

Delay modeling

Q: How long does it take to receive an object from a Web server after sending a request?

Ignoring congestion, delay is influenced by:

- r TCP connection establishment
- r data transmission delay
- r slow start

Notation, assumptions:

- r Assume one link between client and server of rate R
- r S : MSS (bits)
- r O : object size (bits)
- r no retransmissions (no loss, no corruption)

Window size:

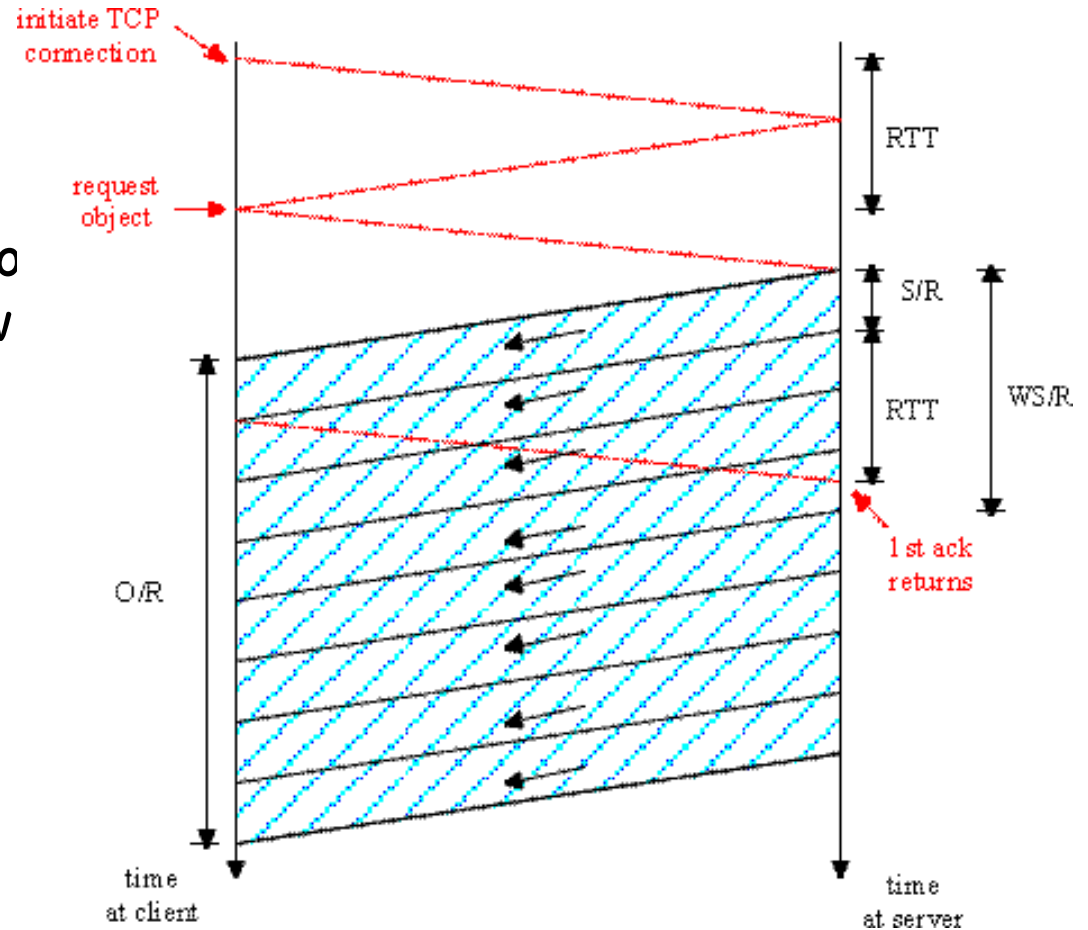
- r First assume: fixed congestion window, W segments
- r Then dynamic window, modeling slow start

Fixed congestion window (1)

First case:

$WS/R > RTT + S/R$: ACK for first segment in window returns before window's worth of data sent

$$\text{delay} = 2RTT + O/R$$

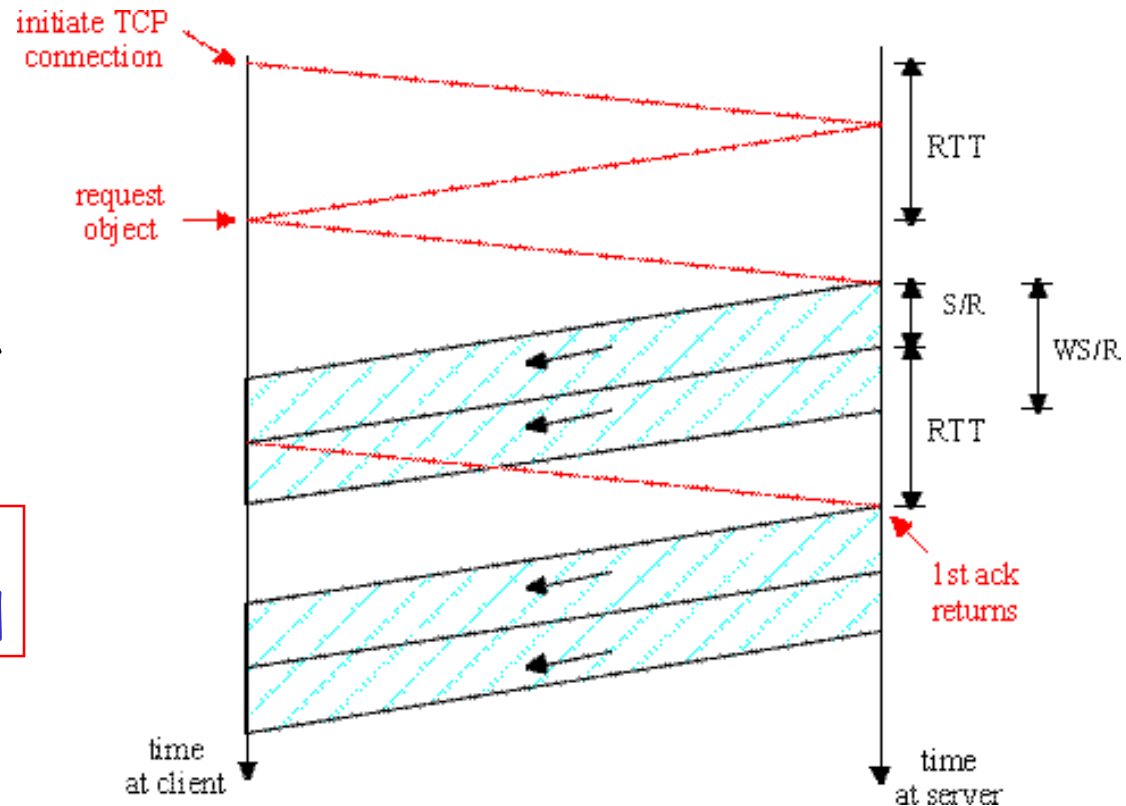


Fixed congestion window (2)

Second case:

- $WS/R < RTT + S/R$: wait for ACK after sending window's worth of data sent

$$\text{delay} = 2RTT + O/R + (K-1)[S/R + RTT - WS/R]$$



TCP Delay Modeling: Slow Start (1)

Now suppose window grows according to slow start

Will show that the delay for one object is:

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

where P is the number of times TCP idles at server:

$$P = \min\{Q, K - 1\}$$

- where Q is the number of times the server idles if the object were of infinite size.
- and K is the number of windows that cover the object.

TCP Delay Modeling: Slow Start (2)

Delay components:

- 2 RTT for connection estab and request
- O/R to transmit object
- time server idles due to slow start

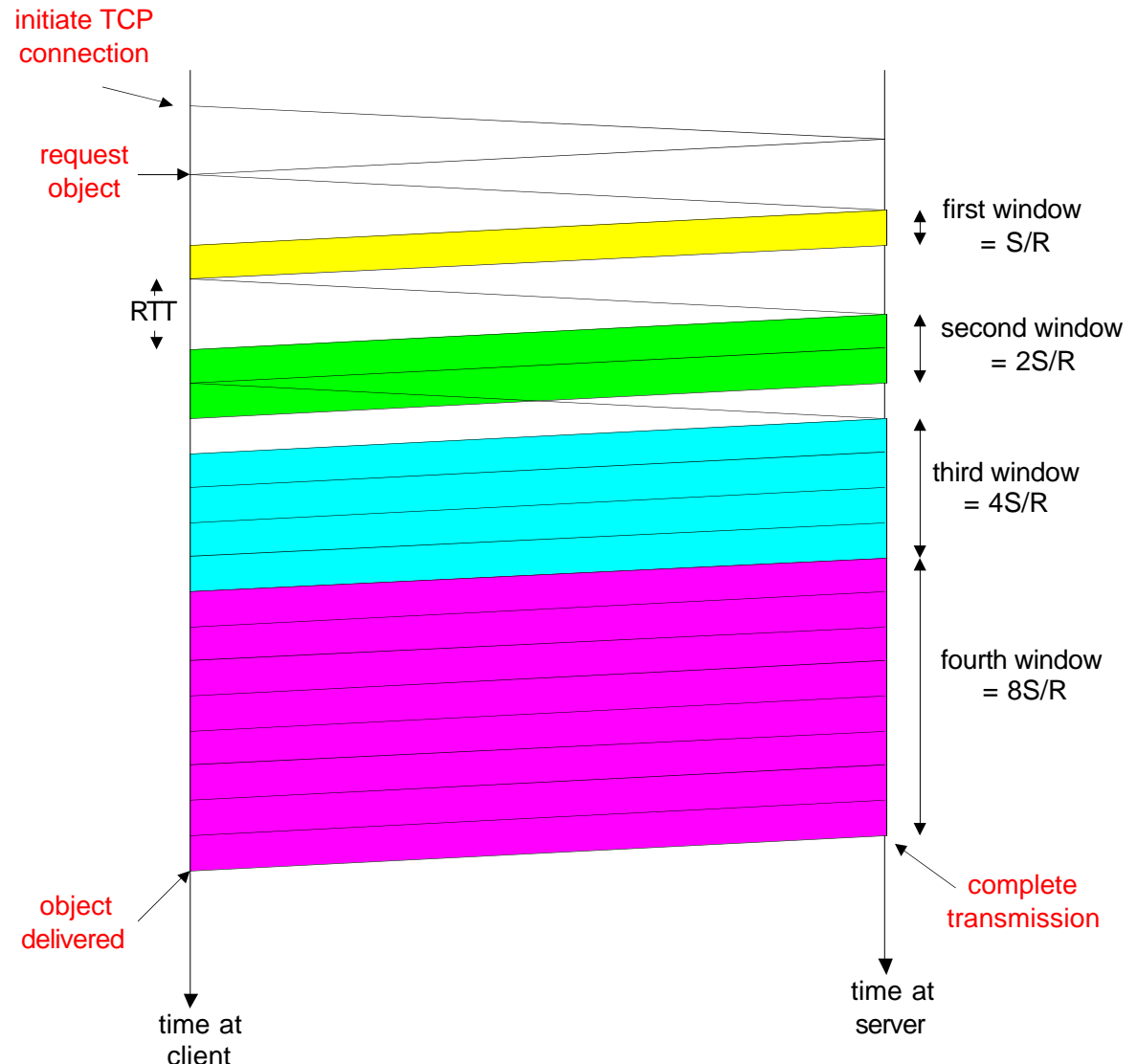
Server idles:

$$P = \min\{K-1, Q\} \text{ times}$$

Example:

- O/S = 15 segments
- K = 4 windows
- Q = 2
- $P = \min\{K-1, Q\} = 2$

Server idles $P=2$ times



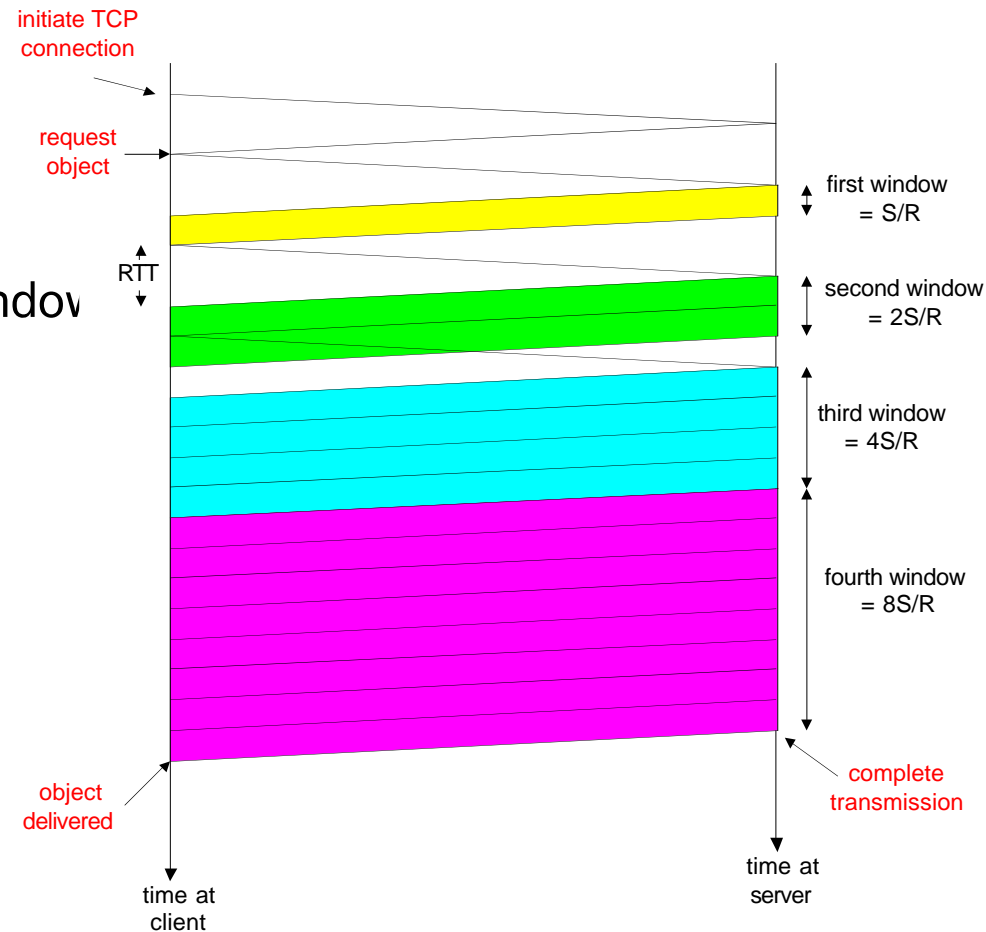
TCP Delay Modeling (3)

$\frac{S}{R} + RTT$ = time from when server starts to send segment
until server receives acknowledgment

$2^{k-1} \frac{S}{R}$ = time to transmit the kth window

$\left[\frac{S}{R} + RTT - 2^{k-1} \frac{S}{R} \right]^+$ = idle time after the kth window

$$\begin{aligned} \text{delay} &= \frac{O}{R} + 2RTT + \sum_{p=1}^P \text{idleTime}_p \\ &= \frac{O}{R} + 2RTT + \sum_{k=1}^P \left[\frac{S}{R} + RTT - 2^{k-1} \frac{S}{R} \right] \\ &= \frac{O}{R} + 2RTT + P \left[RTT + \frac{S}{R} \right] - (2^P - 1) \frac{S}{R} \end{aligned}$$



TCP Delay Modeling (4)

Recall K = number of windows that cover object

How do we calculate K ?

$$\begin{aligned} K &= \min\{k: 2^0 S + 2^1 S + L + 2^{k-1} S \geq O\} \\ &= \min\{k: 2^0 + 2^1 + L/S + 2^{k-1} \geq O/S\} \\ &= \min\{k: 2^k - 1 \geq \frac{O}{S}\} \\ &= \min\{k: k \geq \log_2(\frac{O}{S} + 1)\} \\ &= \left\lceil \log_2(\frac{O}{S} + 1) \right\rceil \end{aligned}$$

Calculation of Q , number of idles for infinite-size object, is similar (see HW).

HTTP Modeling

- r Assume Web page consists of:
 - m 1 base HTML page (of size O bits)
 - m M images (each of size O bits)
- r Non-persistent HTTP:
 - m $M+1$ TCP connections in series
 - m Response time = $(M+1)O/R + (M+1)2RTT + \text{sum of idle times}$
- r Persistent HTTP:
 - m 2 RTT to request and receive base HTML file
 - m 1 RTT to request and receive M images
 - m Response time = $(M+1)O/R + 3RTT + \text{sum of idle times}$
- r Non-persistent HTTP with X parallel connections
 - m Suppose M/X integer.
 - m 1 TCP connection for base file
 - m M/X sets of parallel connections for images.
 - m Response time = $(M+1)O/R + (M/X + 1)2RTT + \text{sum of idle times}$