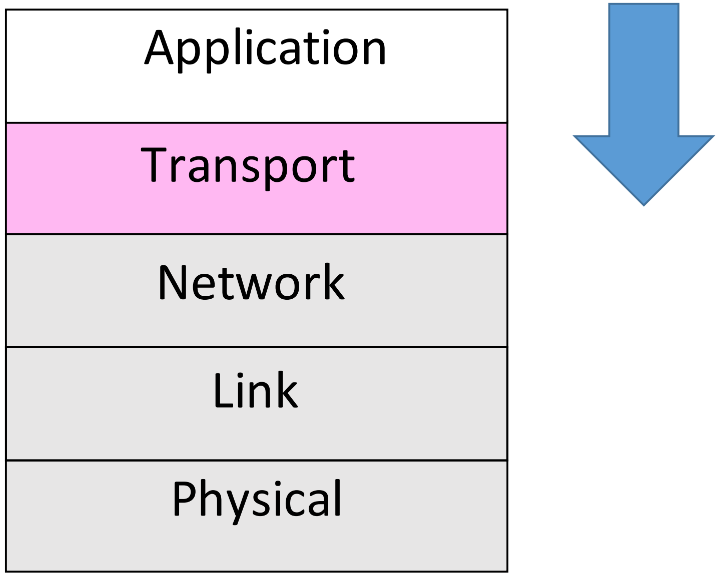
Transport Layer (TCP/UDP)

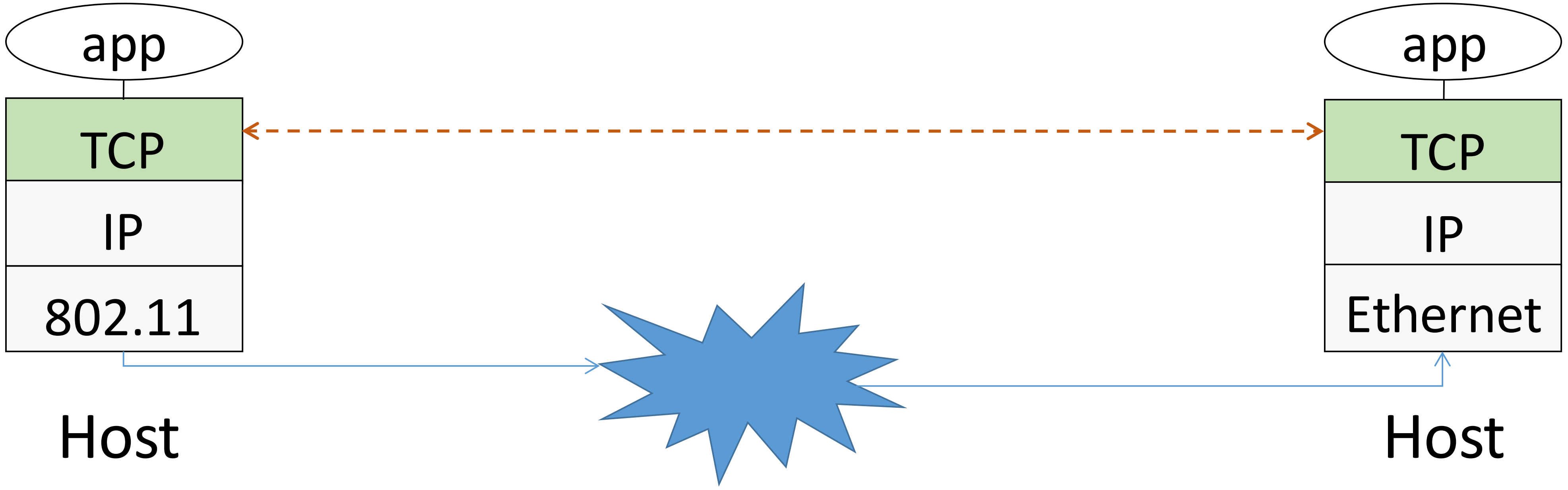
Where we are in the Course

- Moving down to the Transport Layer!

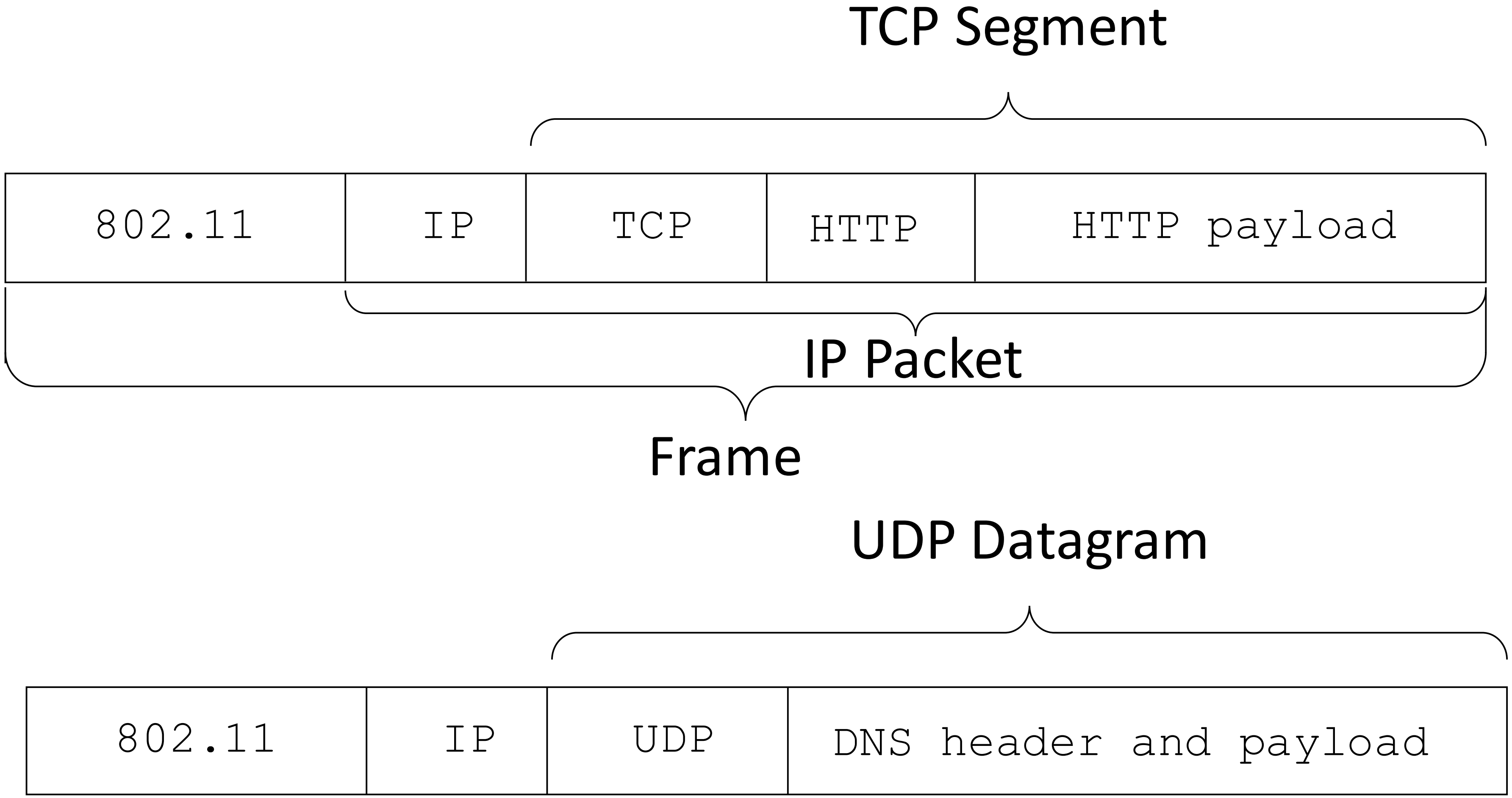


The Transport Layer

- The transport layer provides *end-to-end* connectivity
- To the transport layer, its payload is just bytes



Encapsulation



Transport Layer Services

- Provide different kinds of data delivery across the network to applications

	Unreliable	**Reliable**
Packets	Datagrams (UDP)	
Bytestream		Streams (TCP)

- *Could there be protocols in the two empty boxes?*

Comparison of Internet Transports: Function

TCP	UDP
Streams	Datagrams
Connections	Connectionless
Bytes are delivered to receiving app reliably (once, and in order)	Packets may be lost, reordered, duplicated (but not corrupted)
Arbitrary length content	Fixed maximum datagram size

Comparison of Internet Transports: Performance

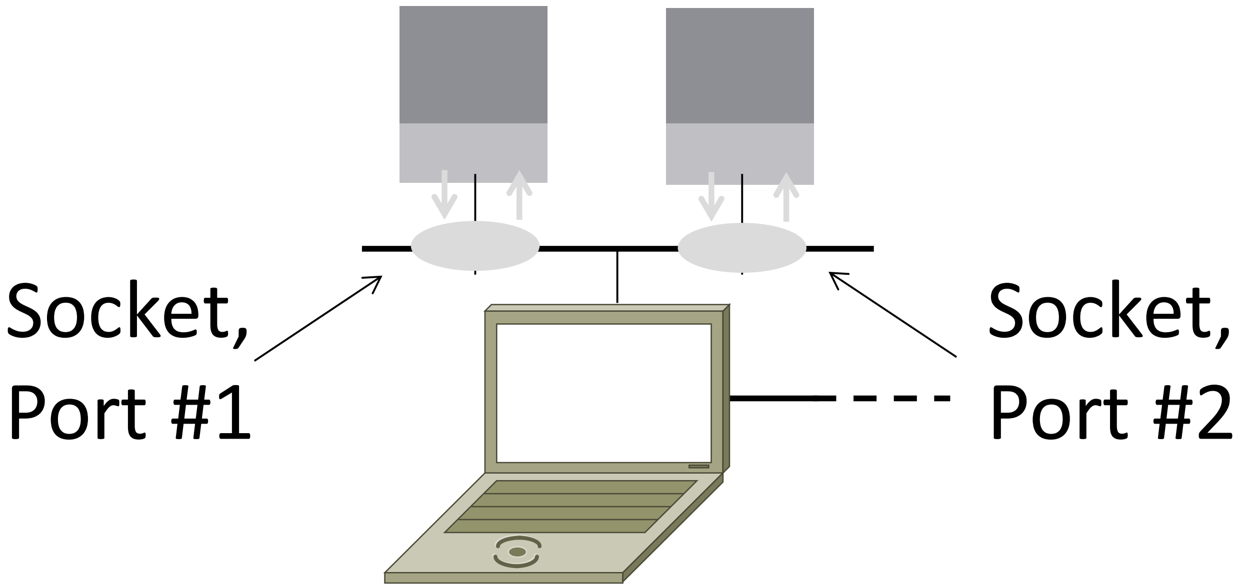
TCP	UDP
Connection latency	No delay
Segment delivery latency (“nagling”)	Datagram is sent now
Flow control matches sender’s rate to receiver’s capability	No flow control (can lead to many lost datagrams)
Congestion control matches sender’s rate to network’s capability	No congestion control (can lead to many lost datagrams)

Socket API

- Simple OS abstraction to use the network
  - The “network” API (really Transport service) used to write all Internet apps
  - Part of all major OSes and languages; originally Berkeley (Unix) ~1983
- Supports both Internet transport services (TCP and UDP)
- The OS provides sockets; the Internet provides the port abstraction

Socket API

- Sockets are associated with ("bound to") Internet ports



Note: A language layer can obscure this interface

Socket API

Same API used for Streams and Datagrams

Primitive	Meaning
SOCKET	Create a new communication endpoint
BIND	Associate a local address (port) with a socket
LISTEN	Announce willingness to accept connections
ACCEPT	Passively establish an incoming connection
CONNECT	Actively attempt to establish a connection
SEND(TO)	Send some data over the socket
RECEIVE(FROM)	Receive some data over the socket
CLOSE	Release the socket

Only needed for Streams

To/From for Datagrams

Ports

- Application process is identified by the tuple
  <IP address, protocol, and port>
  - Ports are 16-bit integers representing local “mailboxes” that a process leases
- Servers often bind to “well-known ports”
  - numbered below 1024
  - require administrative privileges (“privileged ports”)
- Clients often assigned “ephemeral” ports
  - Chosen by OS, used temporarily

Some Well-Known Ports

Port	Protocol	Use
20, 21	FTP	File transfer
22	SSH	Remote login, replacement for Telnet
25	SMTP	Email
80	HTTP	World Wide Web
110	POP-3	Remote email access
143	IMAP	Remote email access
443	HTTPS	Secure Web (HTTP over SSL/TLS)
543	RTSP	Media player control
631	IPP	Printer sharing

Topics

- Service models
  - Socket API and ports
  - Datagrams, Streams
- User Datagram Protocol (UDP)
- Connections (TCP)
- Sliding Window (TCP)
- Flow control (TCP)
- Retransmission timers (TCP)
- Congestion control (TCP)

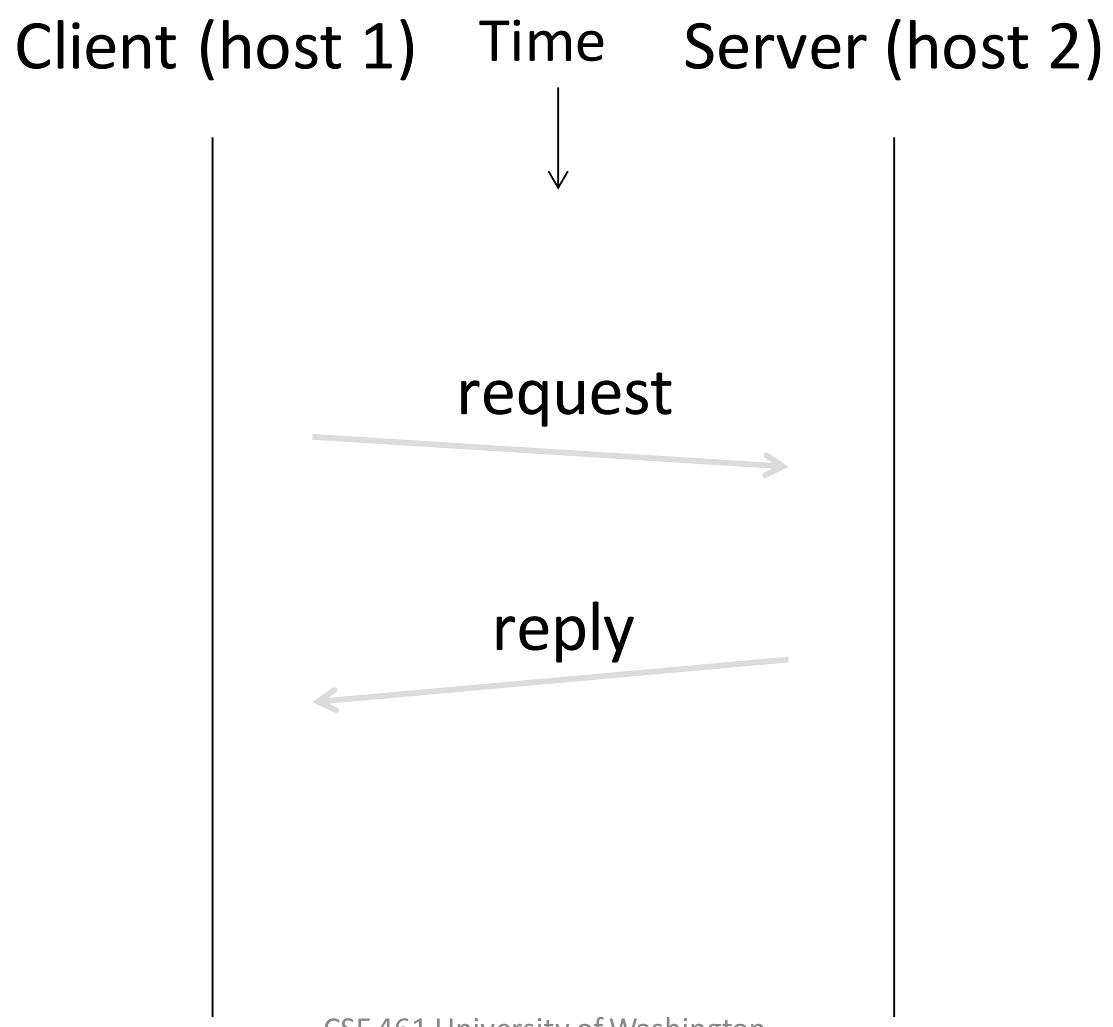
UDP

User Datagram Protocol (UDP)

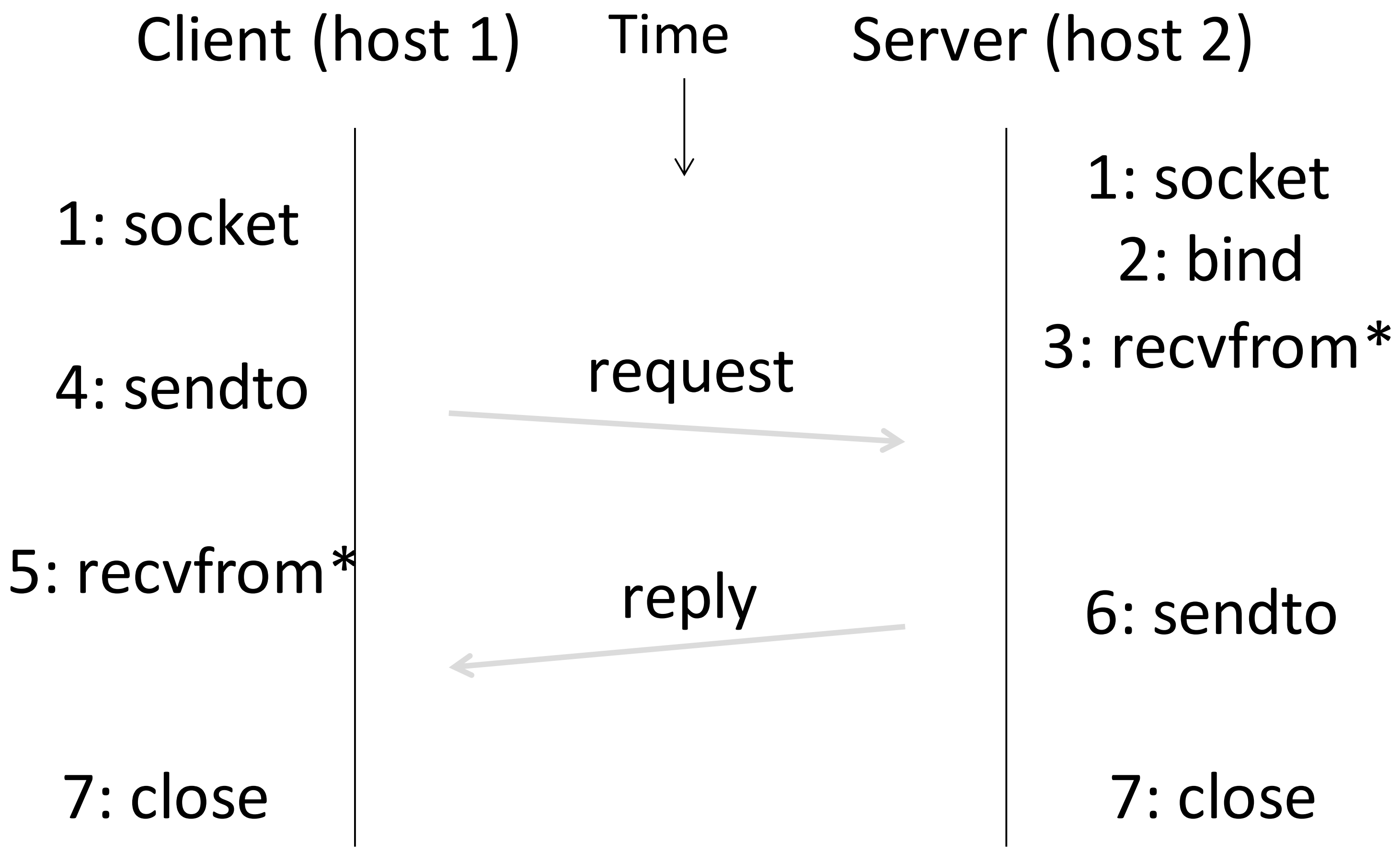
- Used by apps that don't want TCP semantics or for which TCP performance characteristics are unacceptable
  - Voice-over-IP
  - DNS, RPC
  - DHCP

(If application wants reliability and messages then it has work to do!)

Datagram Sockets



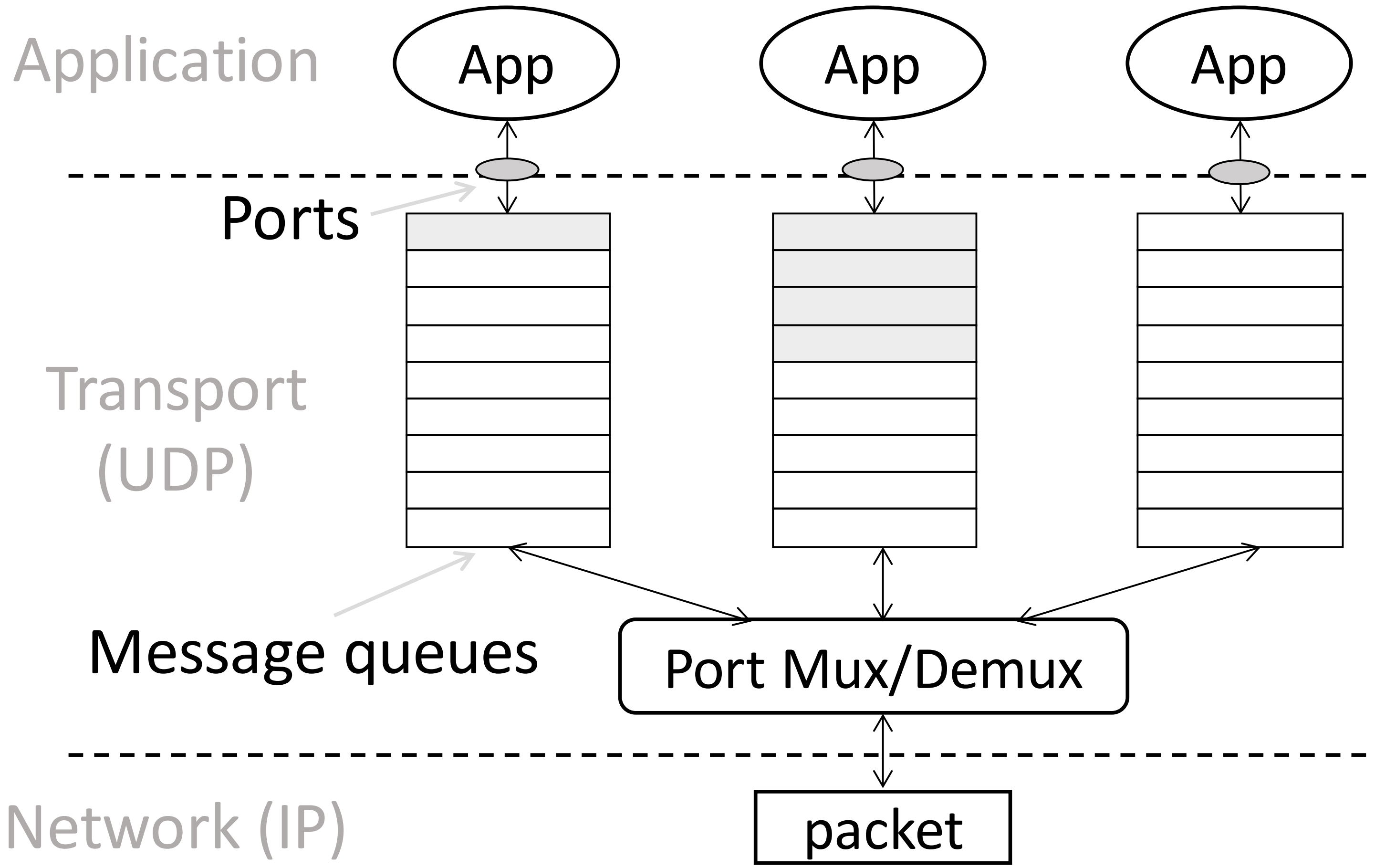
Datagram Sockets



*= call blocks

The protocol implied by this diagram is horribly broken!

UDP Buffering



UDP Header

- Uses ports to identify sending and receiving application processes
- Datagram length limited to 64K
- Checksum (16 bits) for reliability

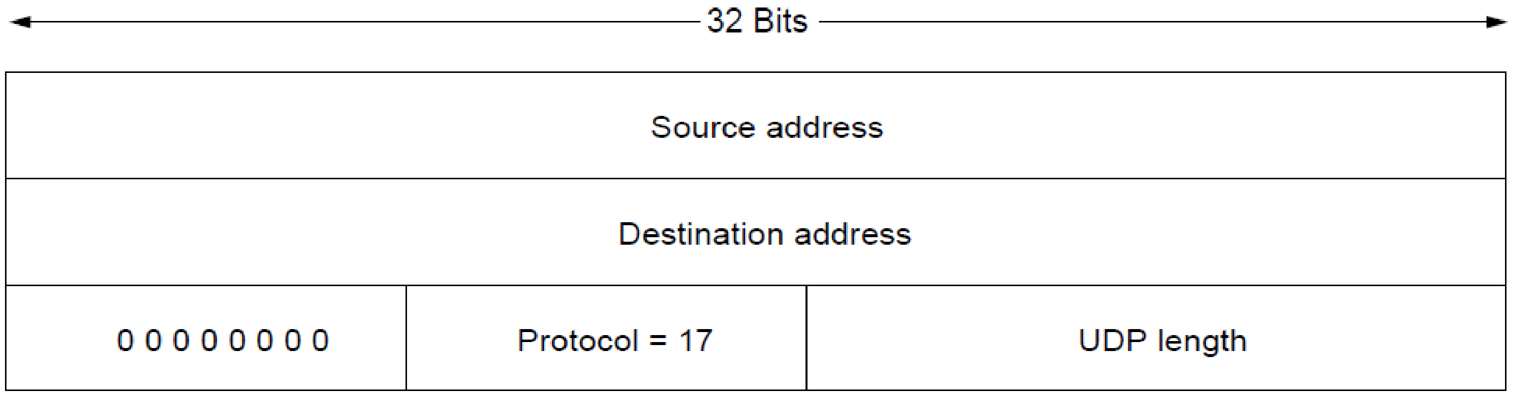
32 Bits

Source port	Destination port
UDP length	UDP checksum

UDP header

UDP Header

- Optional checksum covers UDP segment and IP pseudoheader
  - Checks key IP fields (addresses)
  - Value of zero means “no checksum”



Internet Checksum

- Idea:
  - sender sums up data in N-bit words
    - results in a 16-bit value that is a function of the data

1500 bytes	16 bits

  - receiver performs same summation
  - if value receiver computes doesn’t match value sent by sender, the packet has been corrupted
  - Widely used in, e.g., TCP/IP/UDP

Internet Checksum

- Sum is defined in 1s complement arithmetic (must add back carries)
  - And it's the negative sum
- *"The checksum field is the 16 bit one's complement of the one's complement sum of all 16 bit words …"* – RFC 791

Internet Checksum (2)

Sending:

1. Arrange data in 16-bit words
2. Put zero in checksum position, add
3. Add any carryover back to get 16 bits
4. Negate (complement) to get sum

```
0001
f204
f4f5
f6f7
```

Internet Checksum (3)

Sending:

1. Arrange data in 16-bit words
2. Put zero in checksum position, add
3. Add any carryover back to get 16 bits
4. Negate (complement) to get sum

```
   0001
   f204
   f4f5
   f6f7
+(0000)
------
  2ddf1

   ddf1
+     2
------
   ddf3

   220c
```

Internet Checksum (4)

Receiving:

1. Arrange data in 16-bit words
2. Checksum will be non-zero, add
3. Add any carryover back to get 16 bits
4. Negate the result and check it is 0

```
  0001
  f204
  f4f5
  f6f7
+ 220c
------
```

Internet Checksum (5)

Receiving:

1. Arrange data in 16-bit words
2. Checksum will be non-zero, add
3. Add any carryover back to get 16 bits
4. Negate the result and check it is 0

```
   0001
   f204
   f4f5
   f6f7
+  220c
------
  2fffd

   fffd
+     2
------
   ffff

   0000
```

(Pre-TCP)
Reliability and Retransmissions

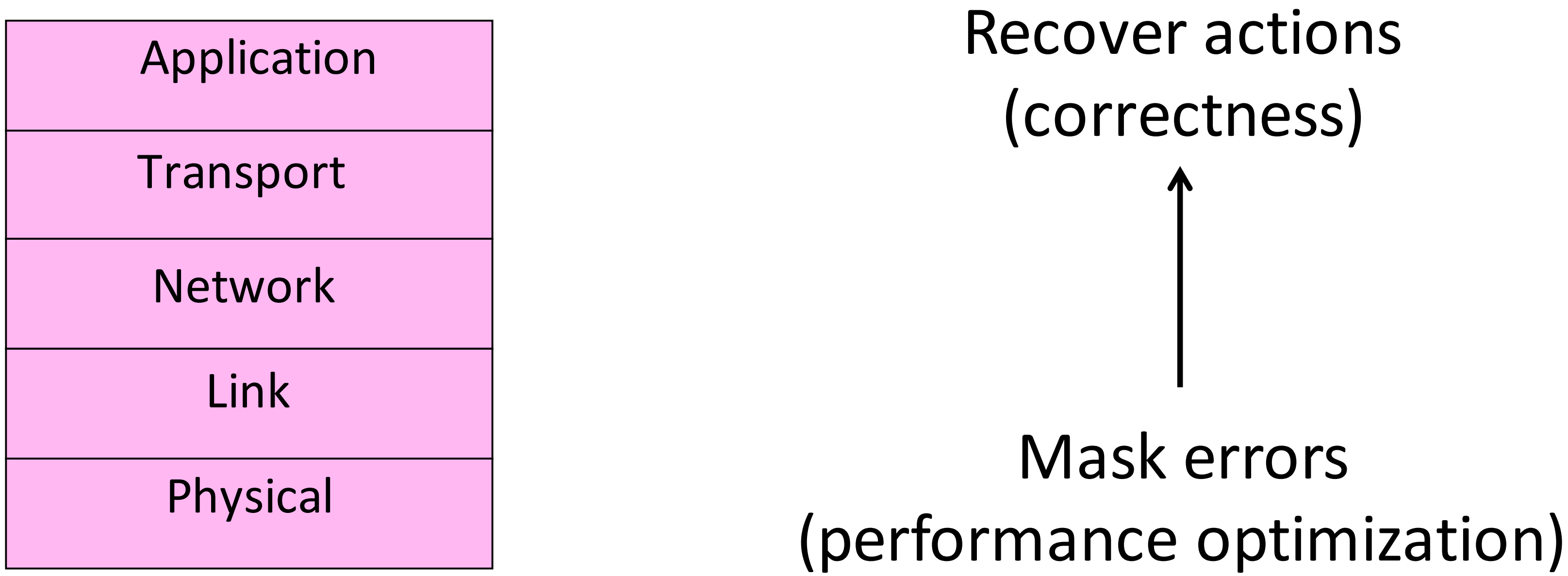
Context on Reliability

- Where in the stack should we place reliability?

Application
Transport
Network
Link
Physical

Context on Reliability (2)

- Everywhere! It is a key issue
  - Different layers contribute differently



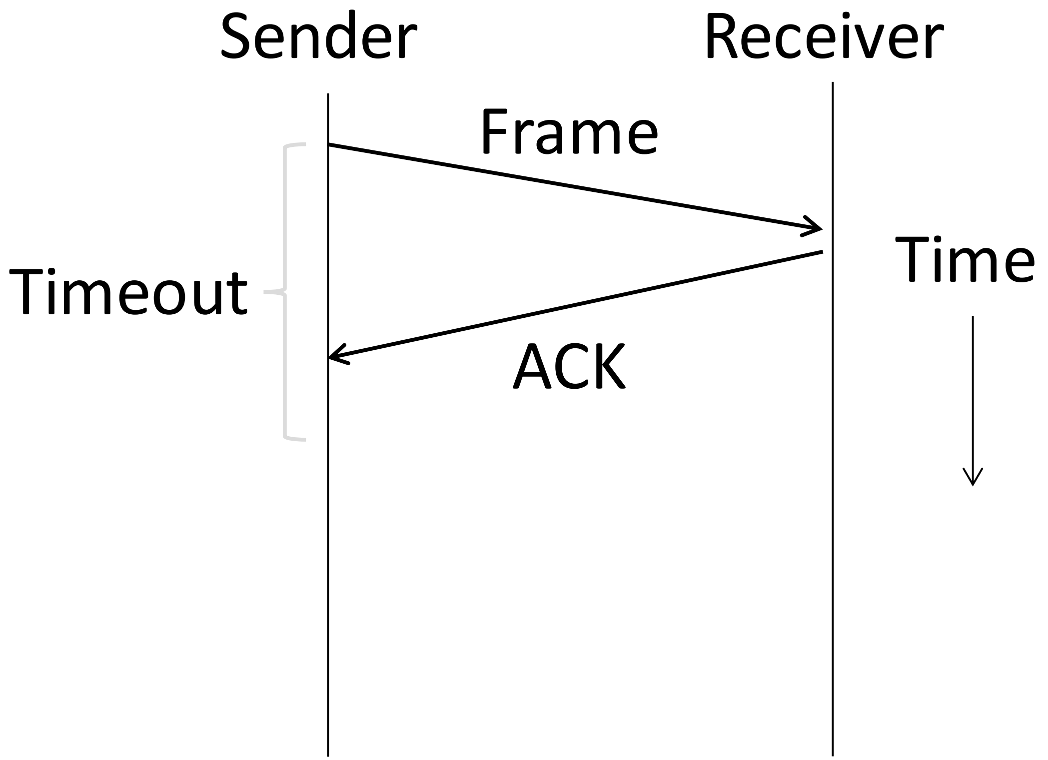
ARQ (Automatic Repeat reQuest)

- ARQ often used when errors are common or must be corrected
  - E.g., WiFi (common) and TCP (must correct)

- Rules at sender and receiver:
  - Receiver automatically acknowledges correct frames with an ACK
    - *positive acknowledgements*
  - Sender automatically resends after a timeout
  - Keep re-sending until an ACK is received

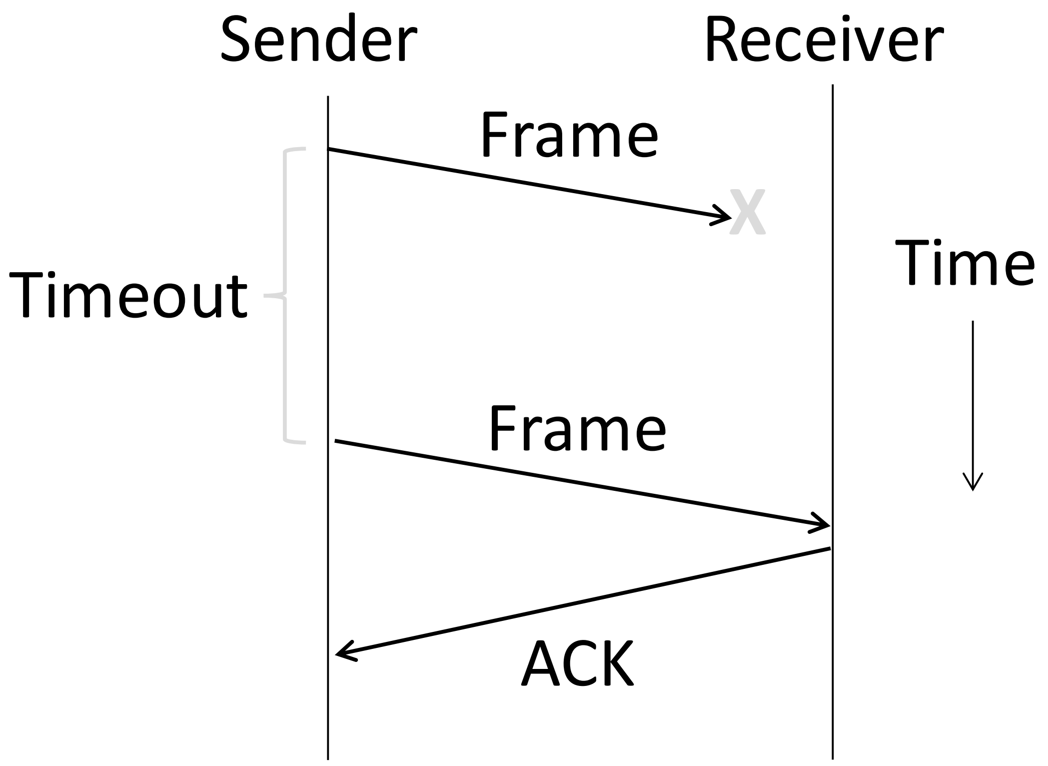
ARQ

- Normal operation (no loss)



ARQ

- Loss and retransmission



So What's Tricky About ARQ?

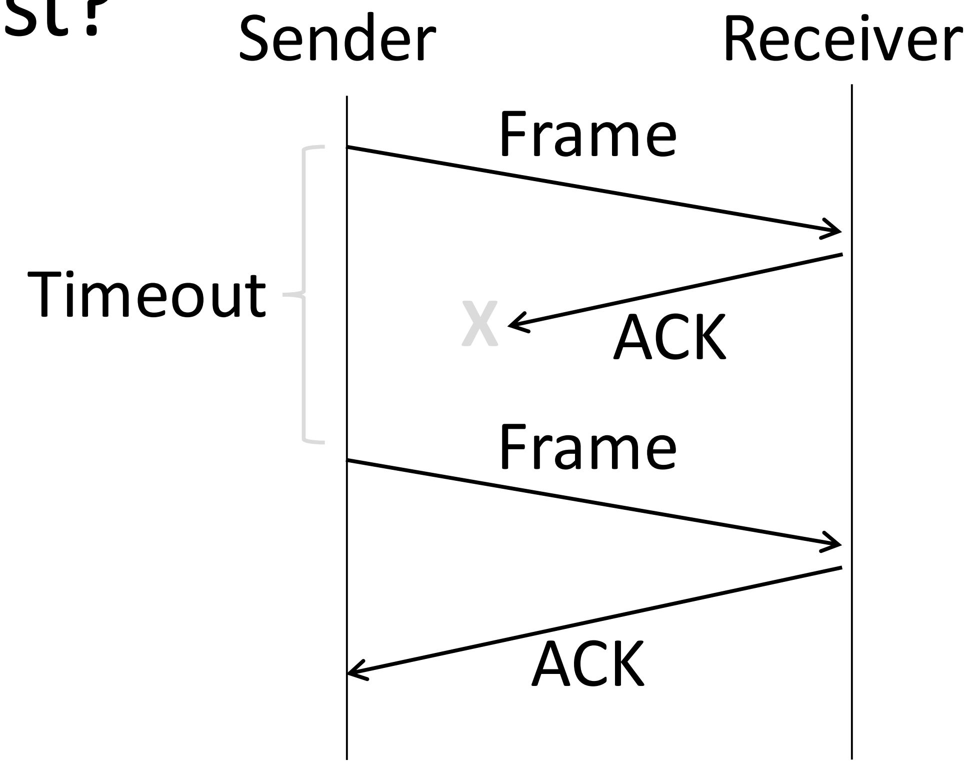
- Two non-trivial issues:
  - How long to set the timeout?
  - How to avoid accepting duplicate frames as new frames
- Want performance in the common case and correctness always

Timeouts

- Timeout should be:
  - Not too big (link goes idle)
  - Not too small (spurious resend)
- Fairly easy on a LAN
  - Clear worst case, little variation
- Fairly difficult over the Internet
  - Much variation, no obvious bound
  - We'll revisit this with TCP (later)

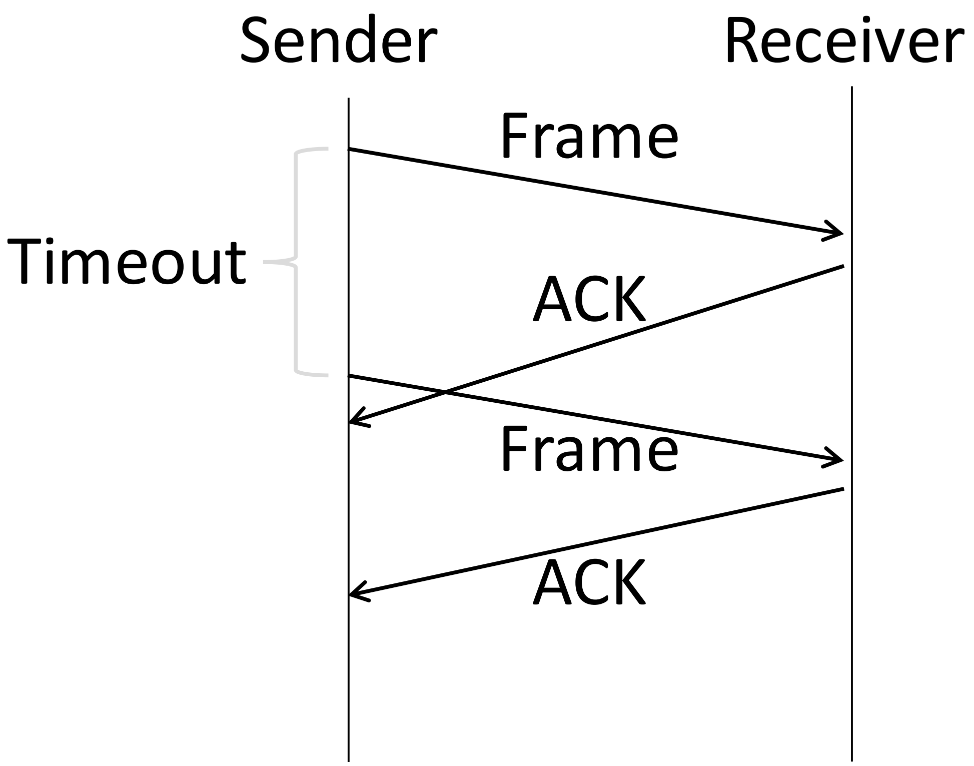
Duplicates

- What happens if an ACK is lost?



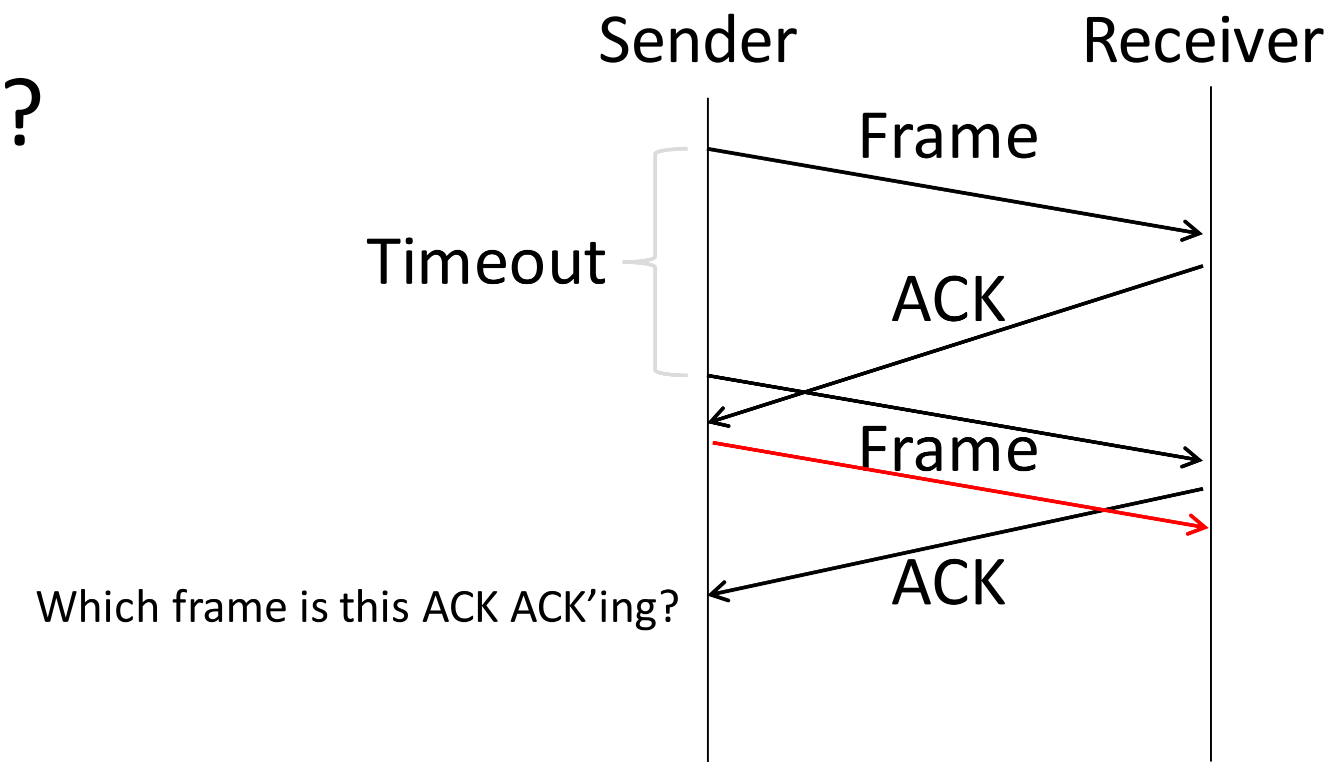
Duplicates

- What happens if the timeout is early?



Duplicates

- What happens if the timeout is early?



Sequence Numbers

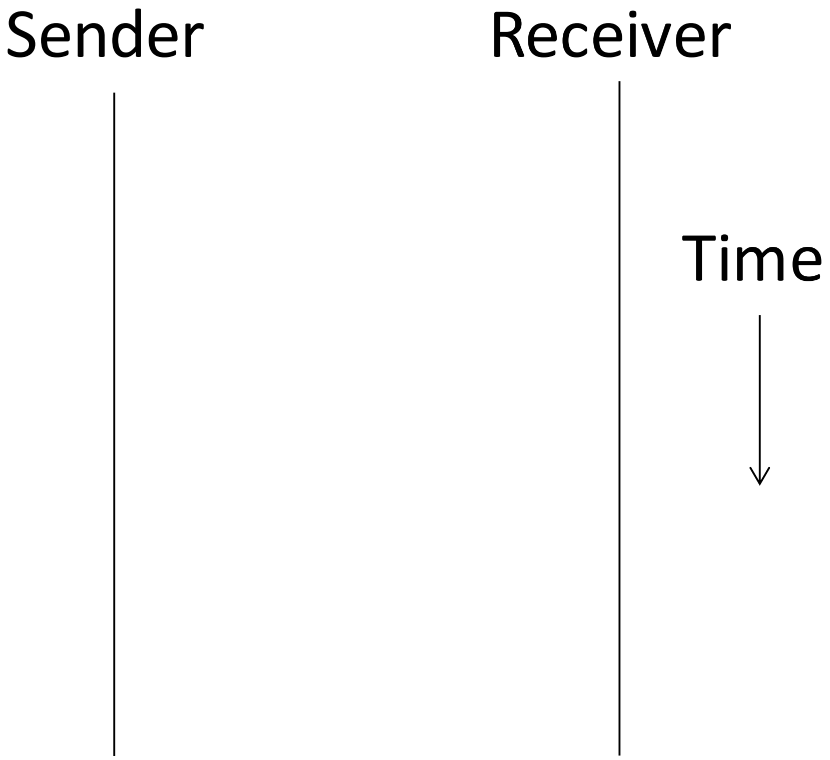
- For correctness, frames and ACKs must both carry sequence numbers

- At an extreme, to distinguish the current frame from the next one, a single bit (two numbers) is sufficient
  - Called <u>Stop-and-Wait</u> protocol

- *In general, the number of packets that can be in flight is limited to half the range of the sequence numbers*

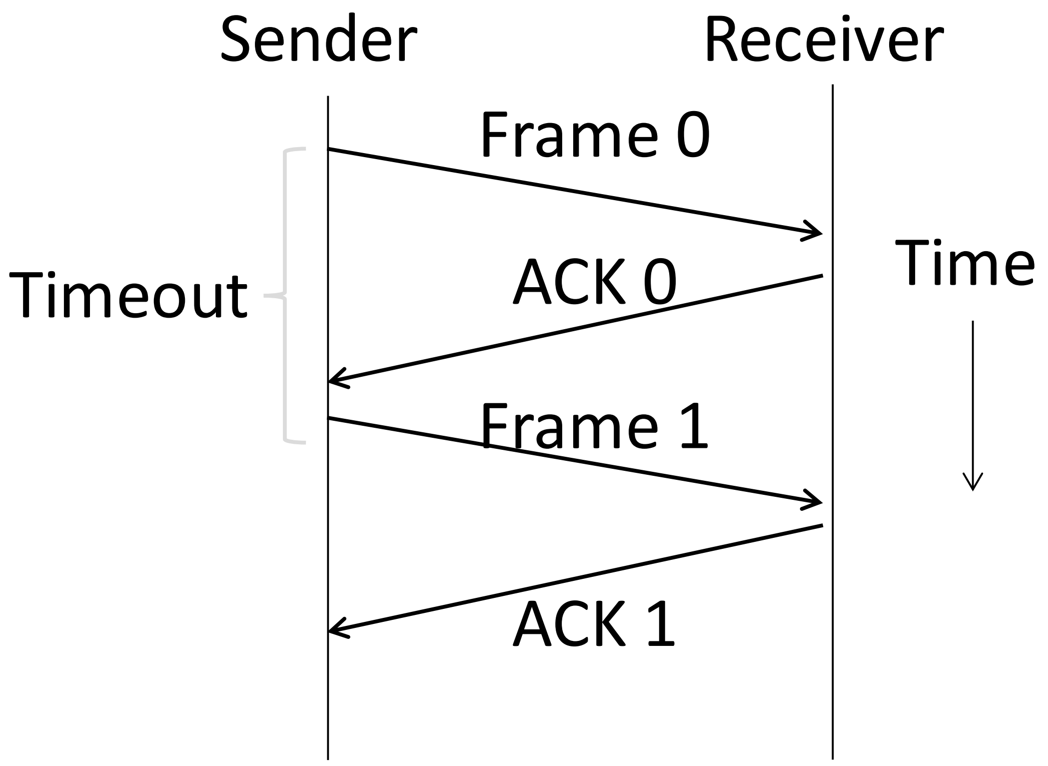
Stop-and-Wait

- In the normal case:



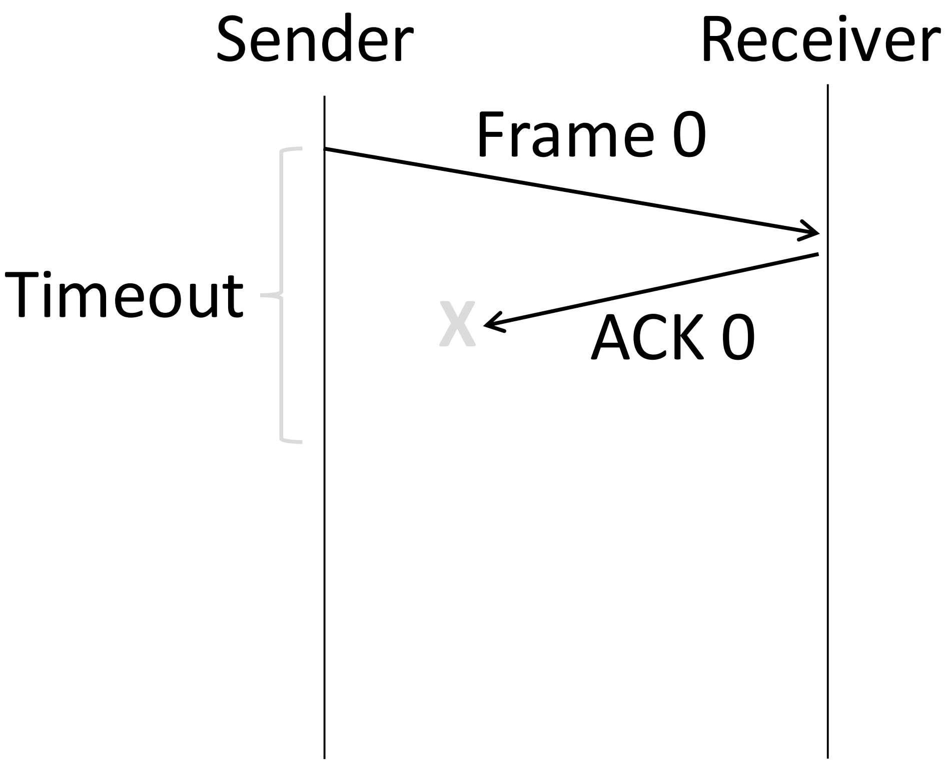
Stop-and-Wait (2)

- In the normal case:



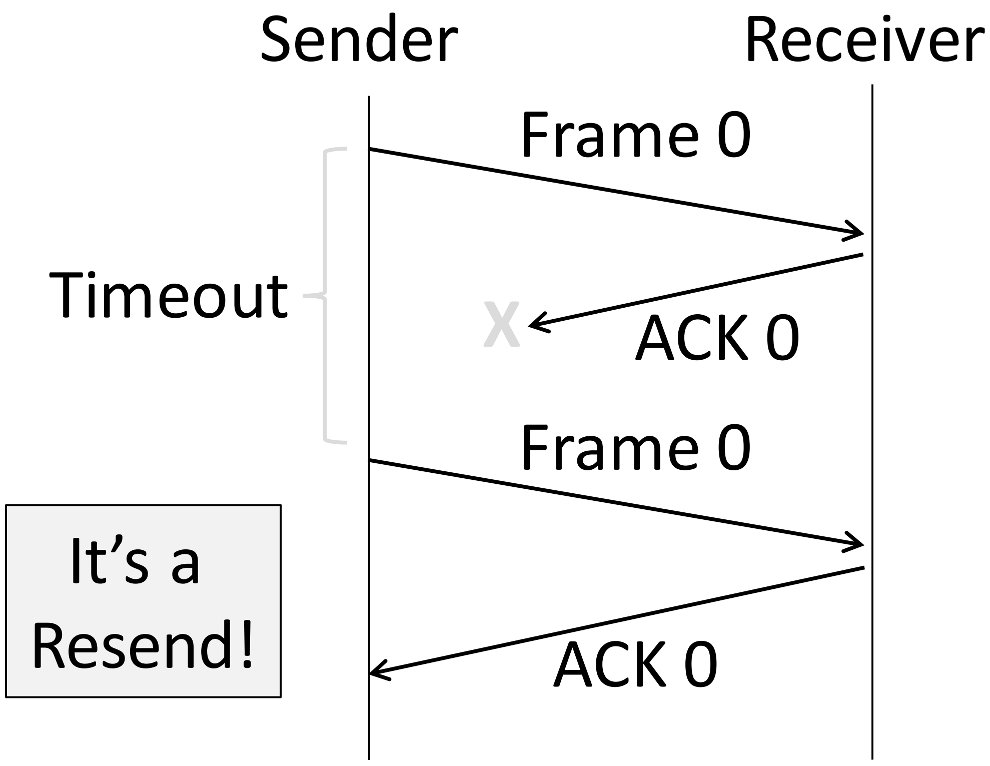
Stop-and-Wait (3)

- With ACK loss:



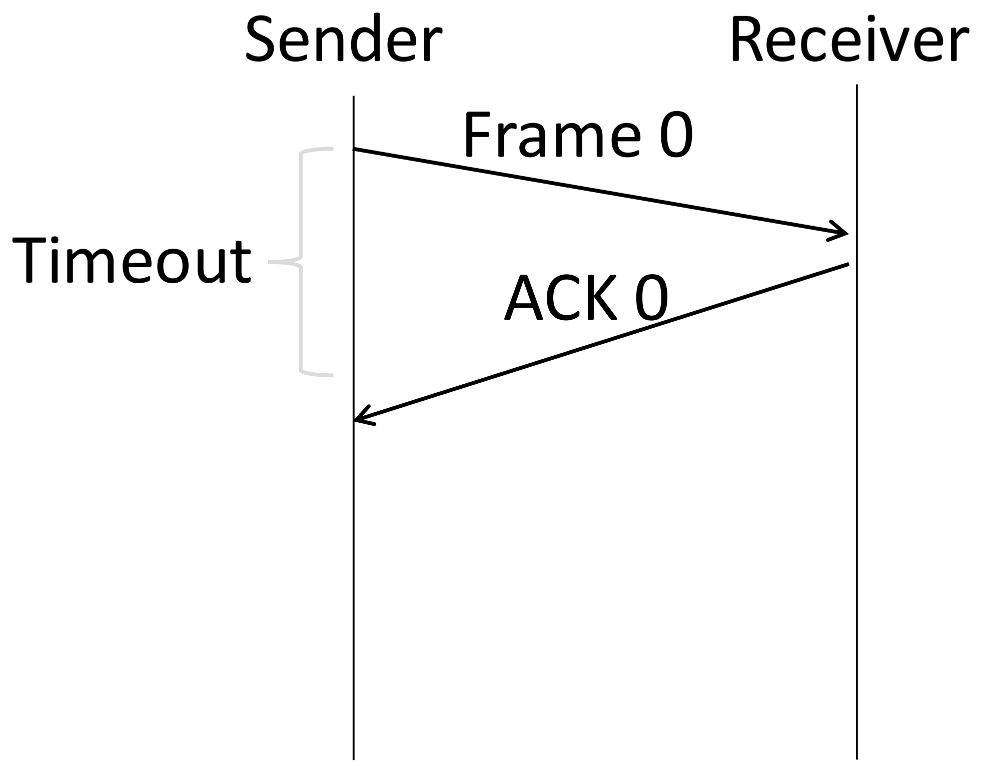
Stop-and-Wait (4)

- With ACK loss:



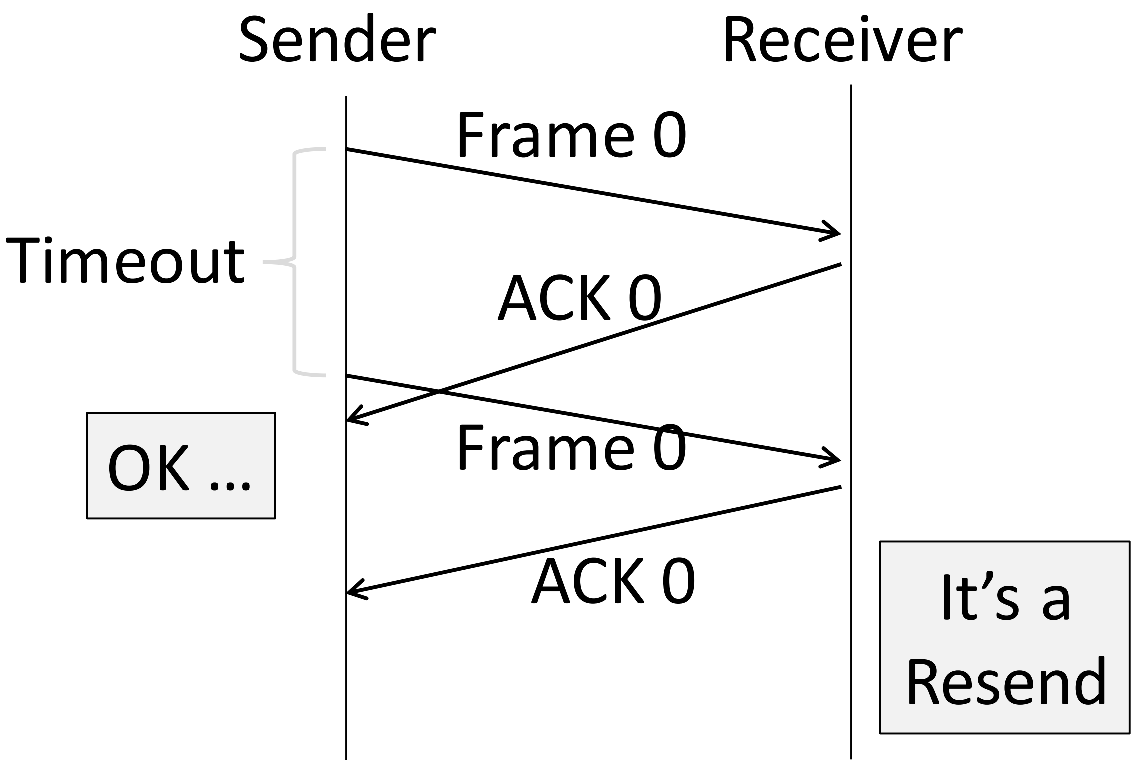
Stop-and-Wait (5)

- With early timeout:



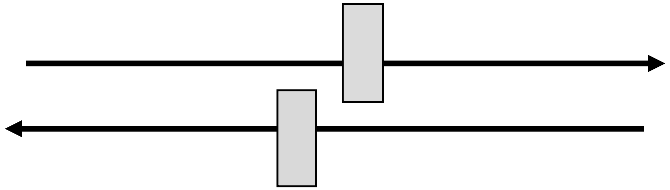
Stop-and-Wait (6)

- With early timeout:



Limitation of Stop-and-Wait

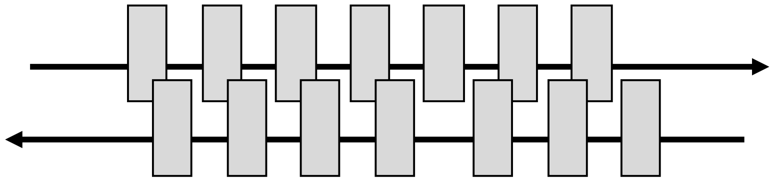
- It allows only a single frame to be outstanding from the sender:
  - Good for LAN, not efficient for high latency communication



- Ex: R=1 Mbps, D = 50 ms
  - How many frames/sec? If R=10 Mbps?

Sliding Window

- Generalization of stop-and-wait
  - Allows W frames to be outstanding
  - Can send W frames per *round trip time* (=2D)

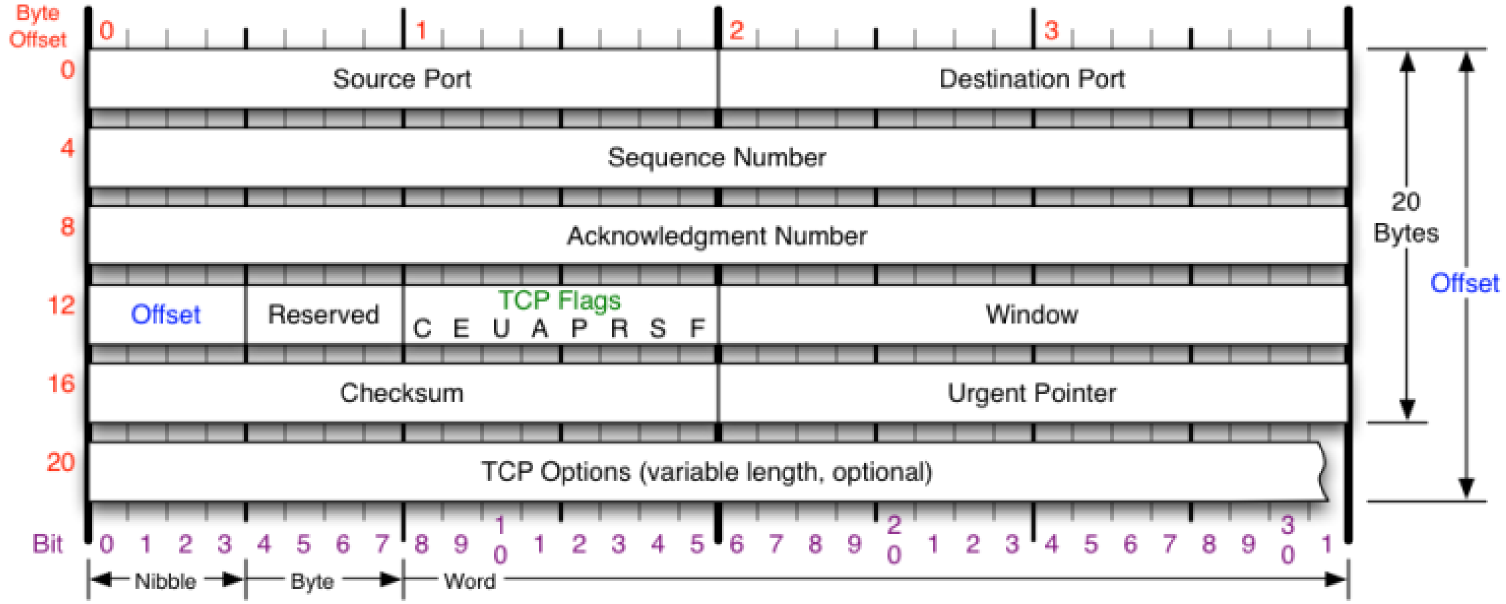


- Various options for numbering frames/ACKs and handling loss
  - Will look at along with

TCP

From https://nmap.org/book/tcpip-ref.html

TCP Header



TCP Flags

C E U A P R S F

Congestion Window

C 0x80 Reduced (CWR)
E 0x40 ECN Echo (ECE)
U 0x20 Urgent
A 0x10 Ack
P 0x08 Push
R 0x04 Reset
S 0x02 Syn
F 0x01 Fin

Congestion Notification

ECN (Explicit Congestion Notification). See RFC 3168 for full details, valid states below.

Packet State	DSB	ECN bits
Syn	0 0	1 1
Syn-Ack	0 0	0 1
Ack	0 1	0 0
No Congestion	0 1	0 0
No Congestion	1 0	0 0
Congestion	1 1	0 0
Receiver Response	1 1	0 1
Sender Response	1 1	1 1

TCP Options

0 End of Options List
1 No Operation (NOP, Pad)
2 Maximum segment size
3 Window Scale
4 Selective ACK ok
8 Timestamp

Checksum

Checksum of entire TCP segment and pseudo header (parts of IP header)

Offset

Number of 32-bit words in TCP header, minimum value of 5. Multiply by 4 to get byte count.

RFC 793

Please refer to RFC 793 for the complete Transmission Control Protocol (TCP) Specification.

TCP Protocol

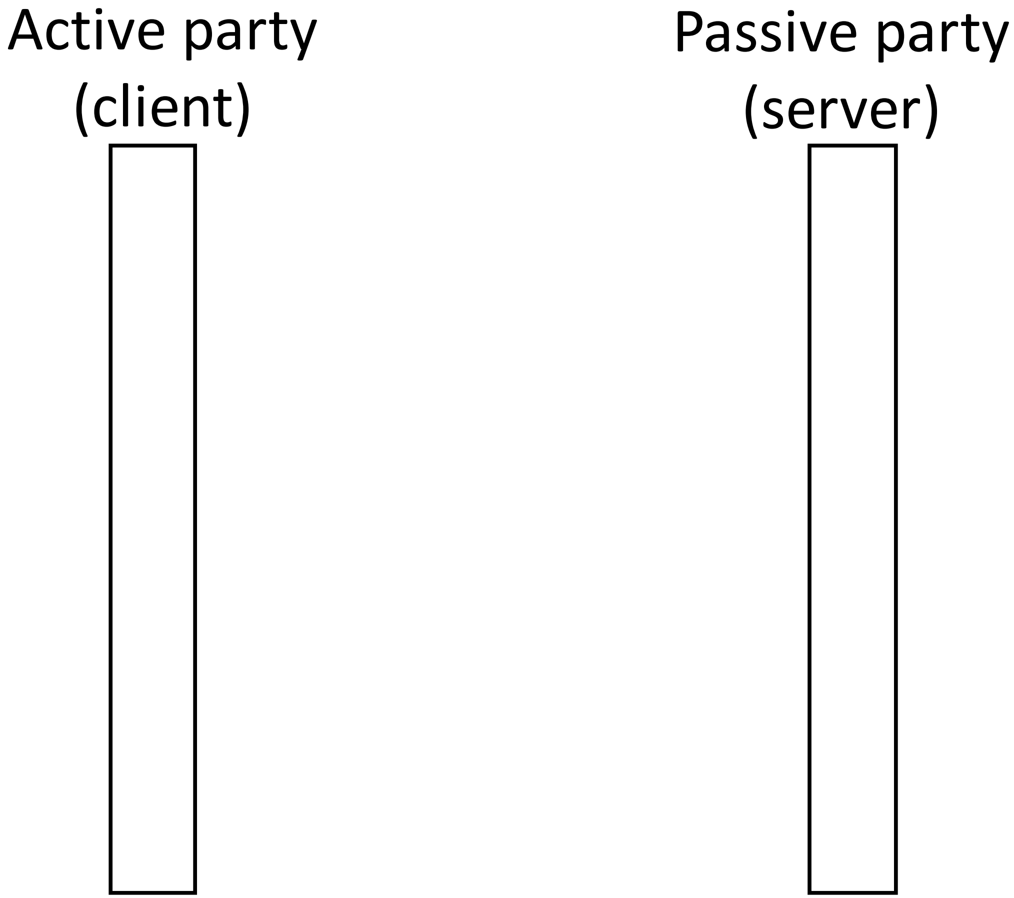
- TCP Consists of 3 primary phases:
  - Connection Establishment (Setup)
  - Sliding Windows/Flow Control
  - Connection Release (Teardown)

Connection Establishment

- Both sender and receiver must be ready before we start the transfer of data
  - Need to agree on a set of parameters
  - e.g., the Maximum Segment Size (MSS)
- This is *signaling*
  - It sets up state at the endpoints
  - Like “dialing” for a telephone call

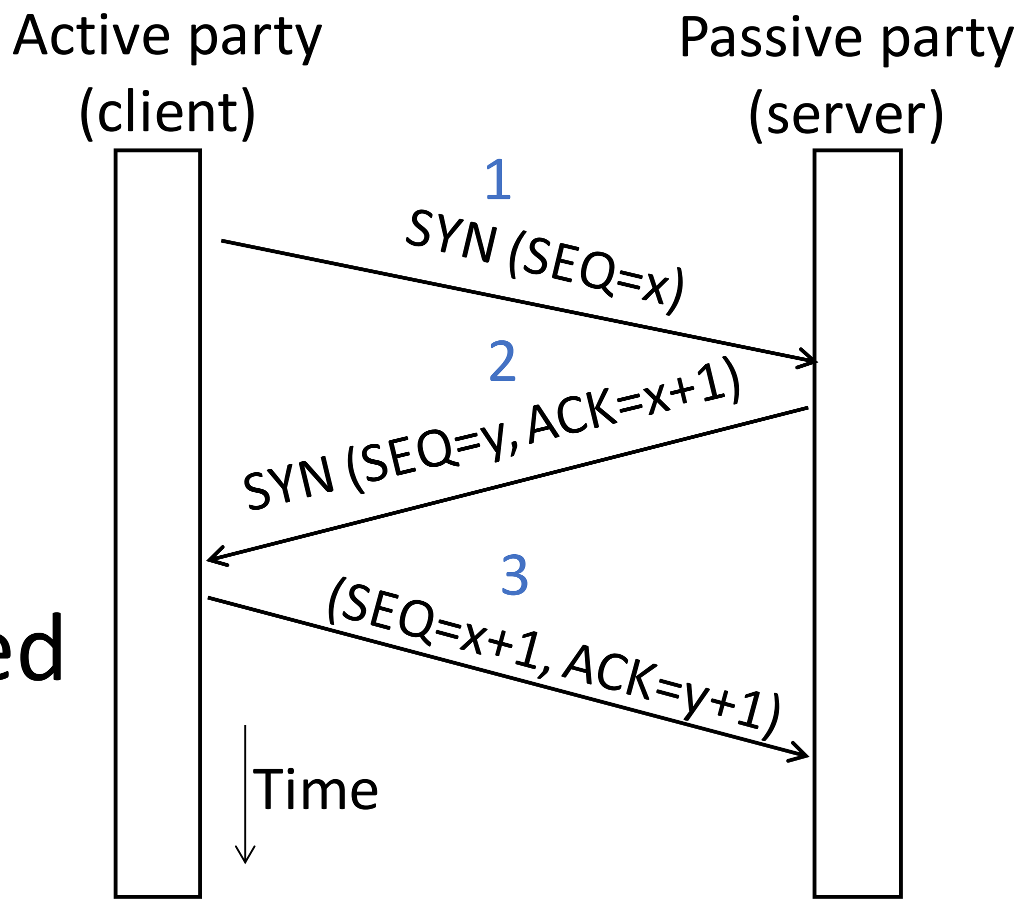
Three-Way Handshake

- Used in TCP; opens connection for data in both directions
- Each side probes the other with a fresh Initial Sequence Number (ISN)
  - Sends on a SYNchronize segment
  - Echo on an ACKnowledge segment
- Chosen to be robust even against delayed duplicates



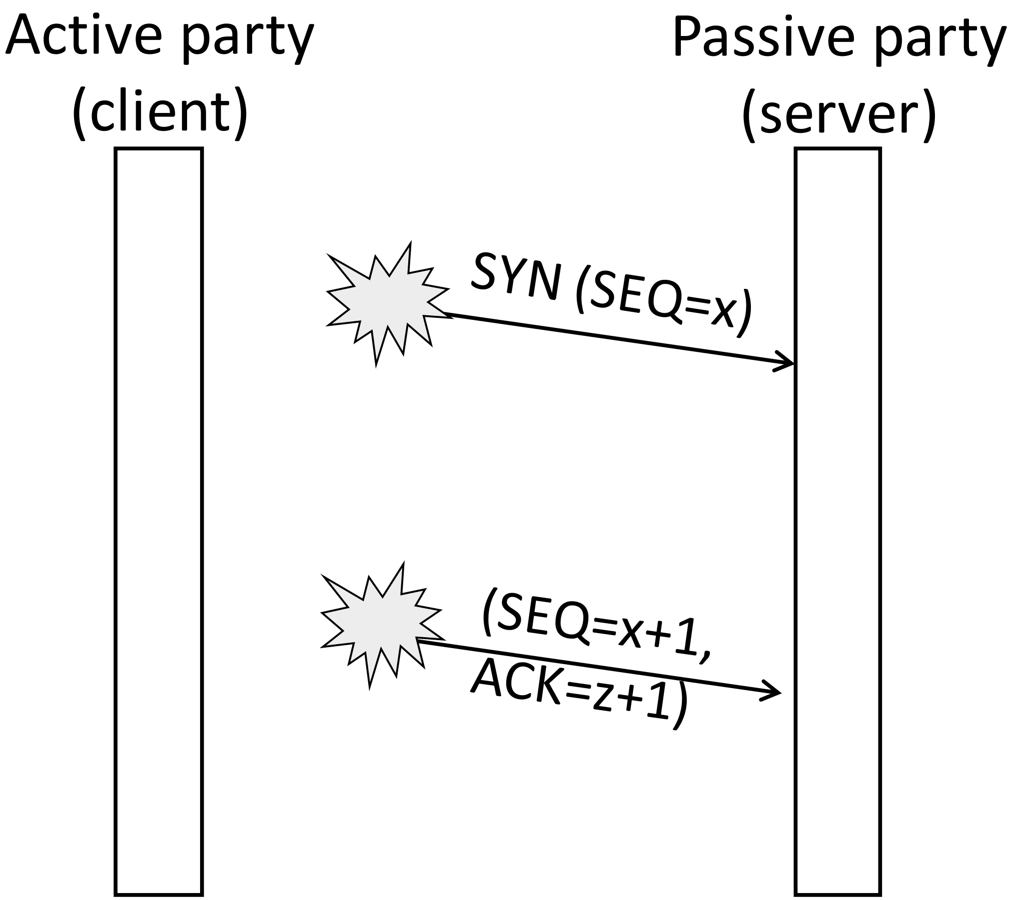
Three-Way Handshake (2)

- Three steps:
  - Client sends SYN(x)
  - Server replies with SYN(y)ACK(x+1)
  - Client replies with ACK(y+1)
  - SYNs are retransmitted if lost
- Sequence and ack numbers carried on further segments



Three-Way Handshake (3)

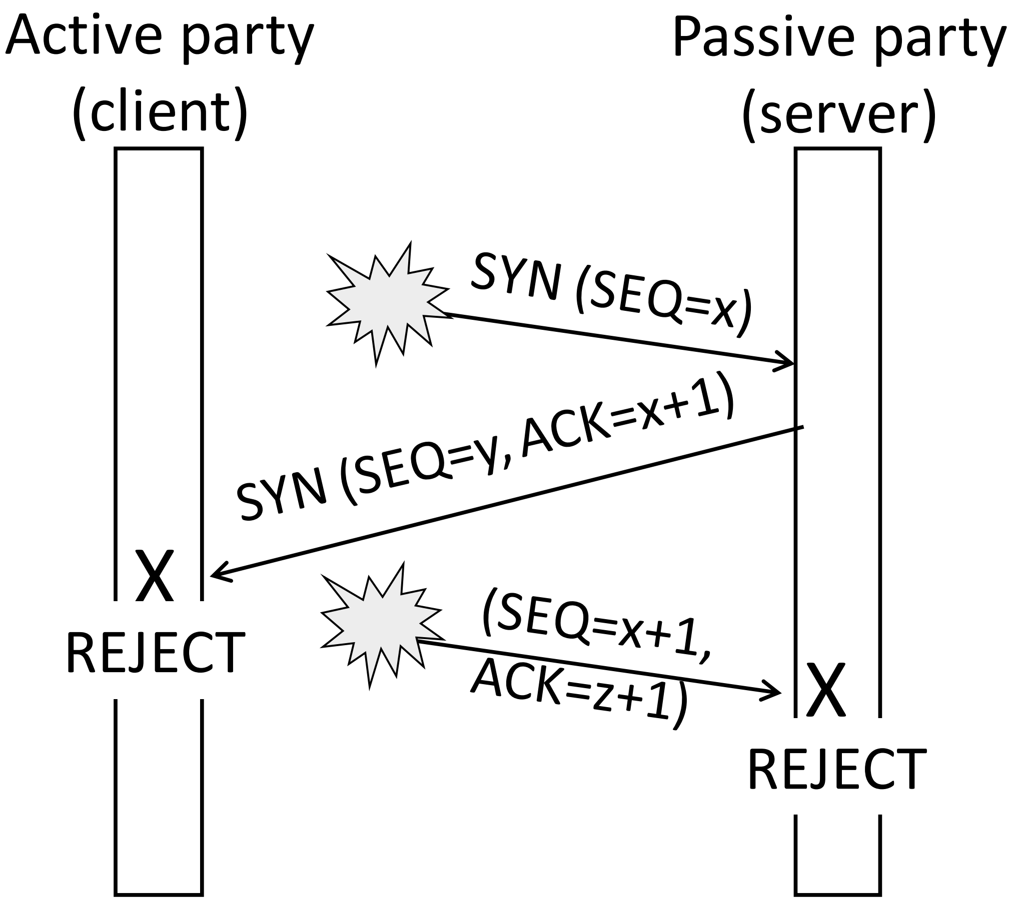
- Suppose delayed, duplicate copies of the SYN and ACK arrive at the server!
  - Improbable, but anyhow …



Three-Way Handshake (4)

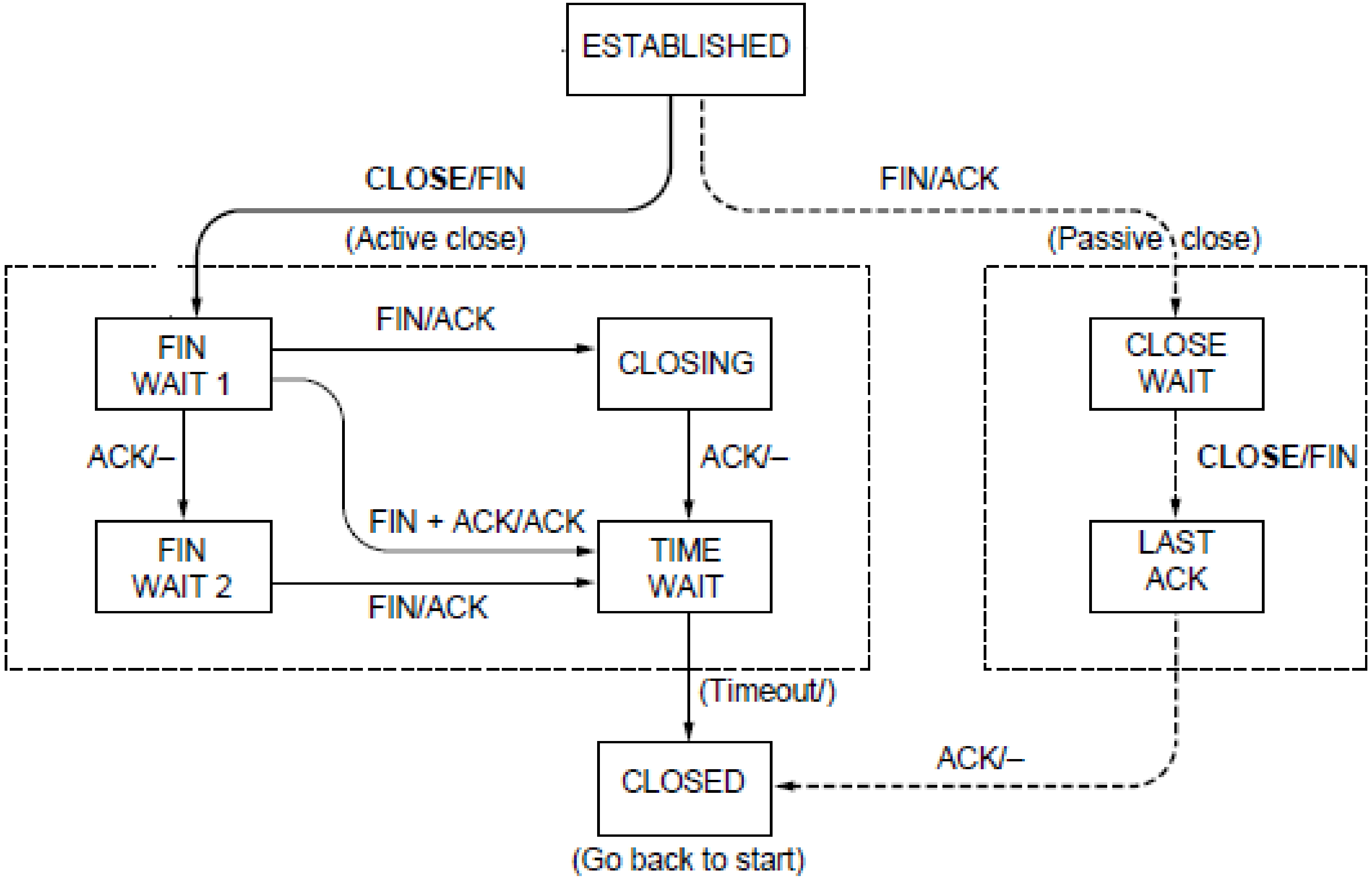
- Suppose delayed, duplicate copies of the SYN and ACK arrive at the server!
  - Improbable, but anyhow …

- Connection will be cleanly rejected on both sides ☺



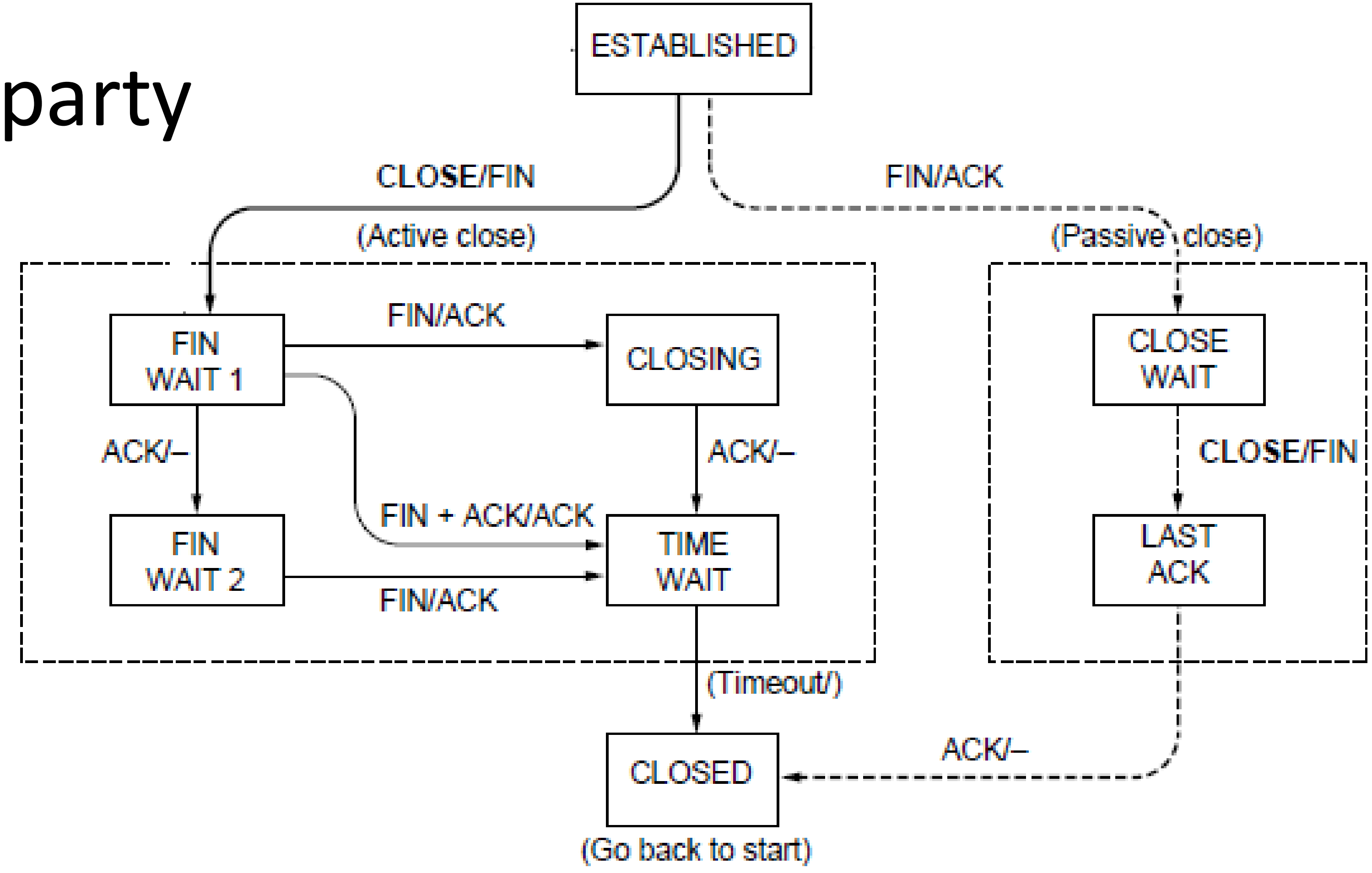
TCP Connection State Machine

Both parties run instances of this state machine



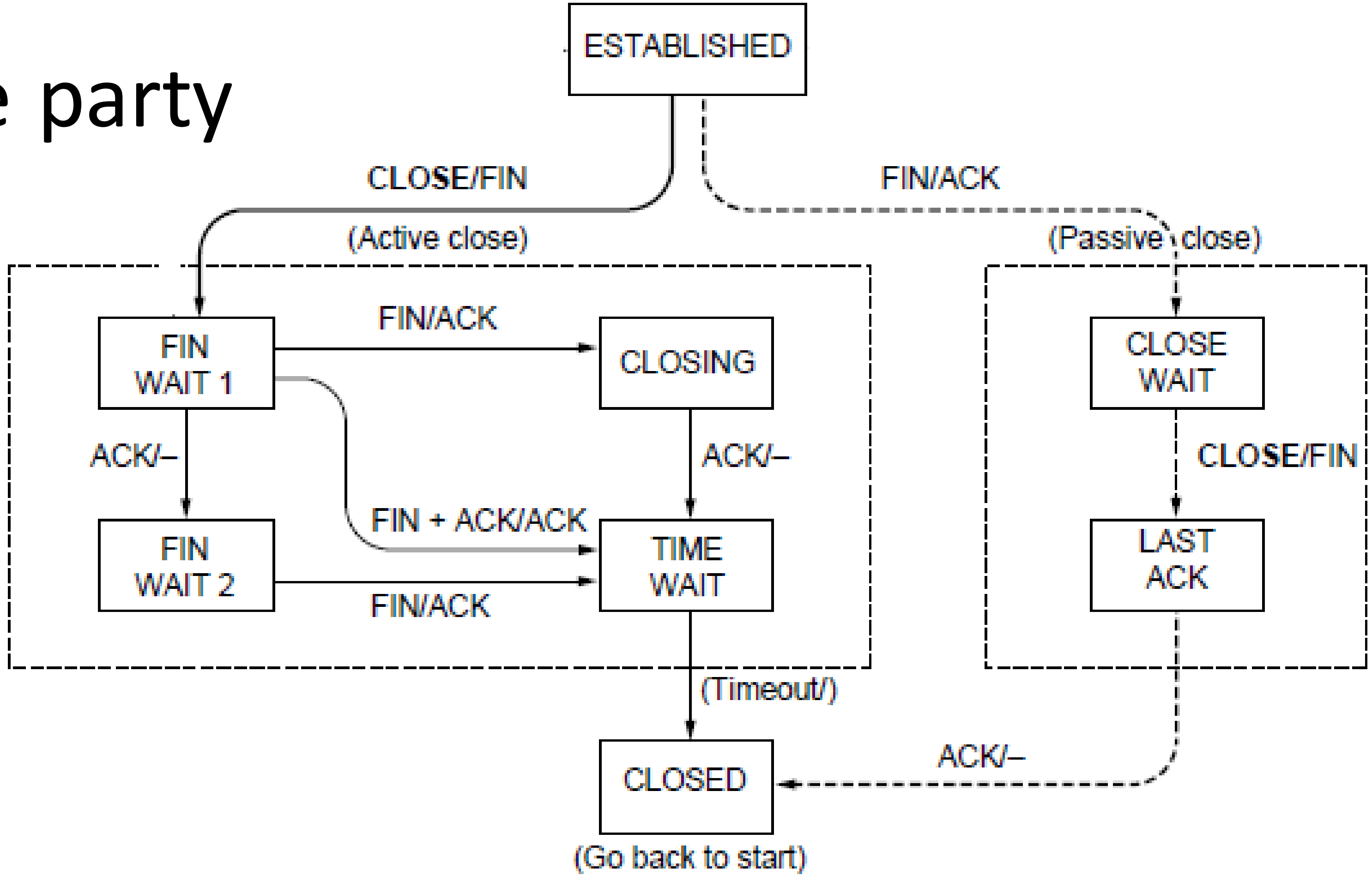
TCP Release

- Follow the active party



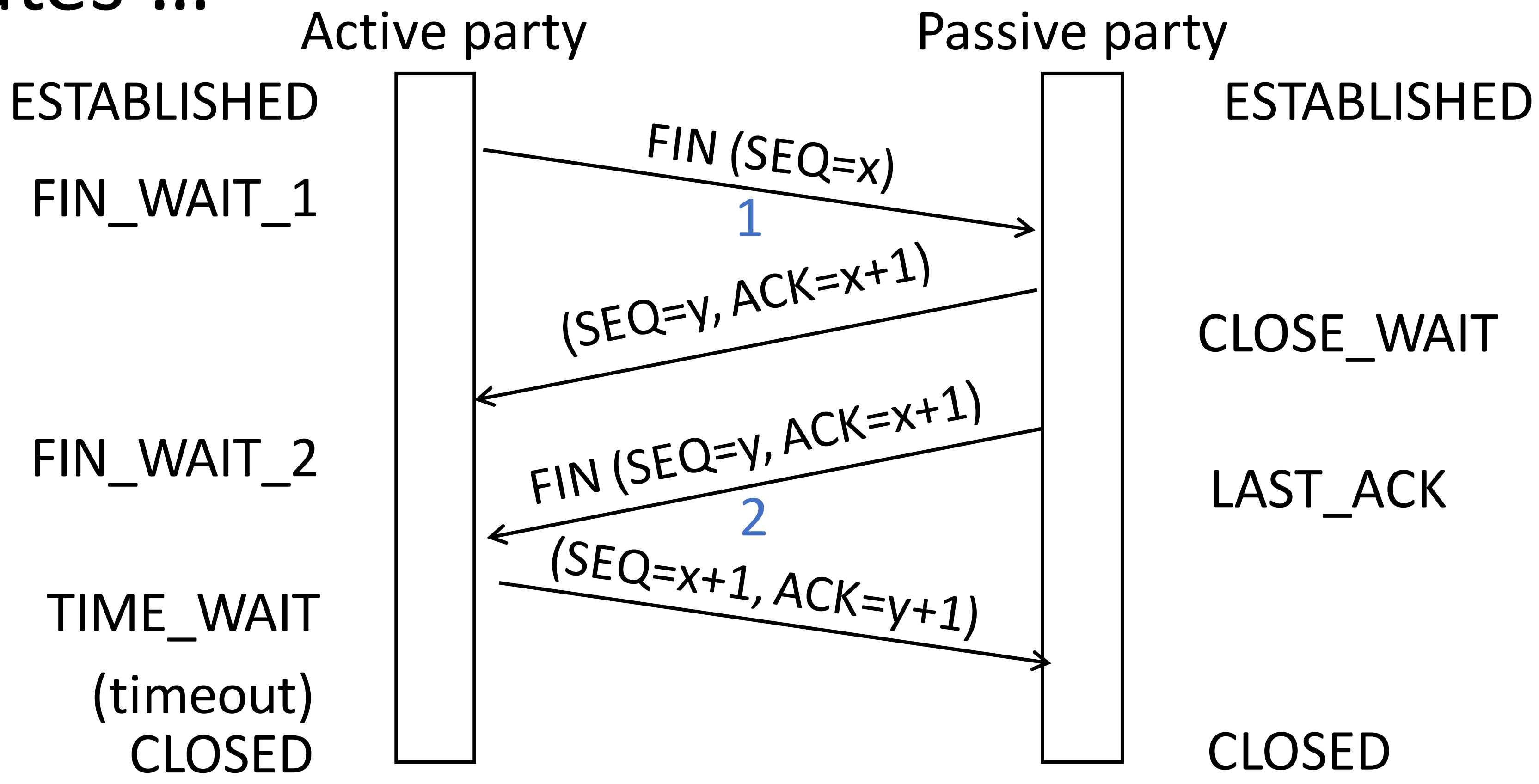
TCP Release

- Follow the passive party



TCP Release

- Again, with states …



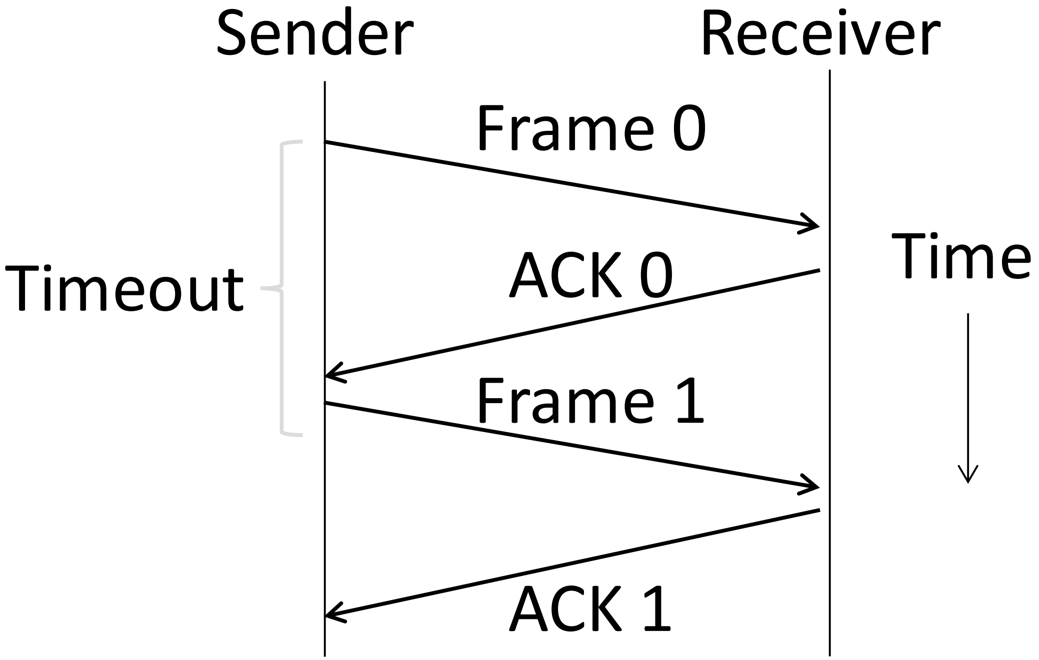
TIME_WAIT State

- Wait a long time after sending all segments and before completing the close
  - Two times the maximum segment lifetime of 60 seconds
- Why?
  - ACK might have been lost, in which case FIN will be resent for an orderly close
  - Could otherwise interfere with a subsequent connection

Flow Control

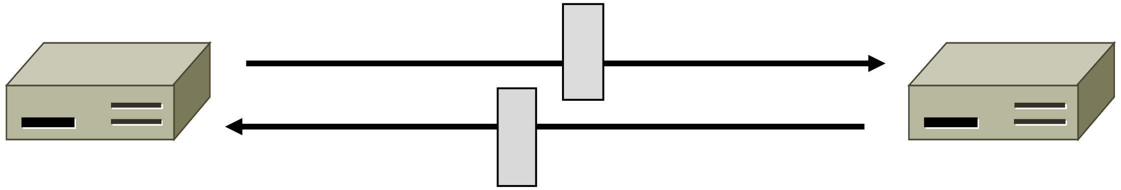
Recall

- ARQ with one message at a time is Stop-and-Wait (normal case below)



Limitation of Stop-and-Wait

- It allows only a single message to be outstanding from the sender:
  - Fine for LAN (only one frame fit)
  - Not efficient for network paths with BD >> 1 packet

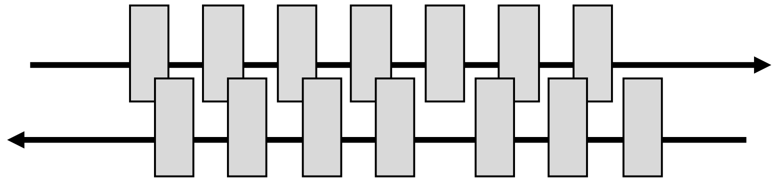


Limitation of Stop-and-Wait (2)

- Example: R=1 Mbps, D = 50 ms
  - RTT (Round Trip Time) = 2D = 100 ms
  - How many packets/sec?
  - What if R=10 Mbps?

Sliding Window

- Generalization of stop-and-wait
  - Allows W packets to be outstanding
  - Can send W packets per RTT (=2D)



  - Pipelining improves performance
  - Need W=2BD to fill network path

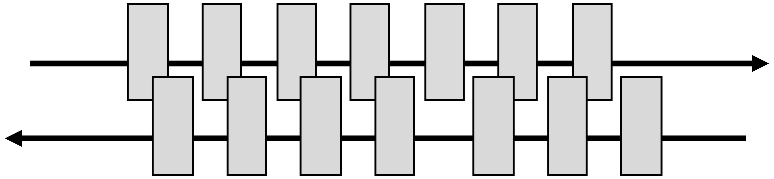
Sliding Window (2)

- What W will use the network capacity?
- Ex: R=1 Mbps, D = 50 ms

- Ex: What if R=10 Mbps?

Sliding Window (3)

- Ex: R=1 Mbps, D = 50 ms
  - 2BD = 10^6 b/sec x 100. 10^{-3} sec = 100 kbit
  - W = 2BD = 10 packets of 1250 bytes



- Ex: What if R=10 Mbps?
  - 2BD = 1000 kbit
  - W = 2BD = 100 packets of 1250 bytes

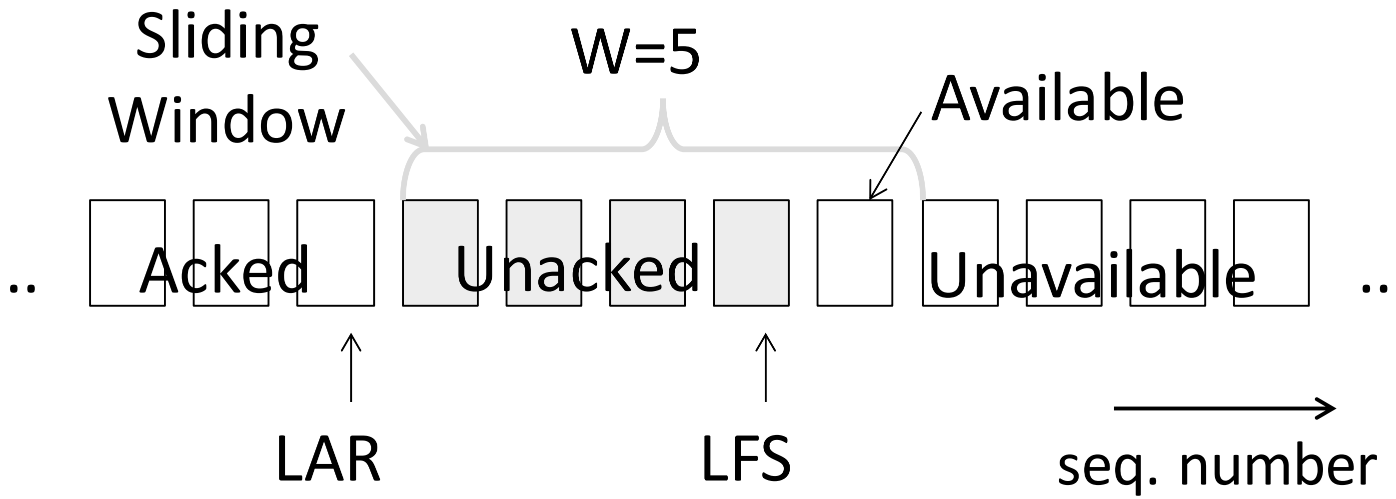
Sliding Window Protocol

- Many variations, depending on how buffers, acknowledgements, and retransmissions are handled

- <u>Go-Back-N</u>
  - Simplest version, can be inefficient
- <u>Selective Repeat</u>
  - More complex, better performance

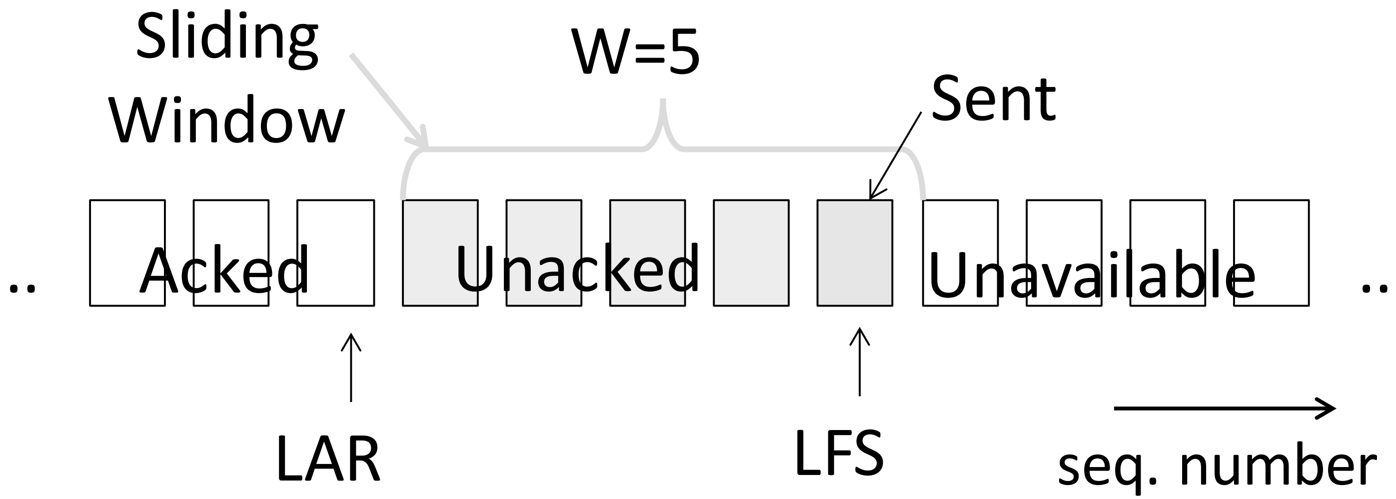
Sliding Window – Sender

- Sender buffers up to W segments until they are acknowledged
  - LFS=LAST FRAME SENT, LAR=LAST ACK REC'D
  - Sends while LFS – LAR ≤ W



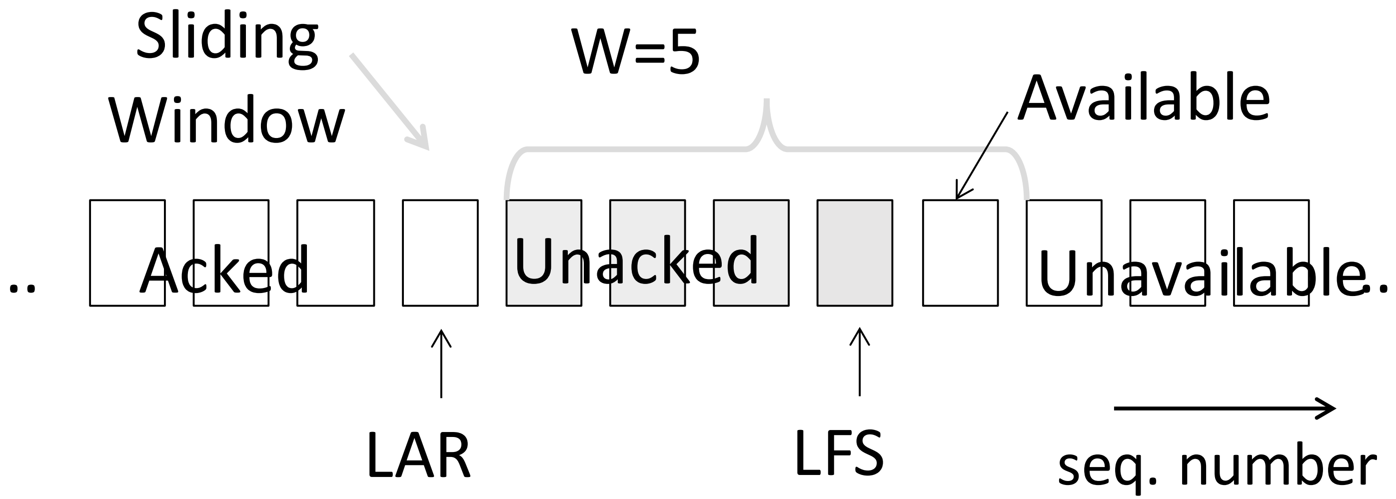
Sliding Window – Sender (2)

- Transport accepts another segment of data from the Application ...
  - Transport sends it (as LFS–LAR → 5)



Sliding Window – Sender (3)

- Next higher ACK arrives from peer…
  - Window advances, buffer is freed
  - LFS–LAR → 4 (can send one more)



Sliding Window – Go-Back-N

- Receiver keeps only a single packet buffer for the next segment
  - State variable, LAS = LAST ACK SENT
- On receive:
  - If seq. number is LAS+1, accept and pass it to app, update LAS, send ACK
  - Otherwise discard (as out of order)

Sliding Window – Selective Repeat

- Receiver passes data to app in order, and buffers out-of-order segments to reduce retransmissions
- ACK conveys highest in-order segment, plus hints about out-of-order segments
- TCP uses a selective repeat design; we'll see the details later

Sliding Window – Selective Repeat (2)

- Buffers W segments, keeps state variable LAS = LAST ACK SENT
- On receive:
  - Buffer segments [LAS+1, LAS+W]
  - Send app in-order segments from LAS+1, and update LAS
  - Send ACK for LAS regardless

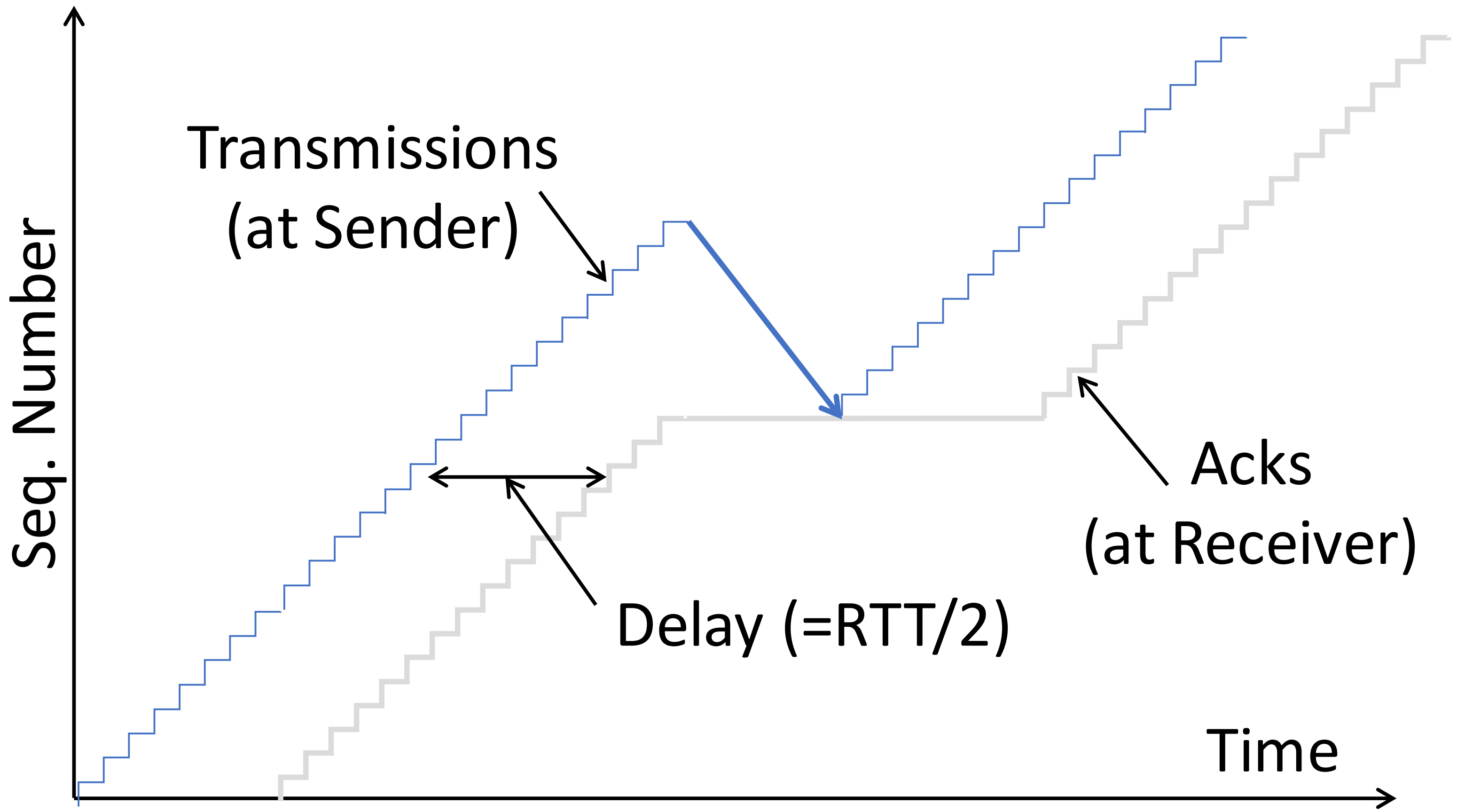
Sliding Window – Retransmissions

- Go-Back-N uses a single timer to detect losses
  - On timeout, resends buffered packets starting at LAR+1
- Selective Repeat uses a timer per unacked segment to detect losses
  - On timeout for segment, resend it
  - Hope to resend fewer segments

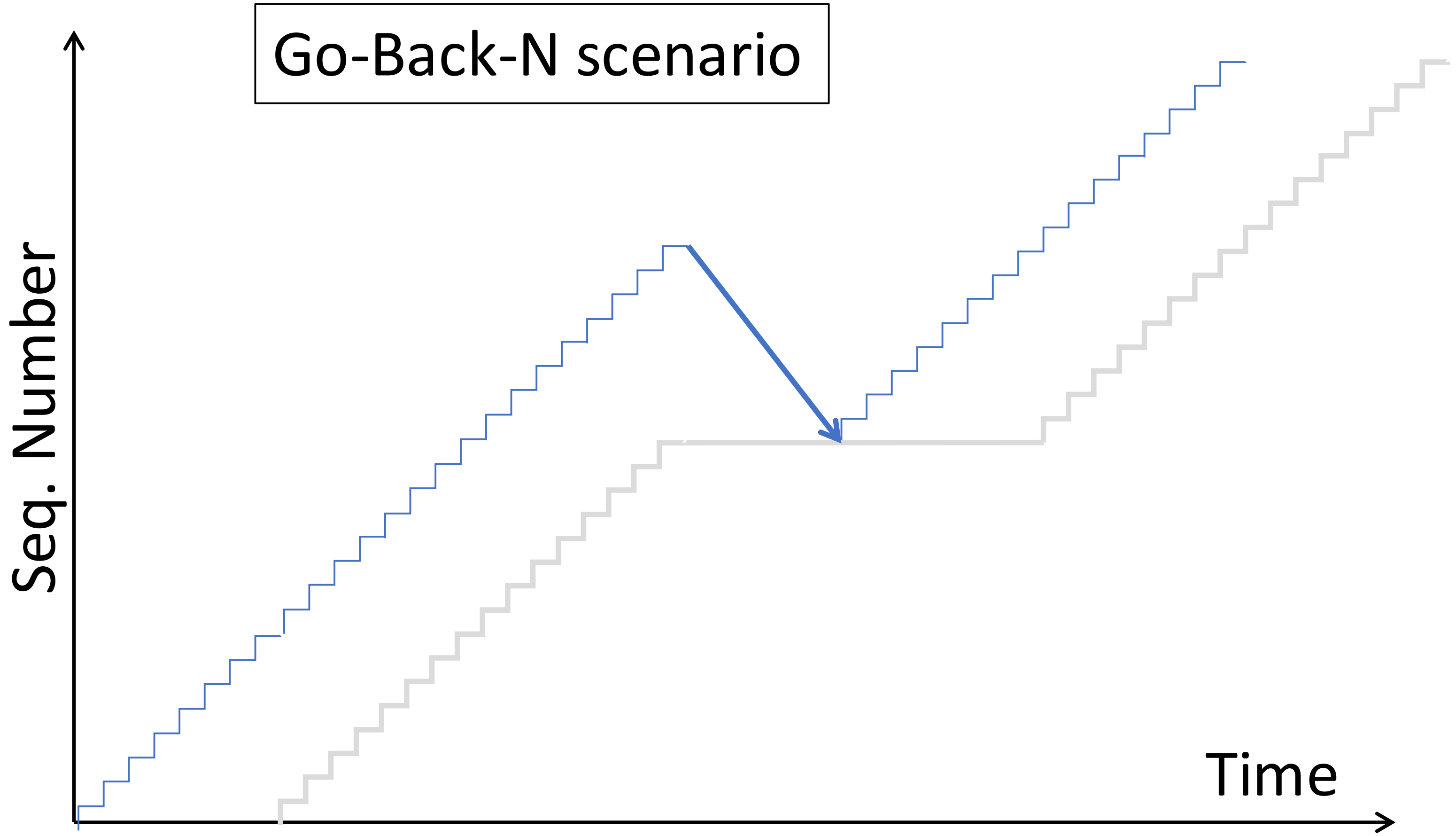
Sequence Numbers

- Need more than 0/1 for Stop-and-Wait …
  - But how many?
- For Selective Repeat, need W numbers for packets, plus W for acks of earlier packets
  - 2W seq. numbers
  - Fewer for Go-Back-N (W+1)
- Typically implement seq. number with an N-bit counter that wraps around at 2^N—1
  - E.g., N=8: …, 253, 254, 255, 0, 1, 2, 3, …

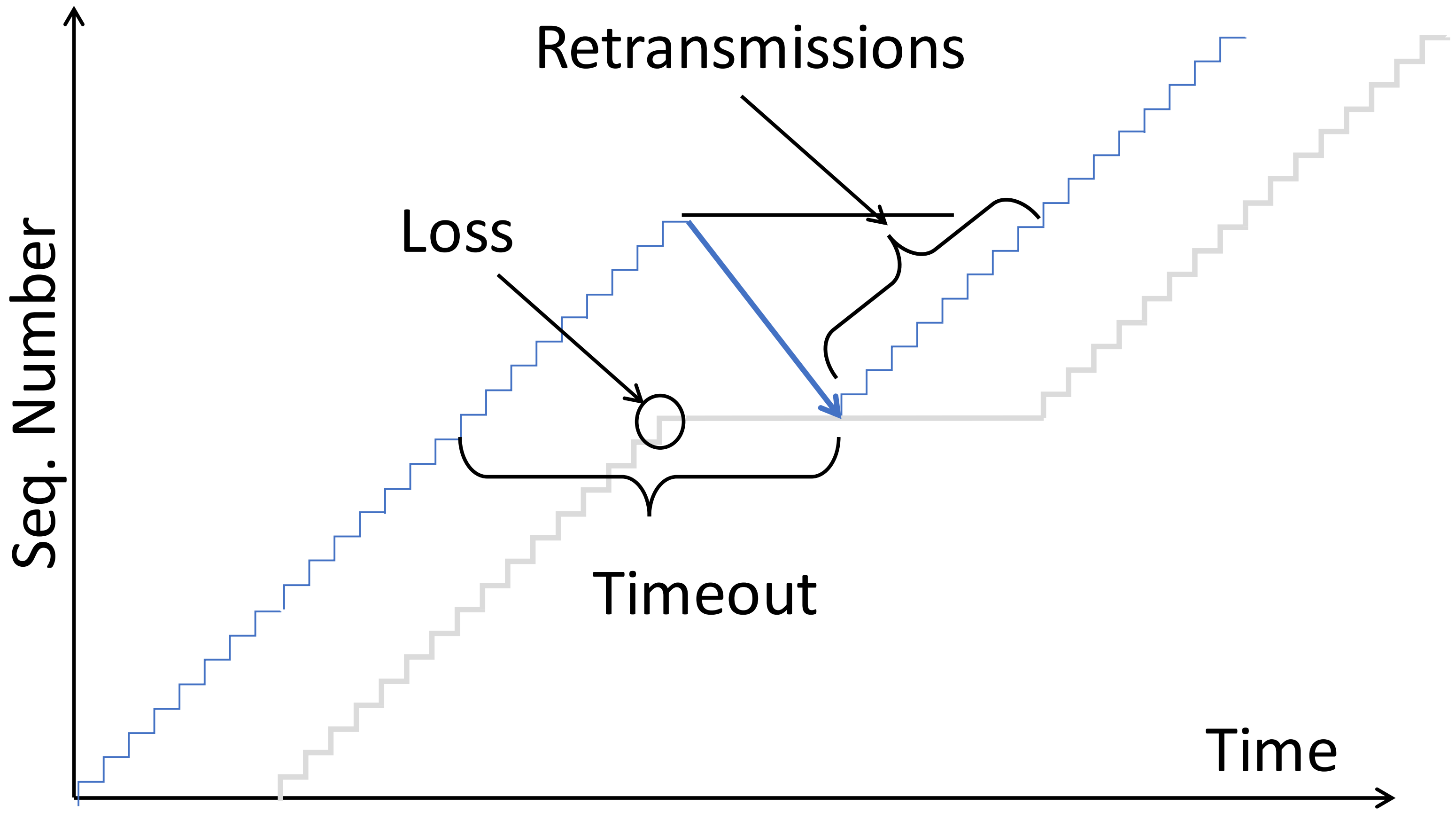
Sequence Time Plot



Sequence Time Plot (2)

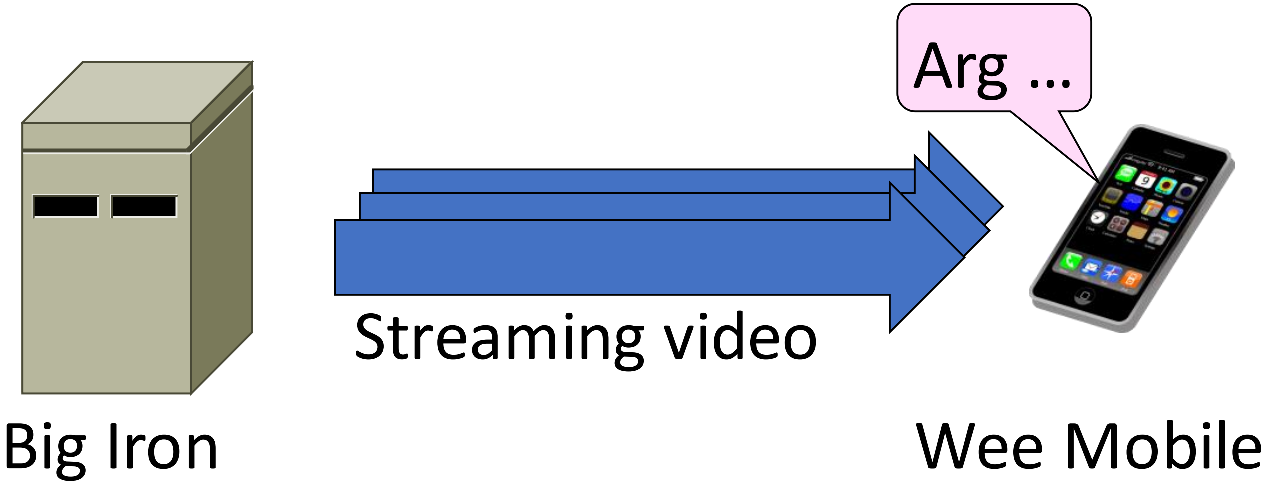


Sequence Time Plot (3)



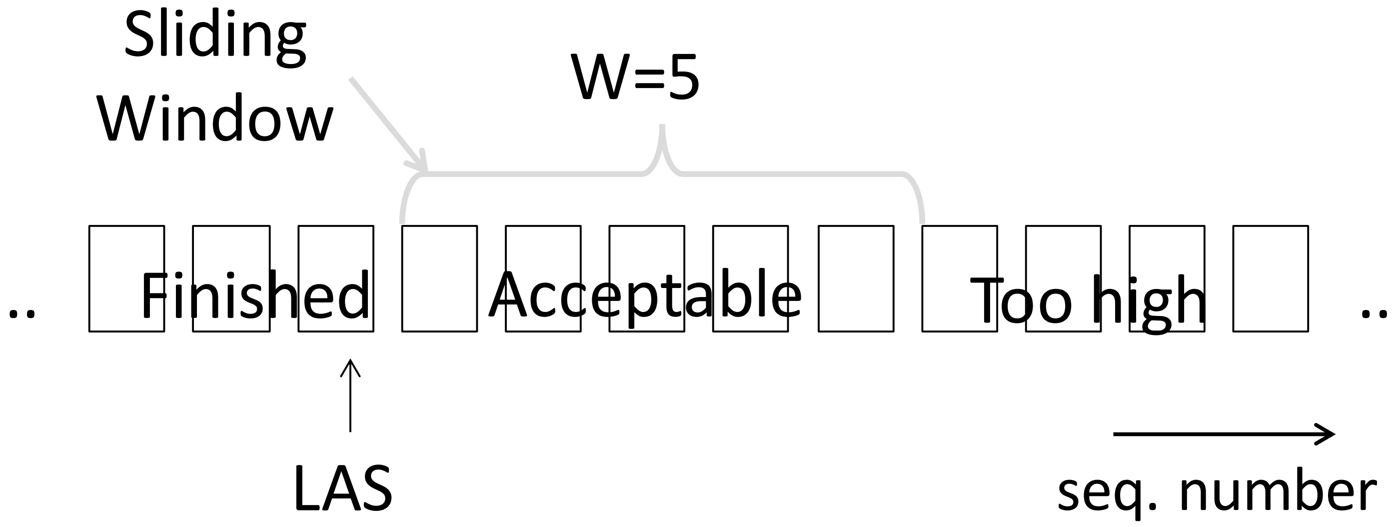
Problem

- Sliding window has pipelining to keep network busy
  - What if the receiver is overloaded?



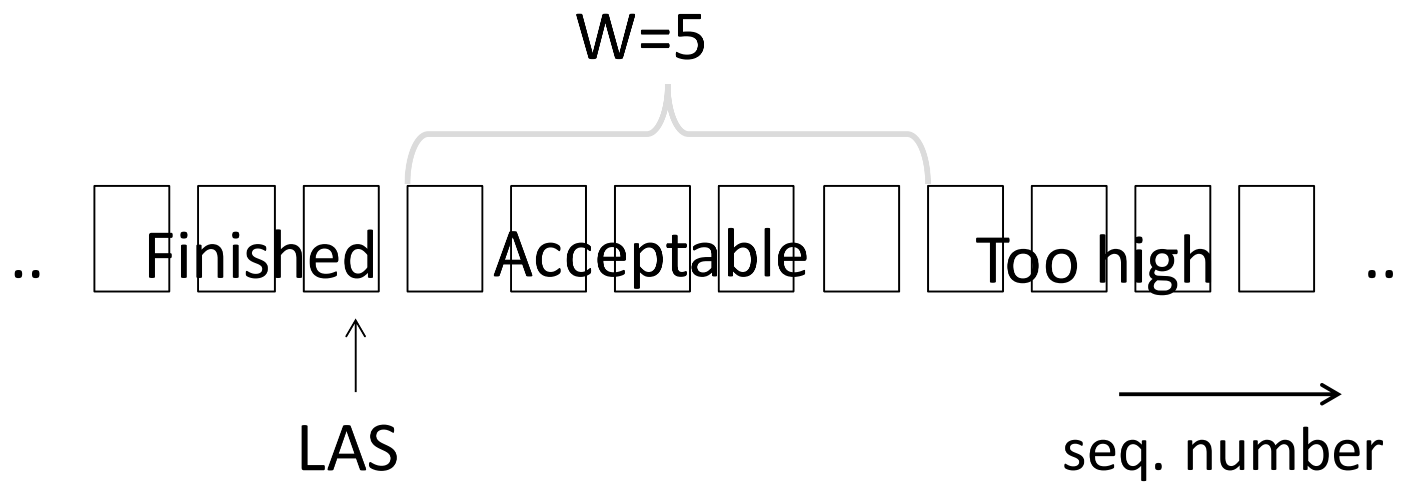
Sliding Window – Receiver

- Consider receiver with W buffers
  - LAS=LAST ACK SENT, app pulls in-order data from buffer with recv() call



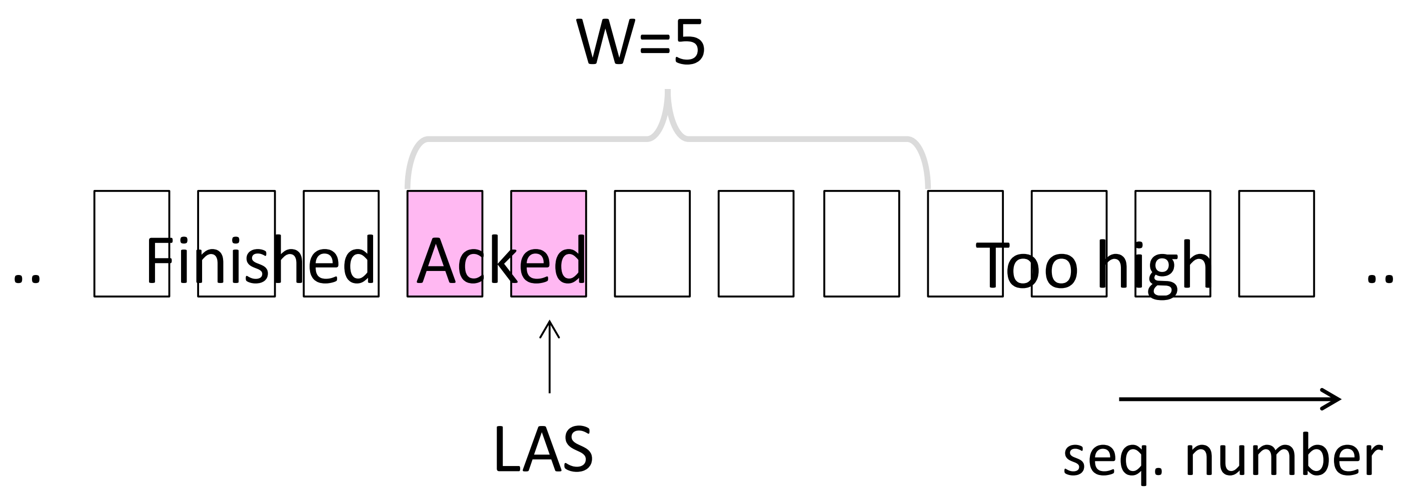
Sliding Window – Receiver (2)

- Suppose the next two segments arrive but app does not call recv()



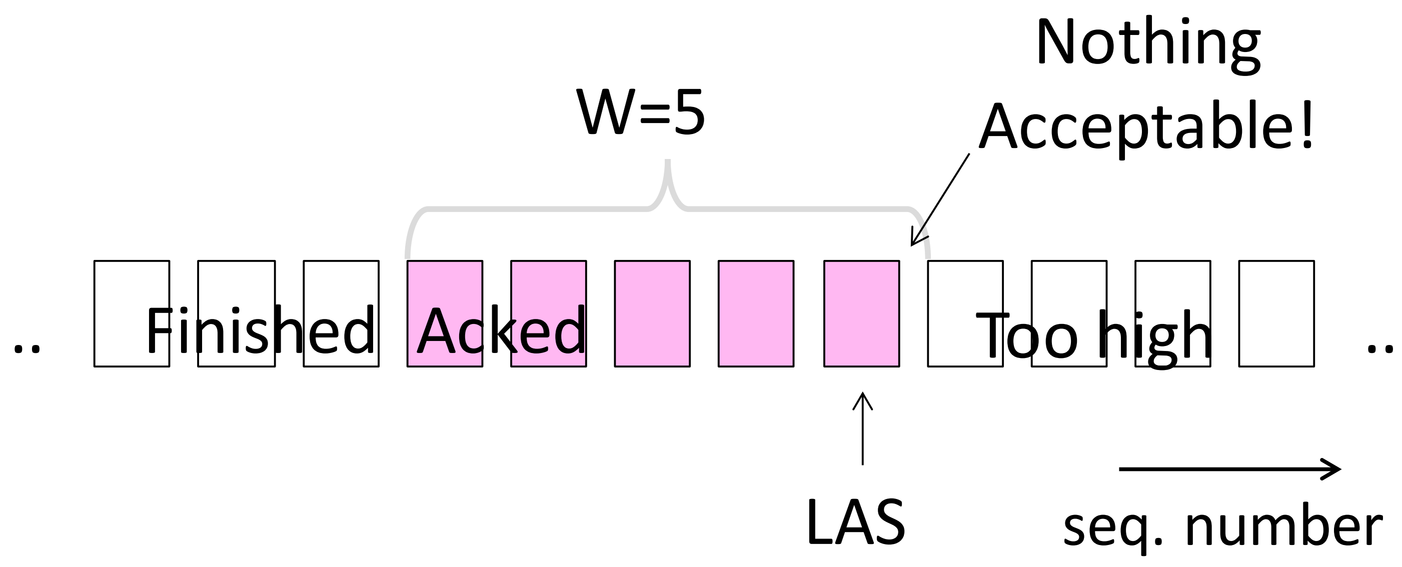
Sliding Window – Receiver (3)

- Suppose the next two segments arrive but app does not call recv()
  - LAS rises, but we can't slide window!



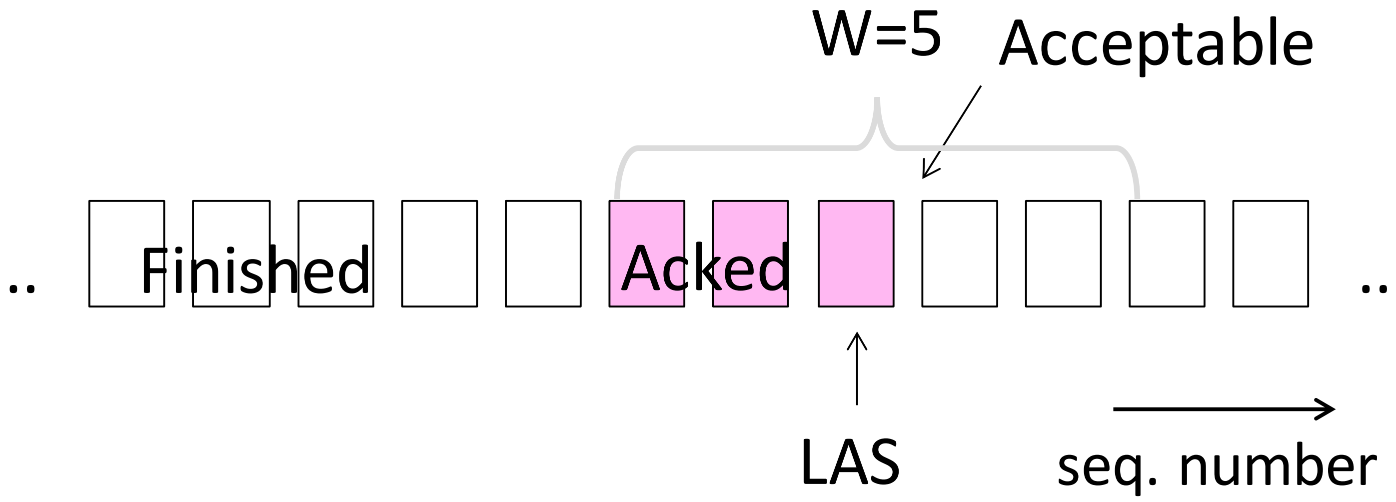
Sliding Window – Receiver (4)

- Further segments arrive (in order) we fill buffer
  - Must drop segments until app recvs!



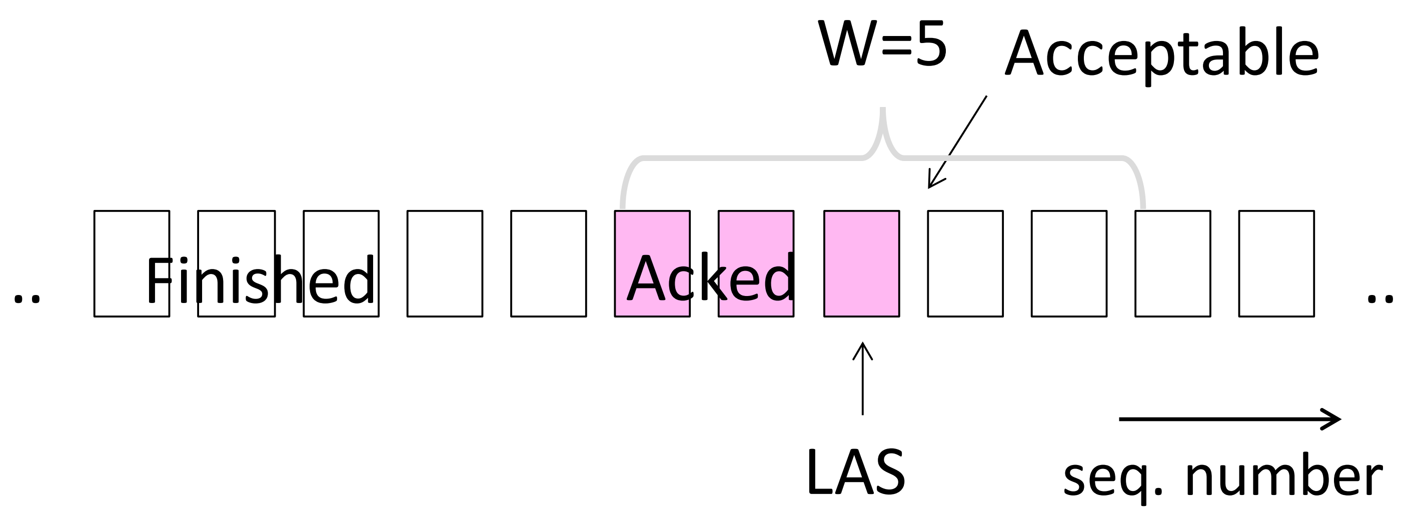
Sliding Window – Receiver (5)

- App recv() takes two segments
  - Window slides (phew)



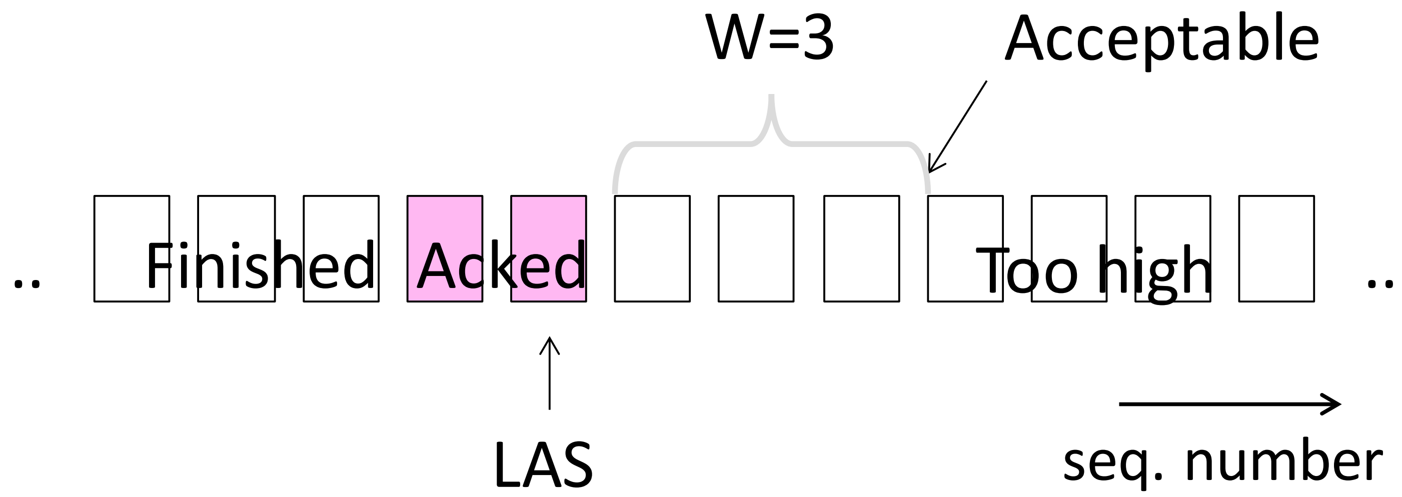
Flow Control

- Avoid loss at receiver by telling sender the available buffer space
  - WIN=#Acceptable, not W (from LAS)



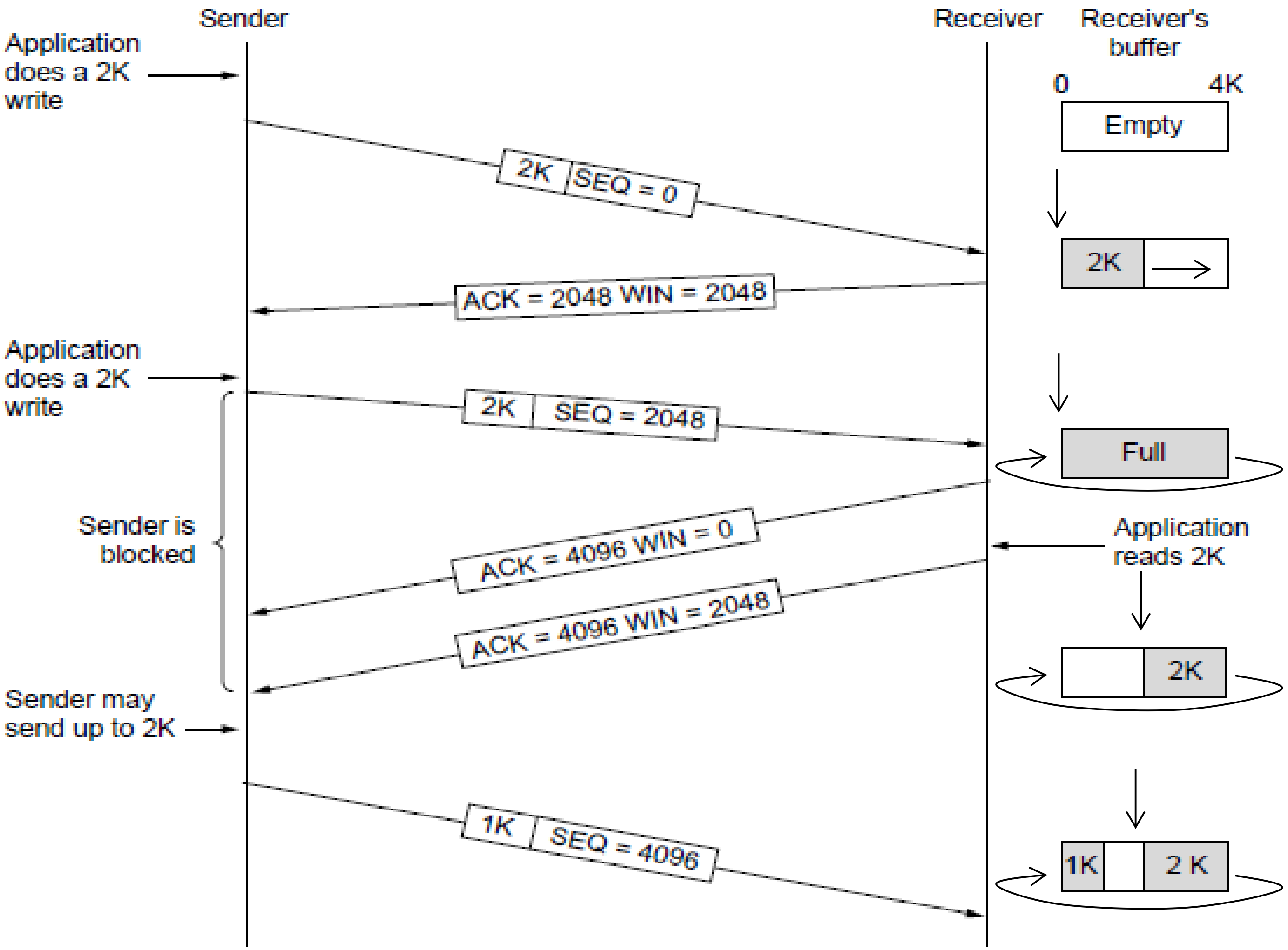
Flow Control (2)

- Sender uses lower of the sliding window and <u>flow control window</u> (WIN) as the effective window size



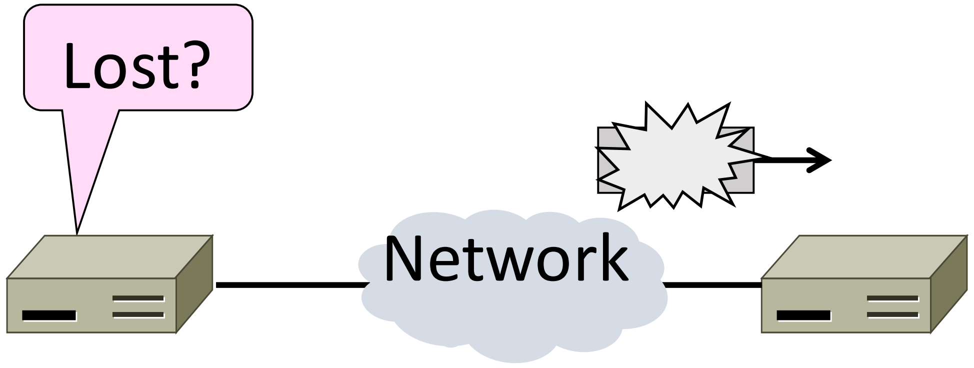
Flow Control (3)

- TCP-style example
  - SEQ/ACK sliding window
  - Flow control with WIN
  - SEQ + length < ACK+WIN
  - 4KB buffer at receiver
  - Circular buffer of bytes



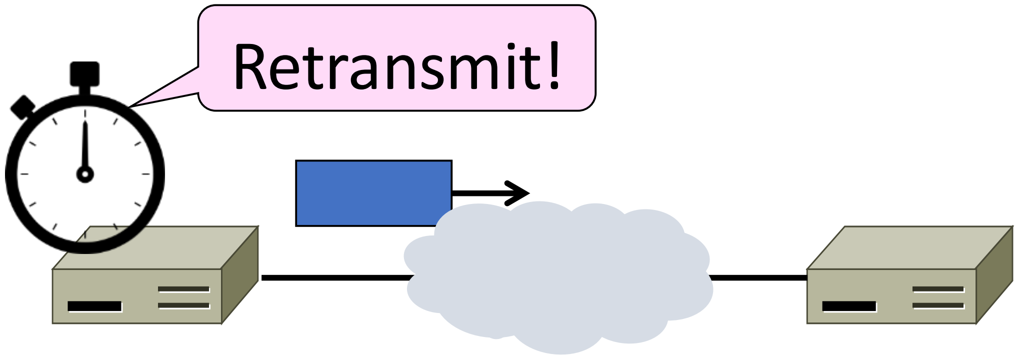
Topic

- How to set the timeout for sending a retransmission
  - Adapting to the network path



Retransmissions

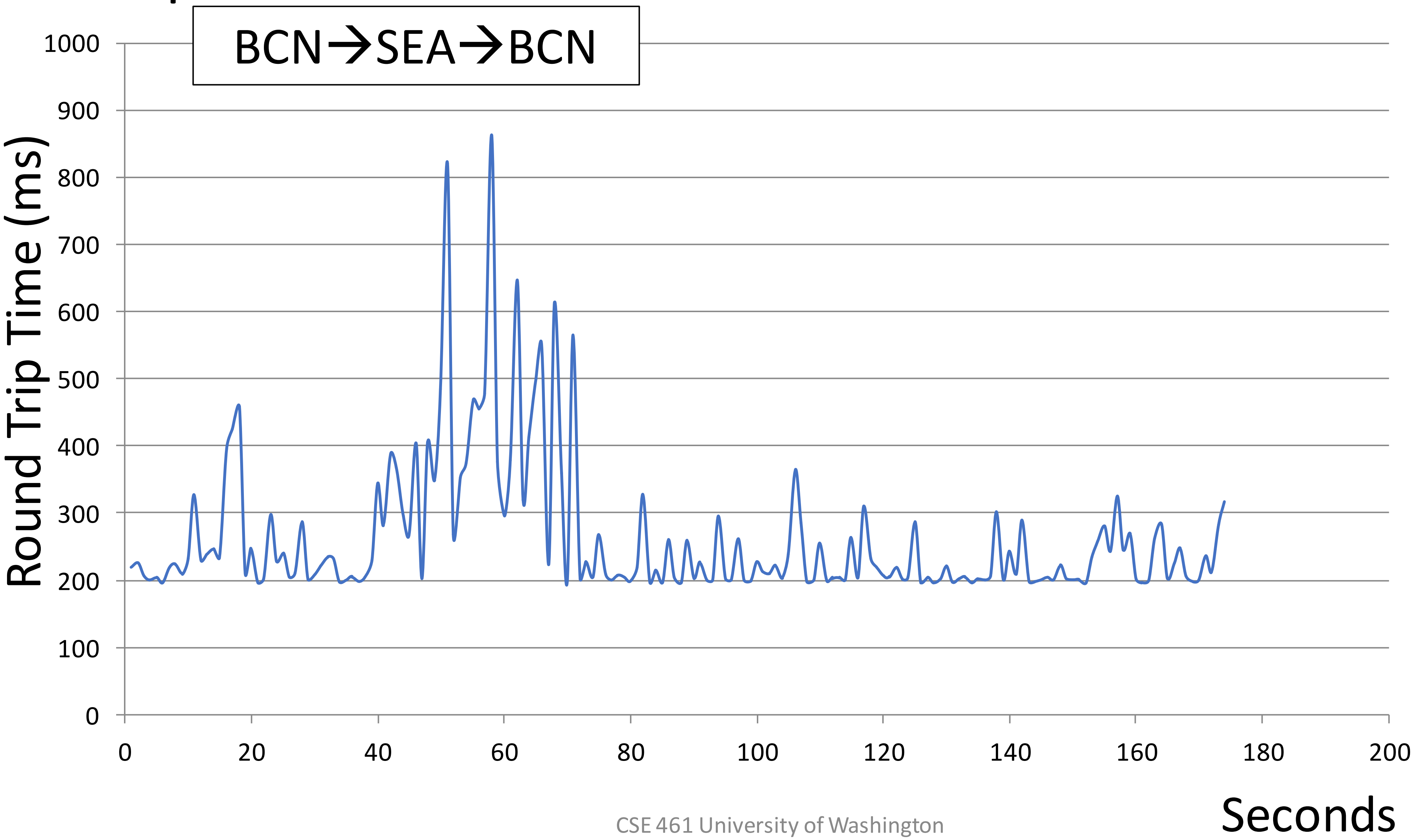
- With sliding window, detecting loss with <u>timeout</u>
  - Set timer when a segment is sent
  - Cancel timer when ack is received
  - If timer fires, <u>retransmit</u> data as lost



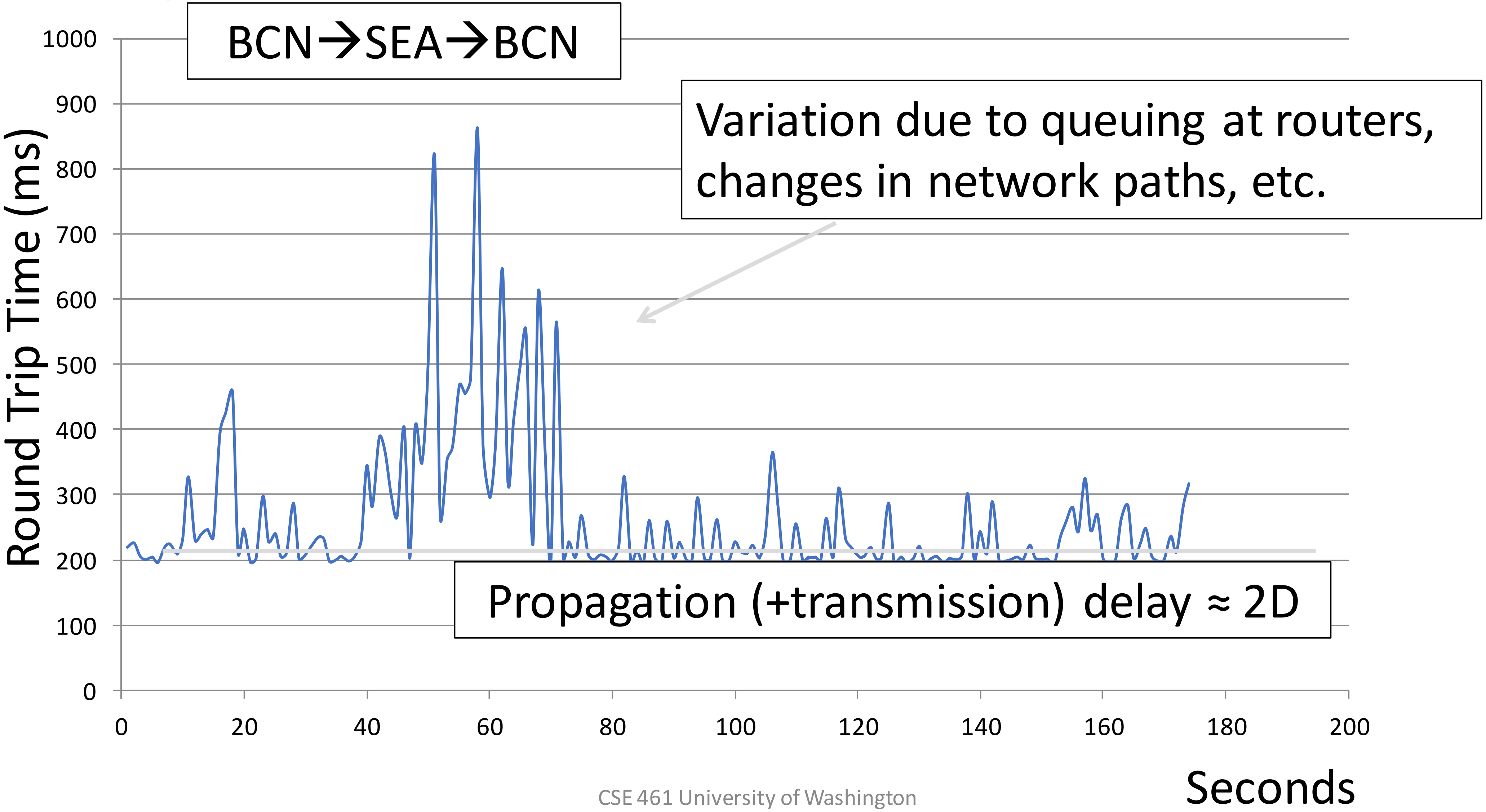
Timeout Problem

- Timeout should be "just right"
  - Too long wastes network capacity
  - Too short leads to spurious resends
  - But what is "just right"?
- Easy to set on a LAN (Link)
  - Short, fixed, predictable RTT
- Hard on the Internet (Transport)
  - Wide range, variable RTT

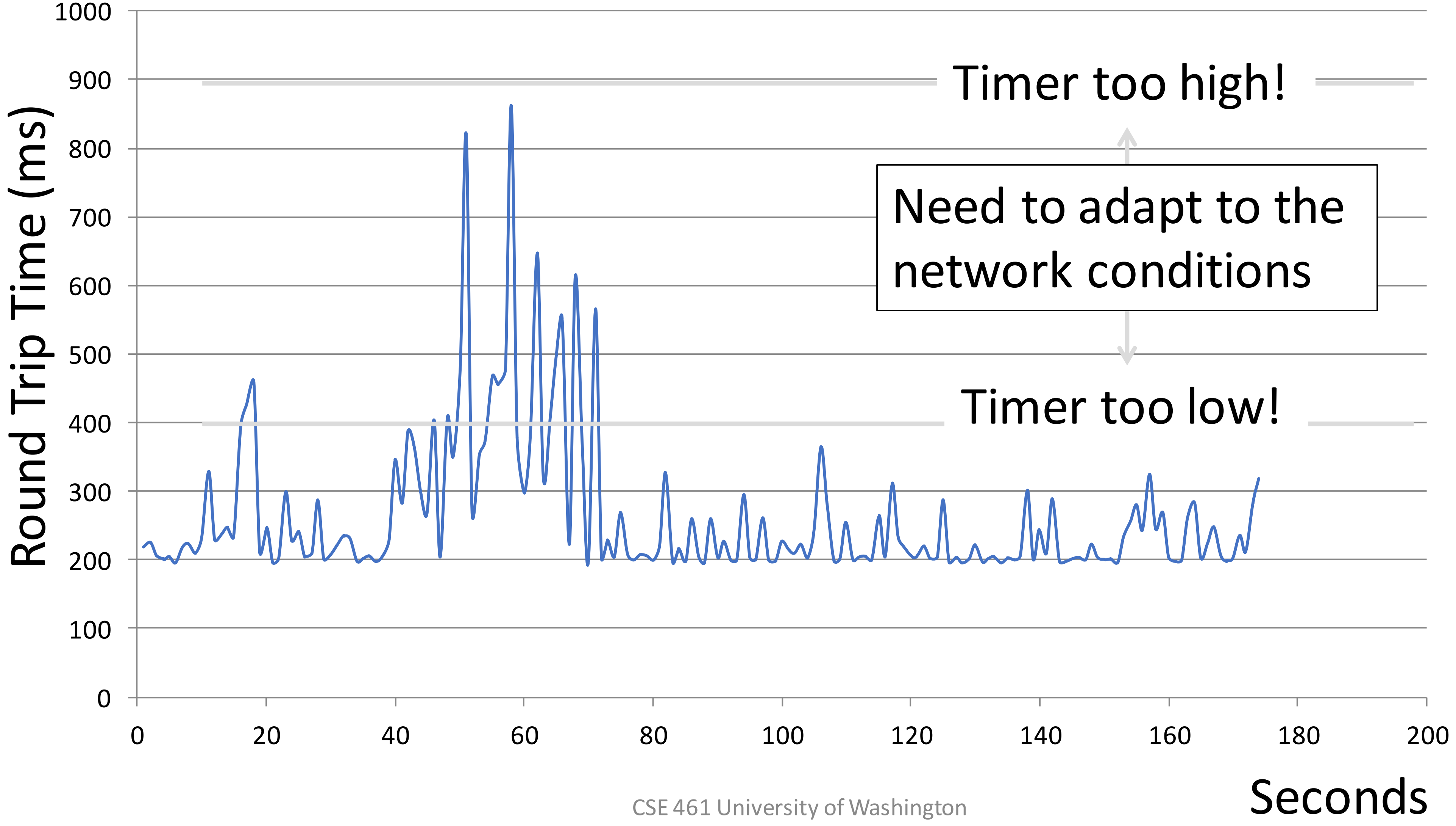
Example of RTTs



Example of RTTs (2)



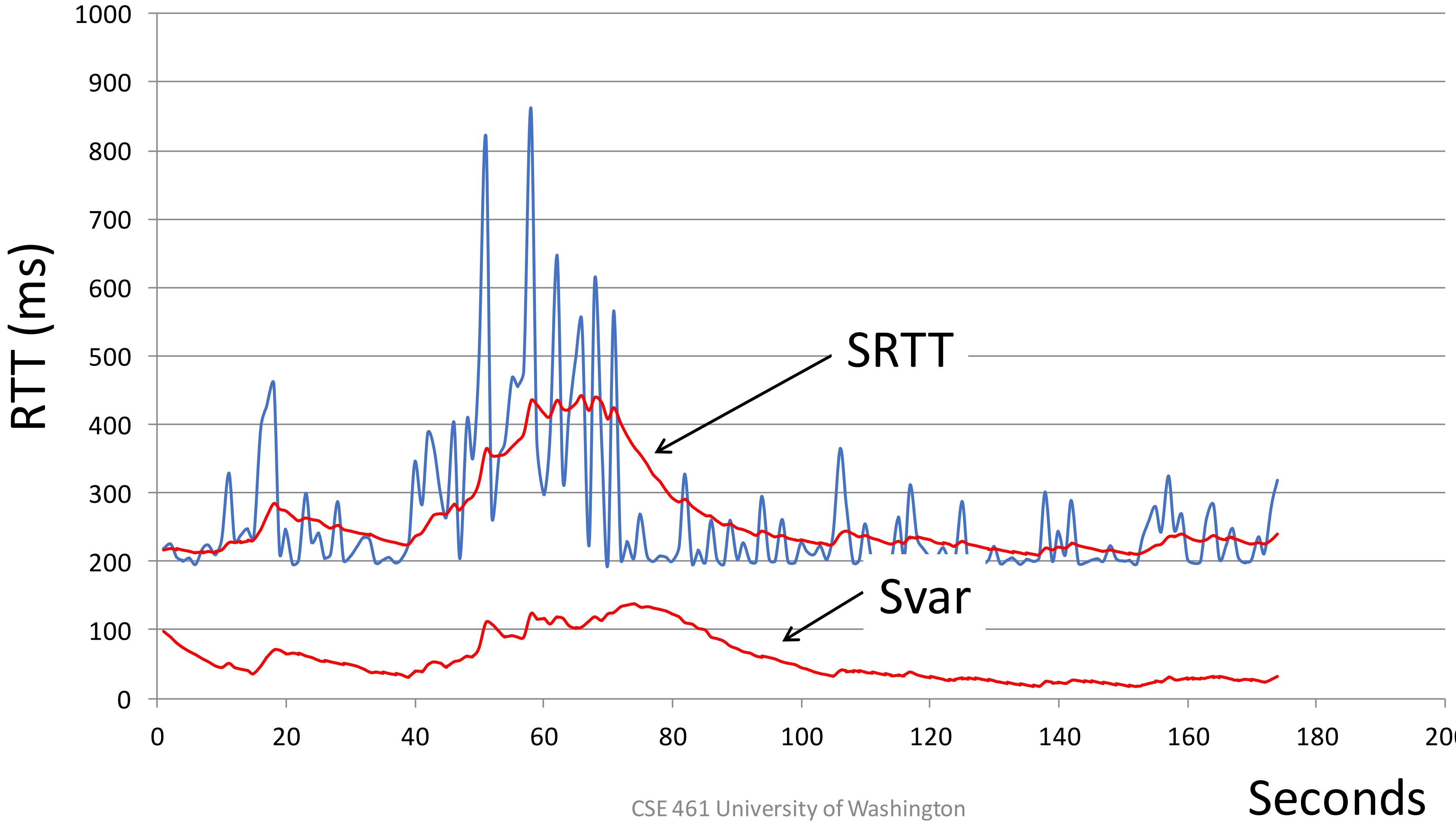
Example of RTTs (3)



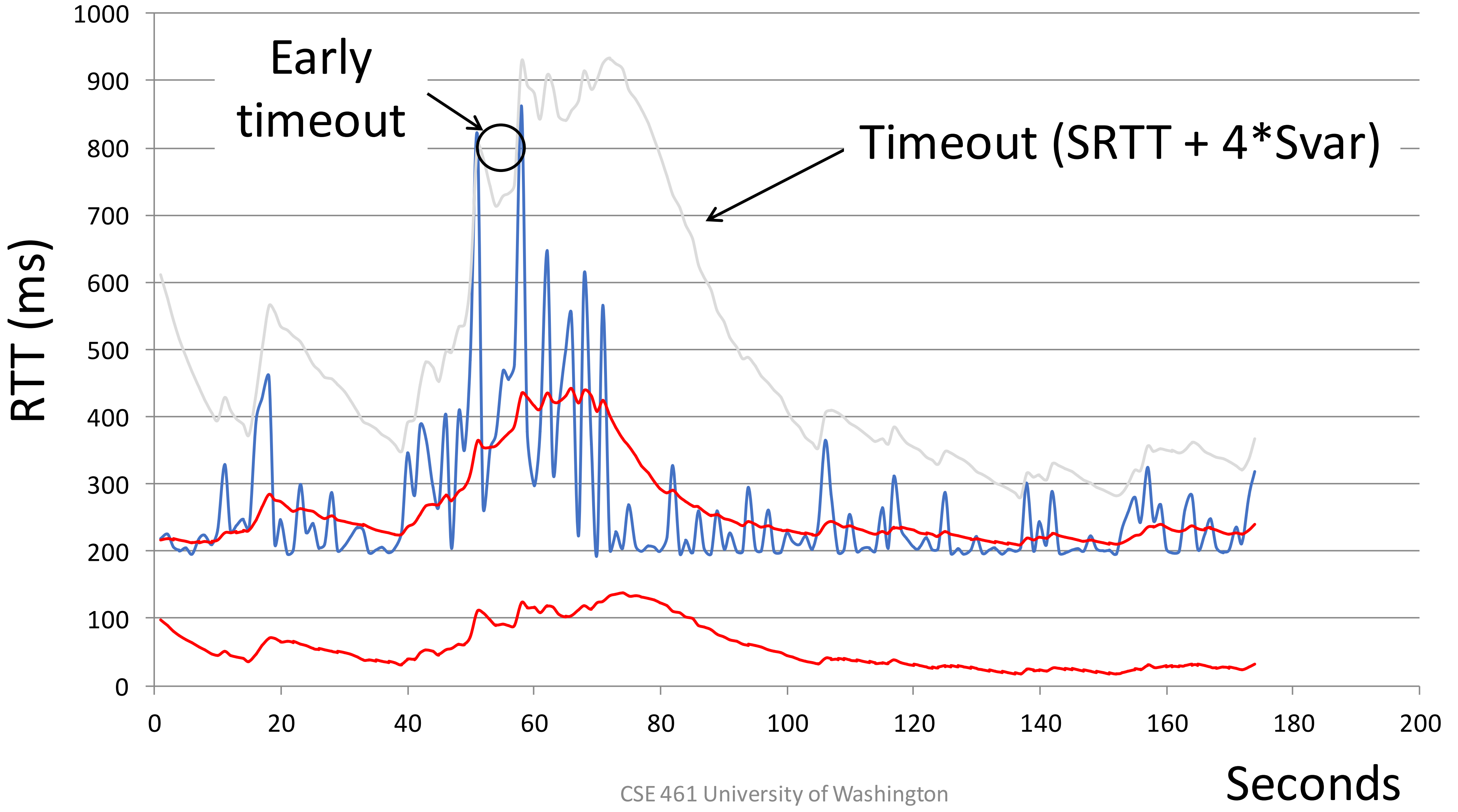
Adaptive Timeout

- Smoothed estimates of the RTT (1) and variance in RTT (2)
  - Update estimates with a moving average
  1. $SRTT_{N+1} = 0.9*SRTT_N + 0.1*RTT_{N+1}$
  2. $Svar_{N+1} = 0.9*Svar_N + 0.1*|RTT_{N+1} - SRTT_{N+1}|$
- Set timeout to a multiple of estimates
  - To estimate the upper RTT in practice
  - TCP $Timeout_N = SRTT_N + 4*Svar_N$

Example of Adaptive Timeout



Example of Adaptive Timeout (2)



Adaptive Timeout (2)

- Simple to compute, does a good job of tracking actual RTT
  - Little "headroom" to lower
  - Yet very few early timeouts
- Turns out to be important for good performance and robustness