurate and stale results
- On a Juniper MX-204 router, SNMP produced a queue occupancy sample every 70 seconds
- Programmable data planes produce a queue occupancy sample per packet



(a)