COMP 431
Internet Services & Protocols

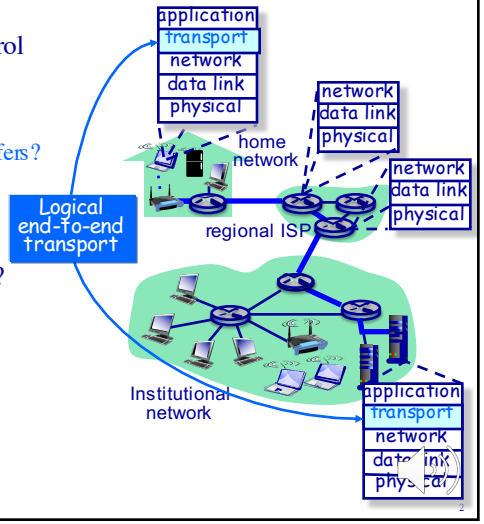
The Transport Layer
TCP Fairness & Performance

Jasleen Kaur

March 26, 2020

Transport Layer Protocols & Services
Performance issues

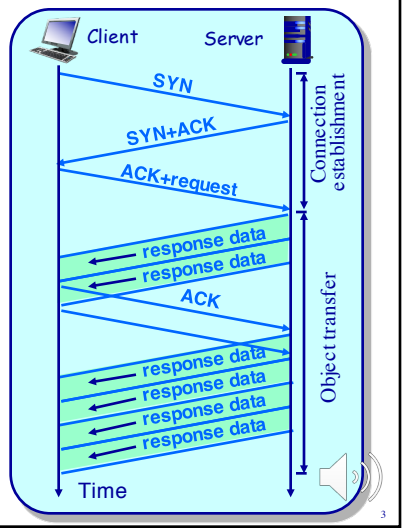
- How does congestion control impact the latency of TCP transfers?
  - » Does TCP give good performance for Web transfers?
- What throughputs are attainable under TCP's congestion control scheme?
  - » What is the impact of slow-start/AIMD congestion control on throughput?



TCP Performance

Latency for TCP transfers

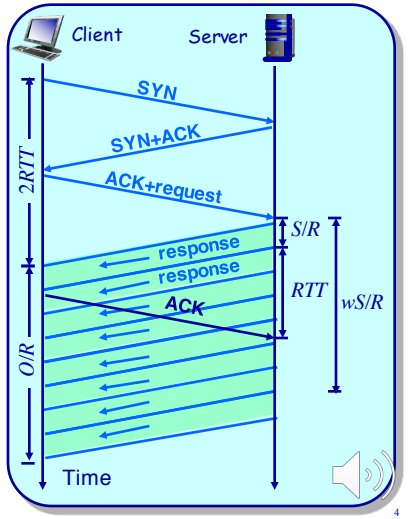
- How long does it take to receive an object from a Web server?
  - » TCP connection establishment overhead
  - » Data transfer delay
    - ❖ Slow start
    - ❖ Congestion avoidance
- Assume one link between client and server with transmission speed *R*
  - » Fixed congestion window of *w* segments
  - » *S = segment size*
  - » *O = object size*
  - » No loss/retransmissions
  - » SYN & ACK transmission times negligible



TCP Latency Analysis

Fixed-sized congestion window of *w* segments

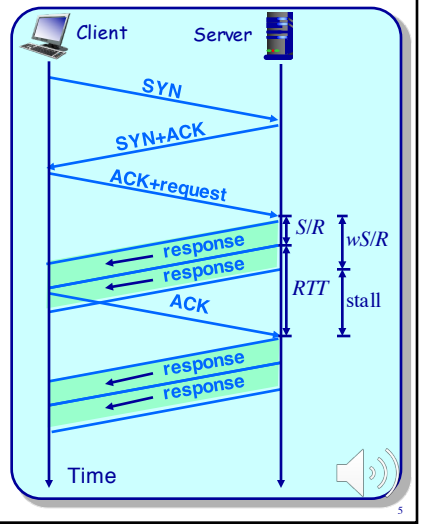
- Assume no congestion or flow control
  - » Receiver has large window
- Case 1: $RTT + S/R < wS/R$
  - » Here ACKs return before the server completes the transmission of a window
  - » (Hence the size of the congestion window does not effect performance)
- $latency = 2RTT + O/R$
  - » Is it possible to do better?



TCP Latency Analysis

Fixed-sized congestion window of *w* segments

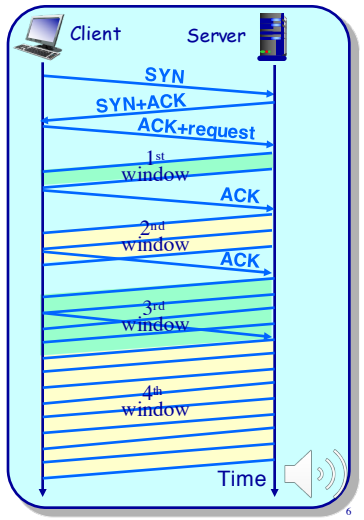
- Case 2: $RTT + S/R > wS/R$
  - » Now the server "stalls" waiting for ACKs to return
  - » $latency = 2RTT + O/R + \text{total stall time}$
  - » $\text{stall time per window} = (S/R + RTT) - wS/R$
- The object requires $k = \lceil O/wS \rceil$ windows to transmit
  - » $\text{total stall time} = (k-1)(S/R + RTT - wS/R)$



TCP Latency Analysis

Adding in the effect of slow-start

- Assume the window grows according to slow start
  - » k = Number of windows required to transmit the object
  - » q = Number of windows transmitted before the server no longer stalls
  - » Actual number of stalls is $P = \text{MIN}\{q, k-1\}$
- $latency = min\_latency + \text{stall time}$
  - » $min\_latency = 2RTT + O/R$
  - » $\text{stall time} = P(...)$



Know for final

TCP Latency Analysis

Adding in the effect of slow-start

- *min_latency* $= 2RTT + O/R$
- *stall time for the* i^{th} *window* =

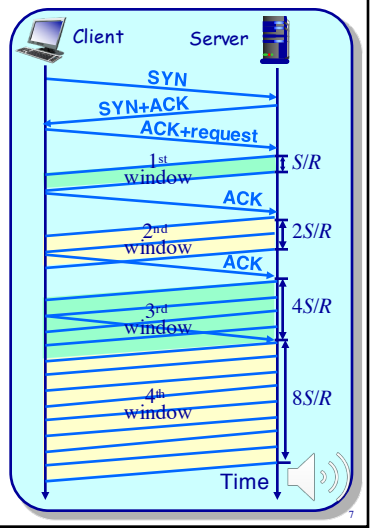
$$= 0 \text{ or } \underbrace{\text{time for the return of the ACK for the first segment in the } i^{th} \text{ window}}_{} - \underbrace{\text{time to transmit the } i^{th} \text{ window}}_{}$$

$$= 0 \text{ or } \left(\frac{S}{R} + RTT\right) - 2^{i-1}\frac{S}{R}$$

- *total stall time* = …

$$= \sum_{i=1}^{P}\left(\frac{S}{R} + RTT - 2^{i-1}\frac{S}{R}\right)$$

$$= P\left(\frac{S}{R} + RTT\right) - (2^{P}-1)\frac{S}{R}$$



TCP Latency Analysis

Summary

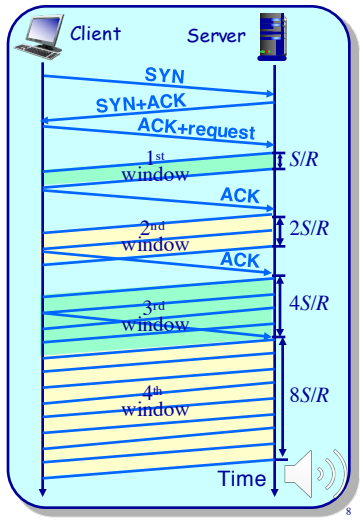
- Latency with a fixed-size window w:
  - » $2RTT + O/R +$

$$(k-1)\left(\frac{S}{R} + RTT - w\frac{S}{R}\right)$$

- Latency under slow start:
  - » $2RTT + O/R +$

$$P\left(\frac{S}{R} + RTT\right) - (2^{P}-1)\frac{S}{R}$$

k = Number of windows required to transmit the object
P= Number of TCP server stalls
= MIN$\{q, k-1\}$



TCP Latency Analysis

Finding q and k

- k is the number of windows required to transmit the object

$$k = \text{MIN}\left\{i:\ 2^0+2^1+2^2+\ldots+2^{i-1} \geq \frac{O}{S}\right\}$$

$$= \text{MIN}\left\{i:\ 2^i-1 \geq \frac{O}{S}\right\}$$

$$= \text{MIN}\left\{i:\ i \geq log_2\left[\frac{O}{S}+1\right]\right\}$$

$$= log_2\left[\frac{O}{S}+1\right]$$

- q is the number of windows required for the sender to fully consume the bandwidth on the link

$$q = \text{MAX}\left\{i:\ RTT + \frac{S}{R} \geq 2^{i-1}\frac{S}{R}\right\}$$

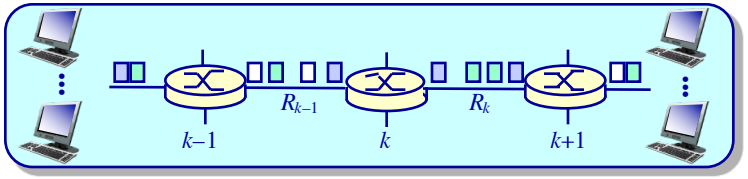
$$= \text{MAX}\left\{i:\ 2^{i-1} \leq 1 + \frac{RTT}{S/R}\right\}$$

$$= \text{MAX}\left\{i:\ i \leq log_2\left[1 + \frac{RTT}{S/R}\right]+1\right\}$$

$$= log_2\left[1 + \frac{RTT}{S/R}\right]+1$$

Congestion Control

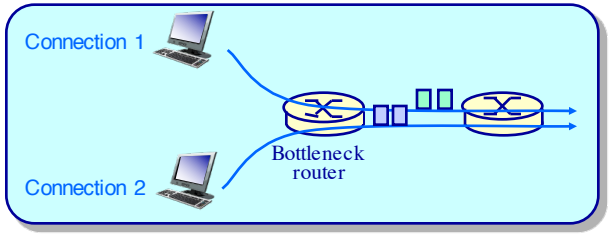
How to Define Fairness?



- If n_k connections share a congested link k with capacity R_k, each connection should receive $r = R_k/n_k$ bandwidth
- But what if a connection can't consume R/n bandwidth?
  - » MAX-MIN fairness:
    - ❖ If a connection receives less bandwidth than it requires, then it receives the same amount of bandwidth as all other unsatisfied connection

TCP Performance

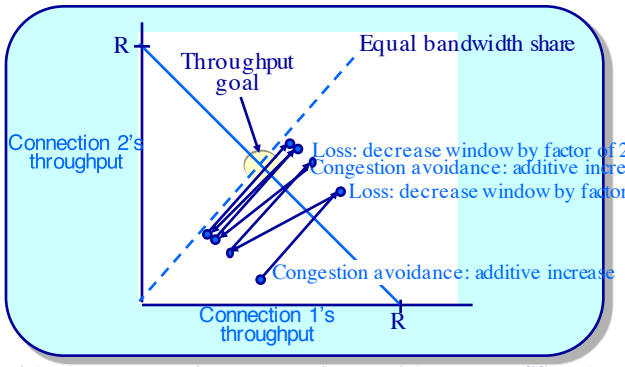
Is TCP throughput fairly realized?



- Simple fairness
  - » If *n* TCP sessions share a bottleneck link, each should get 1/*n* of link capacity
- When a connection slows down, by how much should it slow down?

TCP Throughput

Is TCP fair?



- Consider two competing connections with same *MSS* and *RTT*
  - » Additive increase gives slope of 1, as throughout increases
  - » Multiplicative decrease decreases throughput proportionally

Transport Layer Protocols & Services

Summary

- Fundamental transport layer services
  - » Multiplexing/Demultiplexing
  - » Error detection
  - » Reliable data delivery
  - » Flow control
  - » Congestion control
- Internet transport protocols
  - » UDP
  - » TCP
- Up next: Leaving the network "edge" and diving into the network "core"