15-441

Computer Networking

TCP Connection Management, Error Control

Nov. 1, 2004

Slides – Randy Bryant, Hui Zhang, Ion Stoica, Dave Eckhardt

(Possible) Transport Protocol Functions

Multiplexing/demultiplexing for multiple applications.

- “Port” abstraction abstracts OS notions of “process”

Connection establishment.

- Logical end-to-end connection
- Connection state to optimize performance

Error control.

- Hide unreliability of the network layer from applications
- Many types of errors: corruption, loss, duplication, reordering.

End-to-end flow control.

- Avoid flooding the receiver

Congestion control.

- Avoid flooding the network

Outline

Connection management

- What's a connection?
- How do we get one?
- Why so complicated?
  - Threats

Error control, Flow control

- Stop & Wait vs. sliding window (conceptual and TCP)
- Ack flavors, windows, timeouts, sequence numbers

Next Lecture – Dave Maltz. Mobility

Monday – TCP again

- Congestion control – you will not address in Project 3

Transmission Control Protocol (TCP)

Reliable bi-directional byte stream

Connections established & torn down

- Analogy: setting up & terminating phone call

Multiplexing/ demultiplexing

- Ports at both ends

Error control

- Users see correct, ordered byte sequences

End-end flow control

- Avoid overwhelming machines at each end

Congestion avoidance

- Avoid creating traffic jams within network

Source Port	Dest. Port
Data Sequence #	
Acknowledgment Sequence #	
HL/Flags	Window
D. Checksum	Urgent Pointer
Options..	

TCP Flags

SYN: Synchronize

- Used when setting up connection

FIN: Finish

- Used when tearing down connection

RESET

- I'm lost. Need to abort connection

PUSH

- Signal the receiving application that data is ready

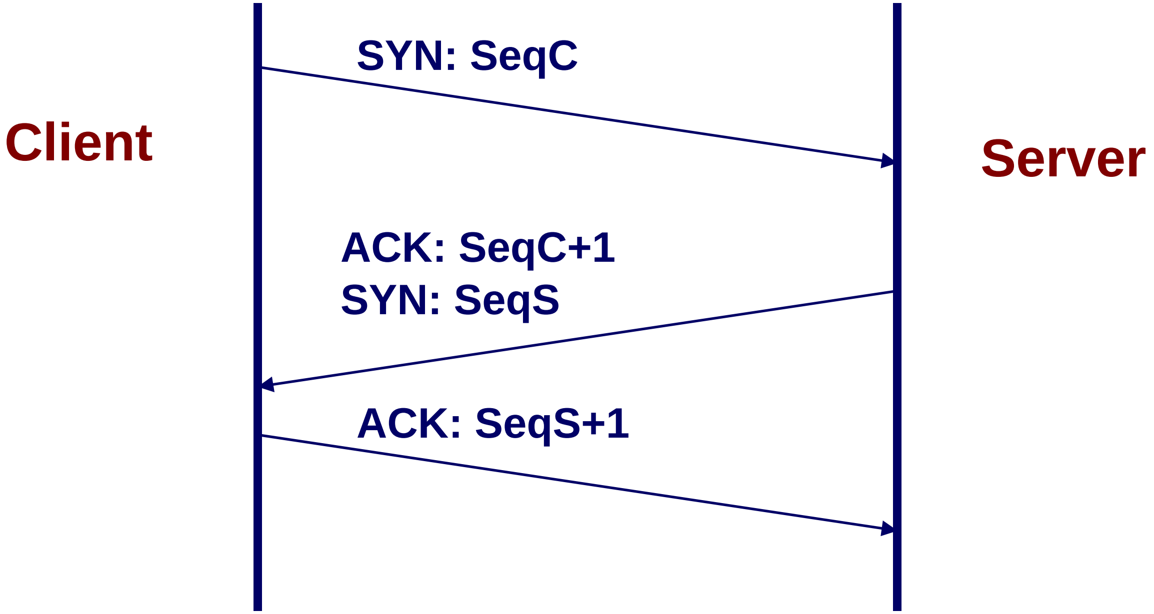
URG: Urgent

- Segment includes “urgent” data

ACK

- Acknowledging received data

Establishing Connection



Three-Way Handshake

- Each side notifies other of starting sequence number it will use for sending
- Each side acknowledges other's sequence number
  - SYN-ACK: Acknowledge sequence number + 1
- Can combine second SYN with first ACK

TCP Session Example

Use windump to trace typical TCP session

Client

- 128.2.222.198:3123
- Randy Bryant's laptop BRYANT-TP2.VLSI using ephemeral port

Server

- 192.216.219.96:80
- Web server at ceiva.com

Task

- Upload digital image to server

TCP Connection Setup Example

```
09:23:33.042318 IP 128.2.222.198.3123 > 192.216.219.96.80: S
  4019802004:4019802004(0) win 65535 <mss 1260,nop,nop,sackOK> (DF)

09:23:33.118329 IP 192.216.219.96.80 > 128.2.222.198.3123: S
  3428951569:3428951569(0) ack 4019802005 win 5840 <mss
1460,nop,nop,sackOK> (DF)

09:23:33.118405 IP 128.2.222.198.3123 > 192.216.219.96.80: . ack
  3428951570 win 65535 (DF)
```

Client SYN

- **SeqC: Seq. #4019802004, window 65535, max. seg. 1260**

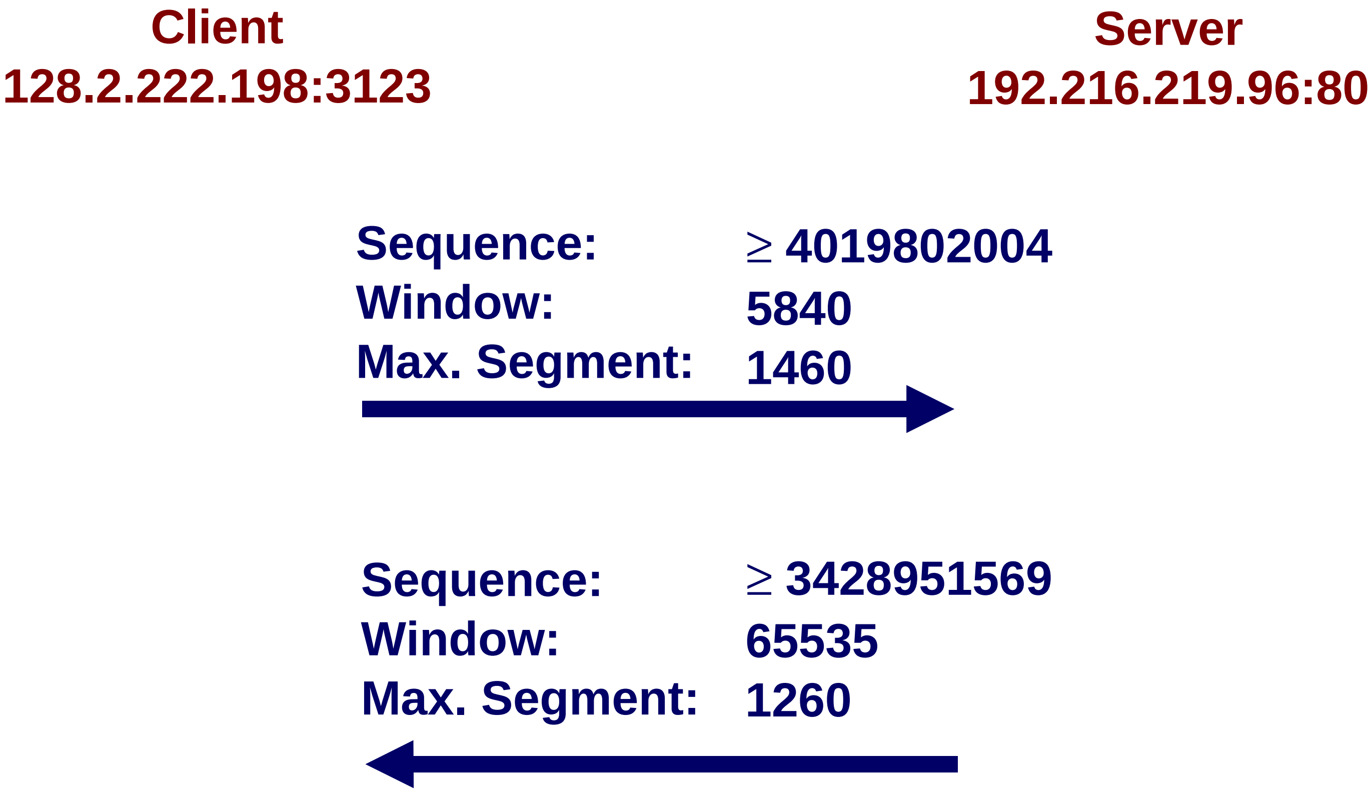
Server SYN-ACK+SYN

- **Receive: #4019802005 (= SeqC+1)**
- **SeqS: Seq. #3428951569, window 5840, max. seg. 1460**

Client SYN-ACK

- **Receive: #3428951570 (= SeqS+1)**

Connection Created



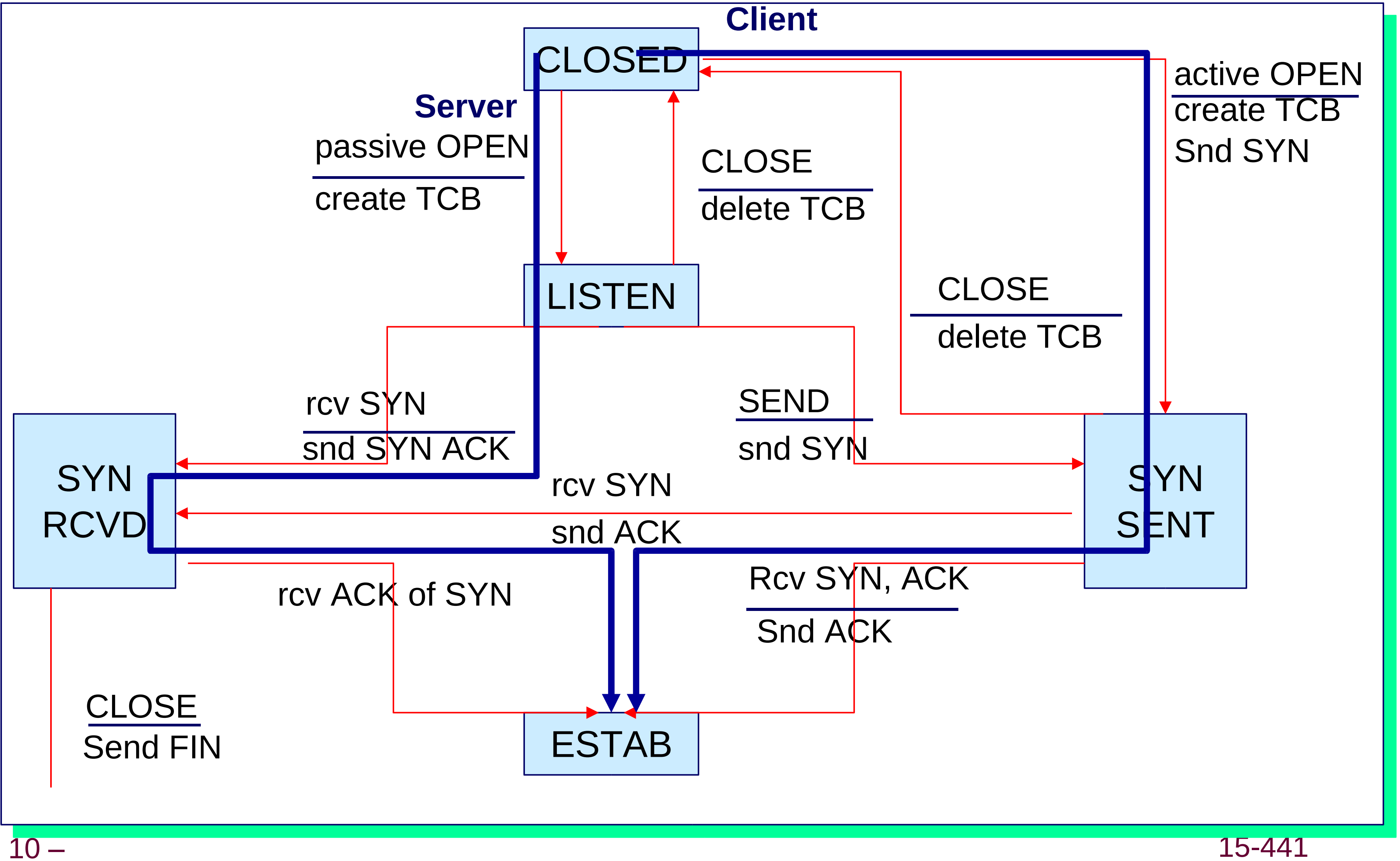
Client
128.2.222.198:3123

Server
192.216.219.96:80

Sequence: ≥ 4019802004
Window: 5840
Max. Segment: 1460

Sequence: ≥ 3428951569
Window: 65535
Max. Segment: 1260

TCP State Diagram: Connection Setup



Handshake – Why So Complicated?

Both sides specify a 32-bit sequence number

- Why can't they just both start with zero?

Recall IP's TTL field

- TTL Max = 255
- Originally expected to be 255 *seconds*!
- Reinterpreted to be 255 hops
- What happens if a *really* old packet arrives?
  - Old connection: IP_1, $Port_1$, IP_2, $Port_2$, [Seq_1], [Seq_2]
  - Which of those will be the same for a new connection?
  - Can you guess how sequence numbers should be chosen?

Error Control – Threats

Network may corrupt frames

- Despite link-level checksum
- Despite switch/router memory ECC
- Example
  - Store packet headers in separate memory from packet bodies
  - Maintain association between header #343 and body #343
    - Most of the time...

Packet-sequencing issues

- Network may duplicate packets (really?)
- Network may re-order packets (why?)
- Network may lose packets (often, actually)

Error Control

Add end-to-end checksum to TCP segments

Packet sequencing problems: per-segment sequence #

- Duplicate: ignore
- Reordered: reorder or drop
- Lost: retransmit

Lost segments detected by sender.

- Receiver won't ACK a lost segment
- Use timeout to detect lack of acknowledgment
- Need estimate of the roundtrip time to set timeout

Retransmission requires sender to keep copy of data.

- Copy is discarded when ACK is received

Error Control Algorithms

Use two basic techniques:

- Acknowledgements (ACKs)
- Timeouts

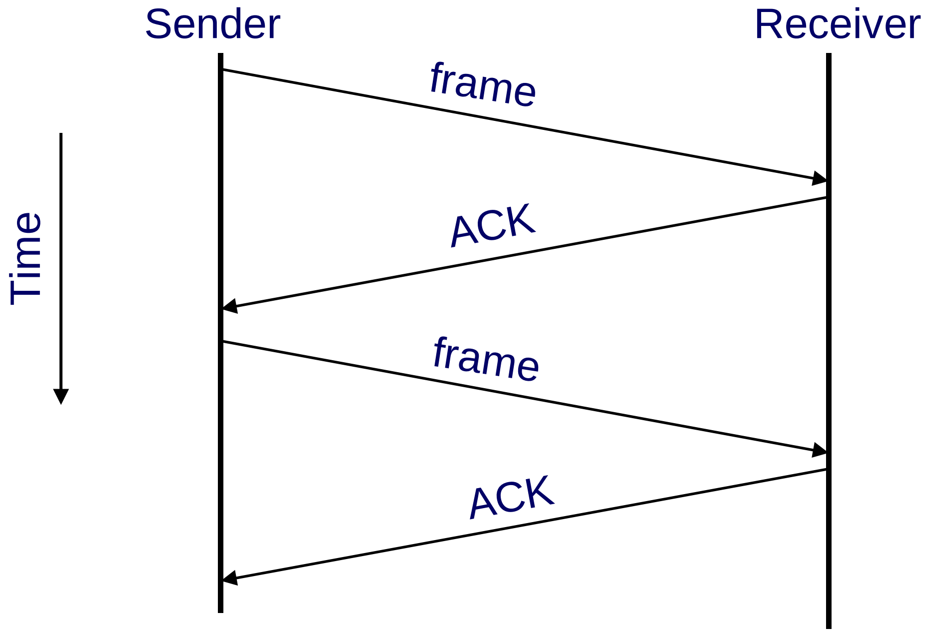
Two examples:

- Stop-and-wait
- Sliding window

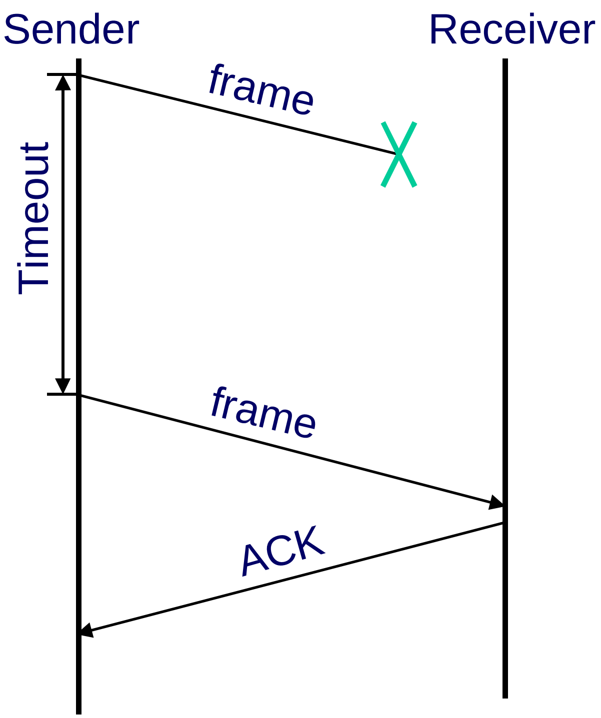
Stop-and-Wait

Receiver: send an acknowledge (ACK) back to the sender upon receiving a packet (frame)

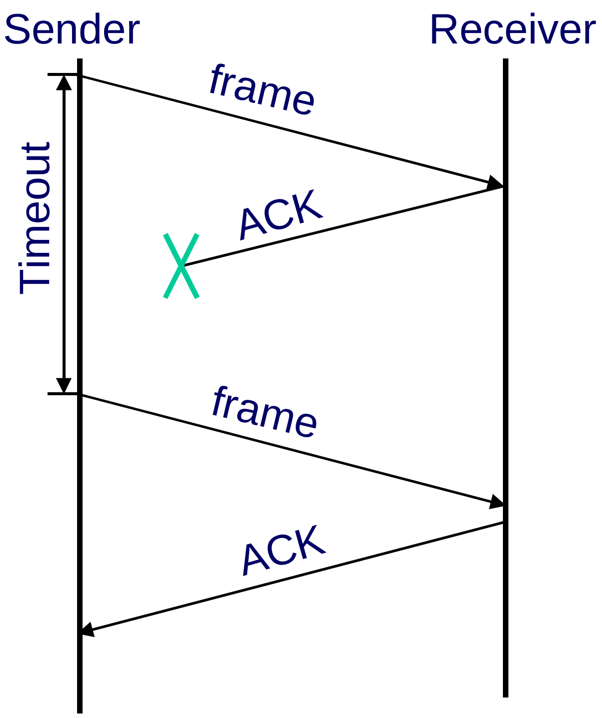
Sender: excepting first packet, send a packet only upon receiving the ACK for the previous packet



What Can Go Wrong?

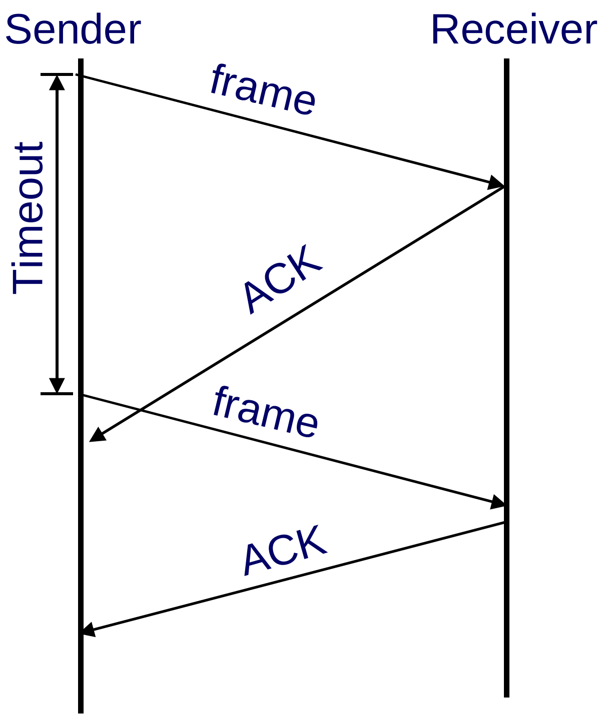


Frame lost - resend it on Timeout



ACK lost - resend packet

Need a mechanism to detect duplicate packet



ACK delayed – resend packet

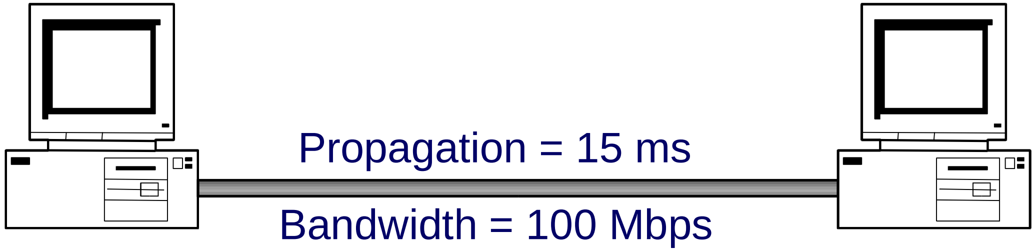
Need a way to differentiate between ACK for current and previous packet – one bit often enough

Stop-and-Wait Disadvantage

May lead to inefficient link utilization

Example

- One-way propagation = 15 ms
- Bandwidth = 100 Mbps
- Packet size = 1000 bytes: transmit = $(8*1000)/10^8$ = 0.08ms
- Neglect queue delay: Latency = approx. 15 ms; RTT = 30 ms



Propagation = 15 ms

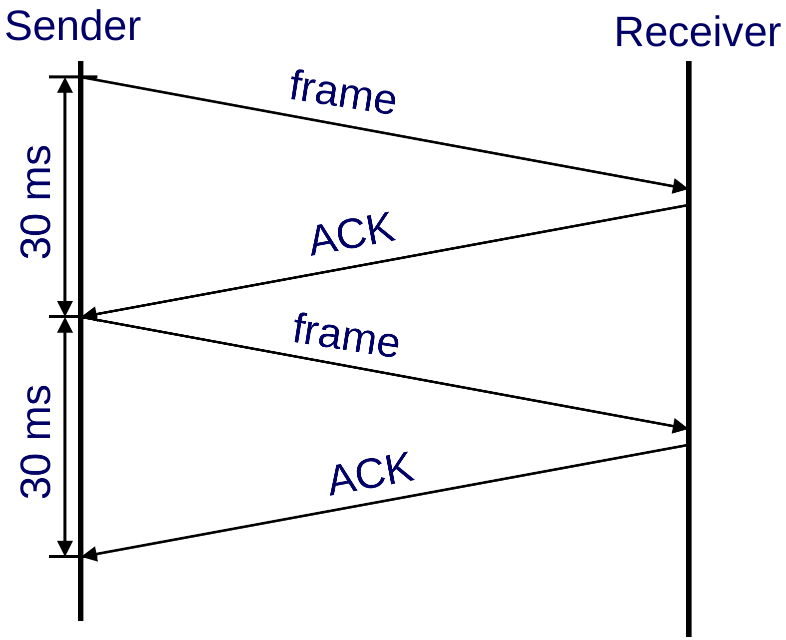
Bandwidth = 100 Mbps

Stop-and-Wait Disadvantage (cont'd)

Send a message every 30 ms

- **Throughput = (8*1000)/0.03 = 0.2666 Mbps**

Thus, the protocol uses less than 0.3% of the link capacity!



Solution

Don’t wait for the ACK of the previous packet before sending the next packet!

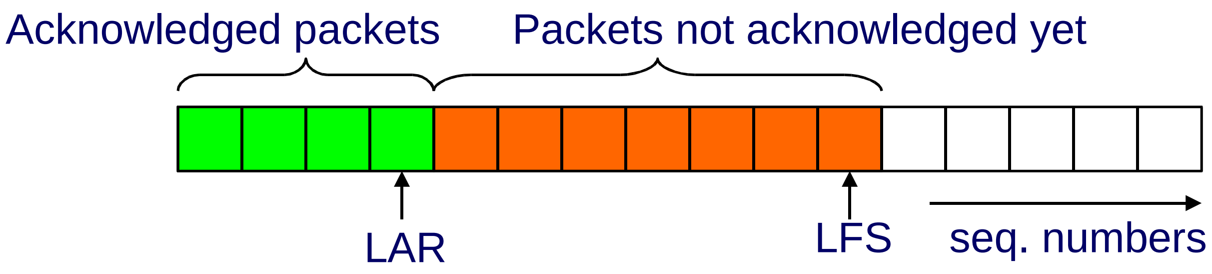
Sliding Window Protocol: Sender

Each packet has a sequence number

- Assume infinite sequence numbers for simplicity

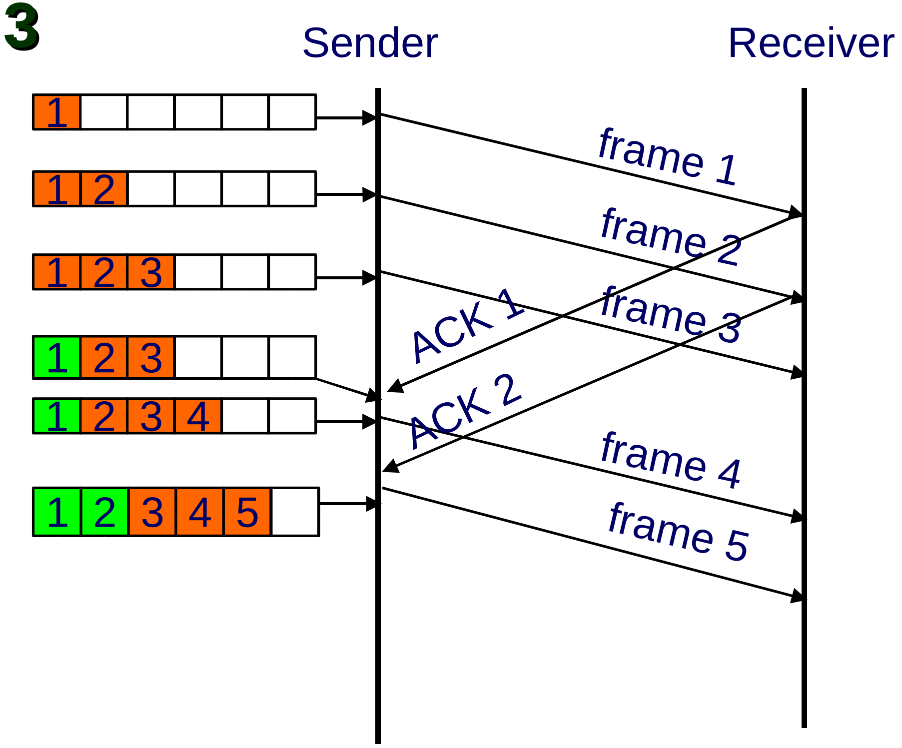
Sender maintains a window of sequence numbers

- SWS (sender window size) – maximum number of packets that can be sent without receiving an ACK
- LAR (last ACK received)
- LFS (last frame sent)



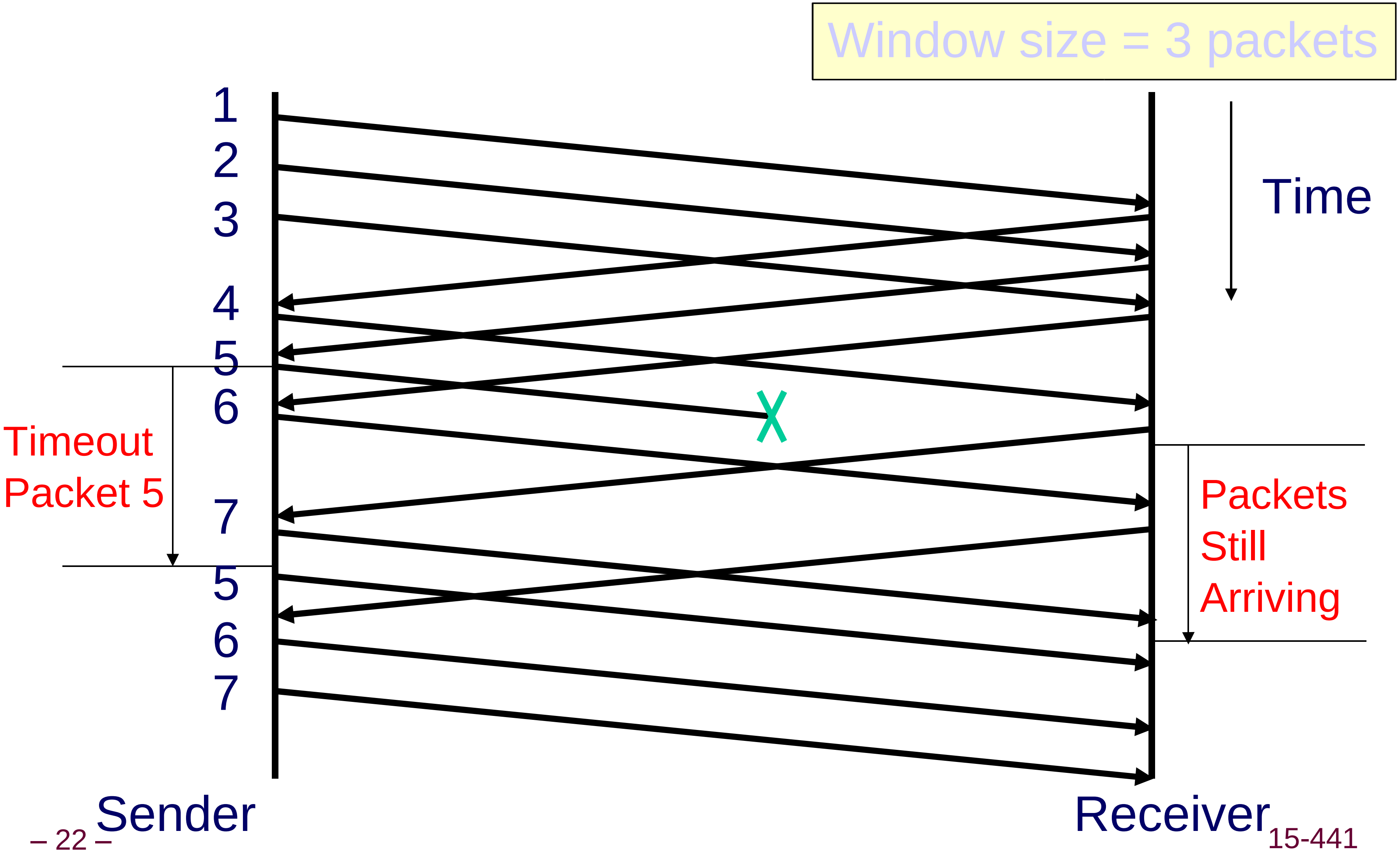
Example

Assume SWS = 3



Note: usually ACK contains the sequence number of the first packet in sequence expected by receiver

Need for Receiver Window



Window size = 3 packets

Sliding Window Protocol: Receiver

Receiver maintains a window of sequence numbers

- **RWS (receiver window size) – maximum number of out-of-sequence packets that can received**
- **LFR (last frame received) – last frame received in sequence**
- **LAF (last acceptable frame)**
- **LAF – LFR <= RWS**

Sliding Window Protocol: Receiver

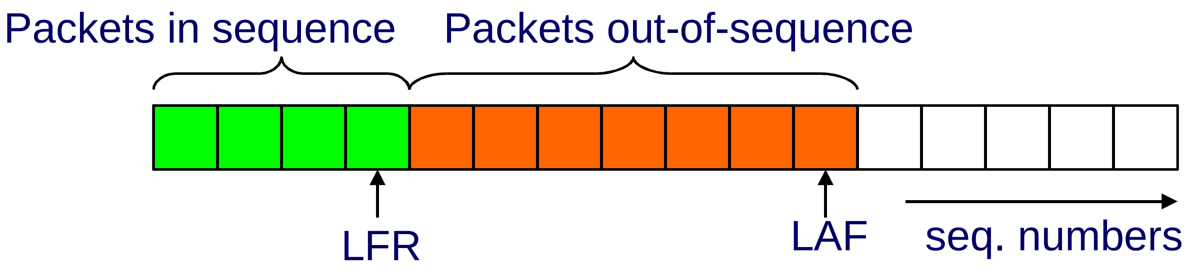
Let seqNum be the sequence number of arriving packet

If (seqNum <= LFR) or (seqNum >= LAF)

- **Discard packet**

Else

- **Accept packet**
- **ACK largest sequence number seqNumToAck, such that all packets with sequence numbers <= seqNumToAck were received**



Choices of Ack

Cumulative ack

- I have received 17..23
- I have [still] received 17..23

Selective ack

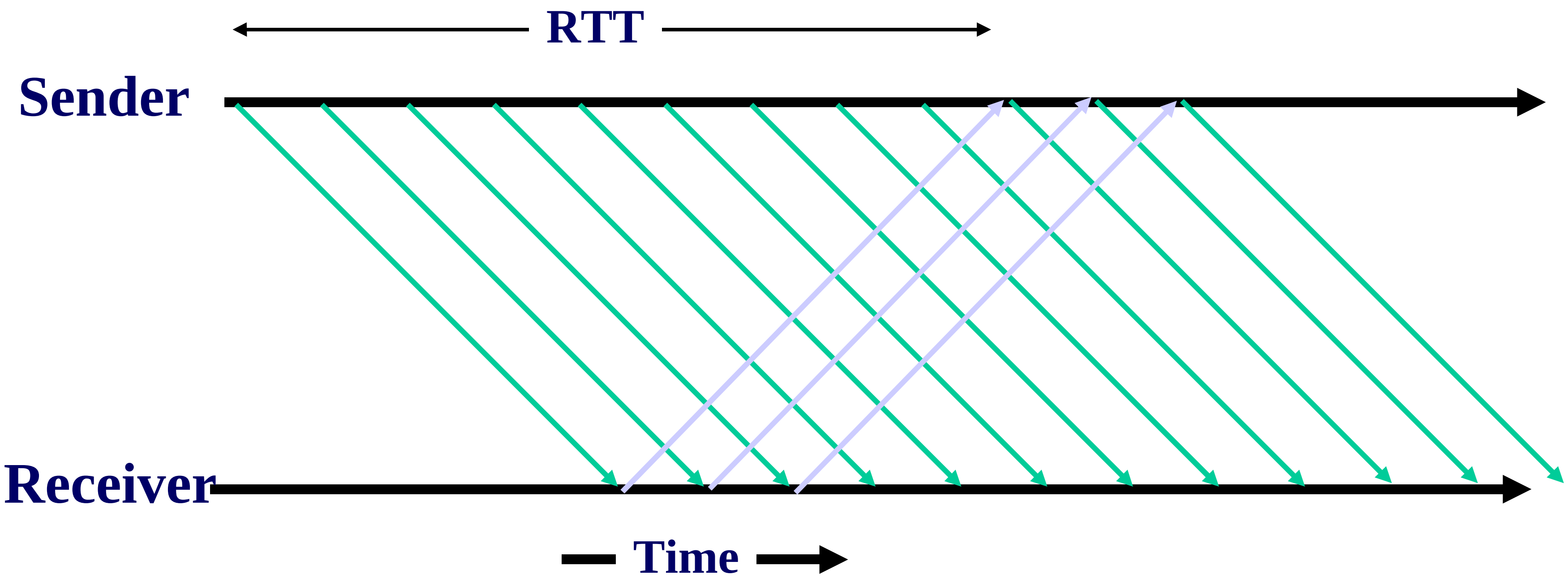
- I received 17-23, 25-27

Negative ack

- I think I'm missing 24...

Tradeoffs?

Choosing Window Size



$$\text{Max Throughput} = \frac{\text{Window Size}}{\text{Roundtrip Time}}$$

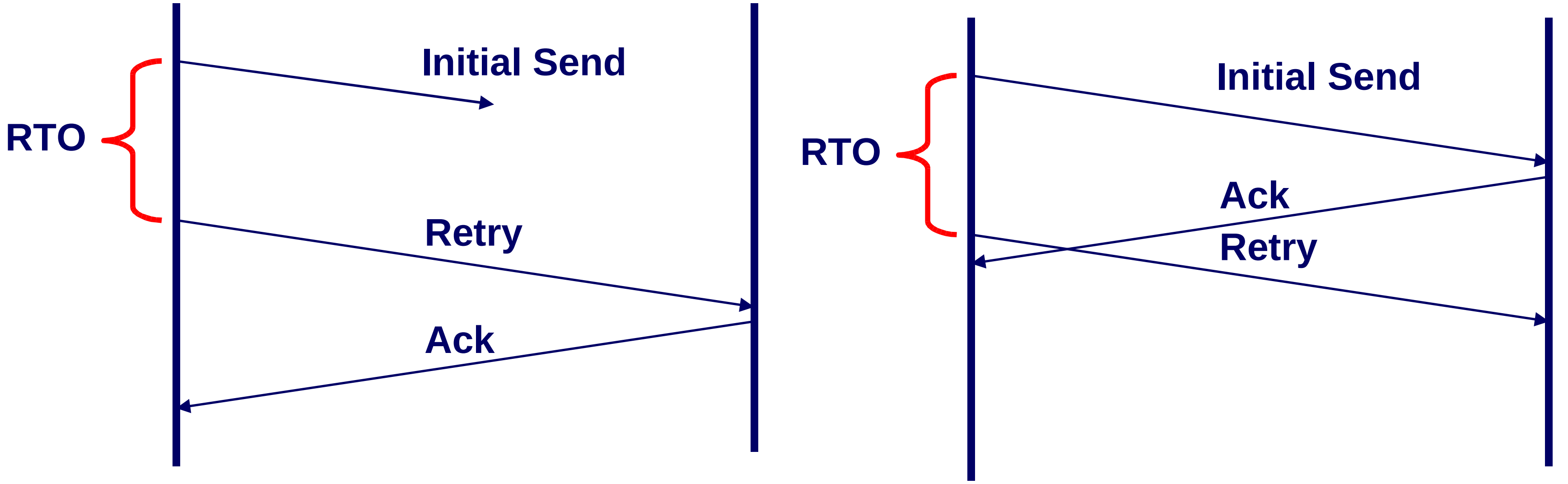
Timeout Value Selection

Long timeout?

Short timeout?

Solution?

Setting Retransmission Timeout (RTO)



Detect dropped packet

RTO too short

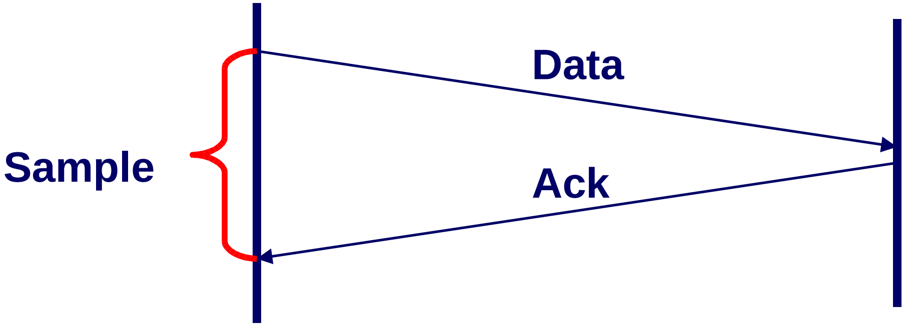
- Time between sending & resending segment

Challenge

- Too long: Add latency to communication when packets dropped
- Too short: Send too many duplicate packets
- General principle: Must be > 1 Round Trip Time (RTT)

Round-trip Time Estimation

Every Data/Ack pair gives new RTT estimate

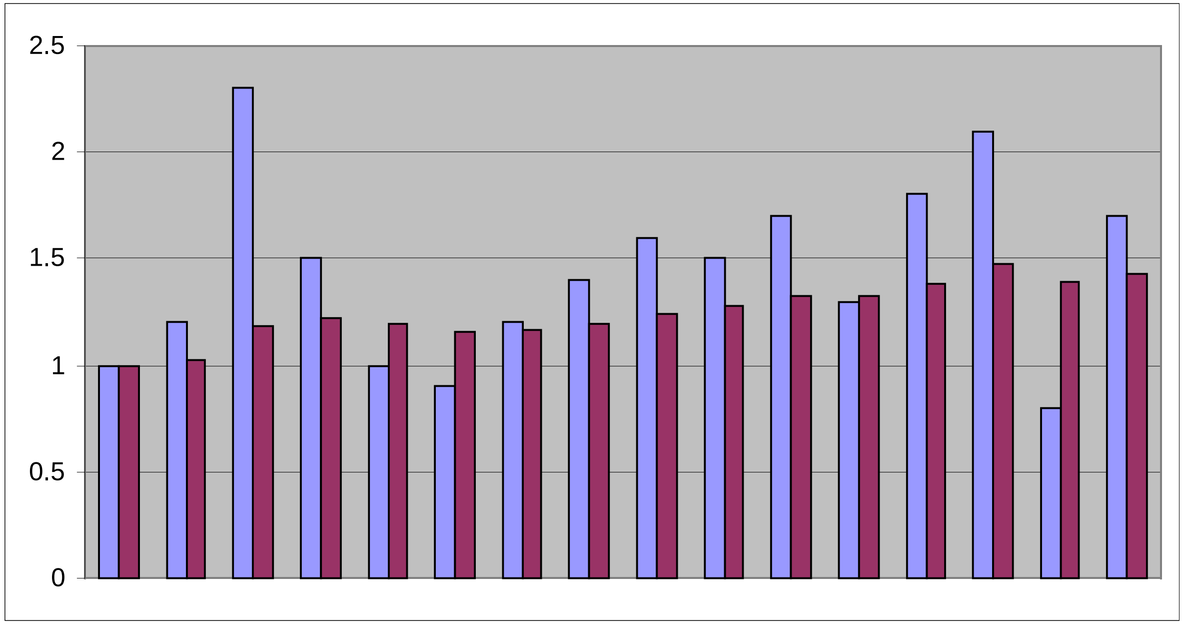


Can Get Lots of Short-Term Fluctuations

Original TCP Round-trip Estimator

Round trip times exponentially averaged:

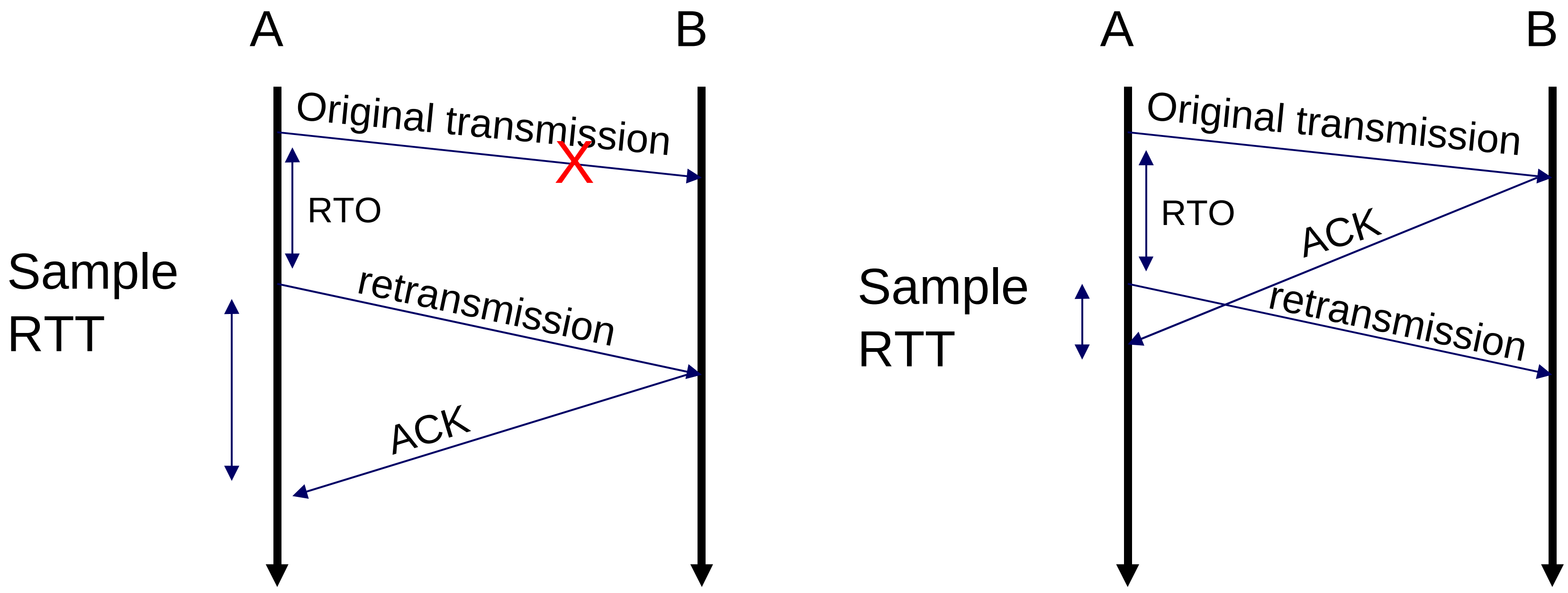
- New RTT = α (old RTT) + (1 - α) (new sample)
- Recommended value for α: 0.8 - 0.9
  - 0.875 for most TCP's



Retransmit timer set to β RTT, where β = 2

- Want to be somewhat conservative about retransmitting

RTT Sample Ambiguity



Karn/Partridge Algorithm

- **Ignore sample for segment that has been retransmitted**
- **Use exponential backoff for retransmissions**
  - **Each time retransmit same segment, double the RTO**
  - **Based upon premise that major congestion is causing packet losses**

Sequence Number Space

Each byte in byte stream is numbered.

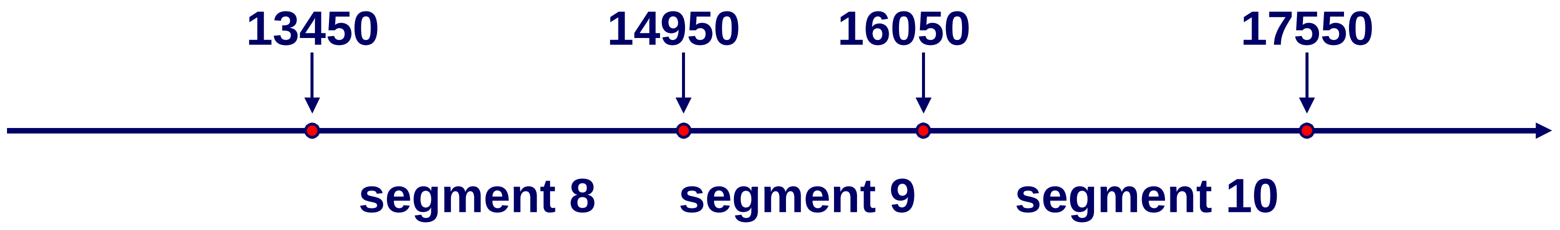
- 32 bit value
- Wraps around
- Initial values selected at start up time

TCP breaks up the byte stream in packets (“segments”)

- Packet size is limited to the Maximum Segment Size

Each segment has a sequence number.

- Indicates where it fits in the byte stream



Finite Length Sequence Number

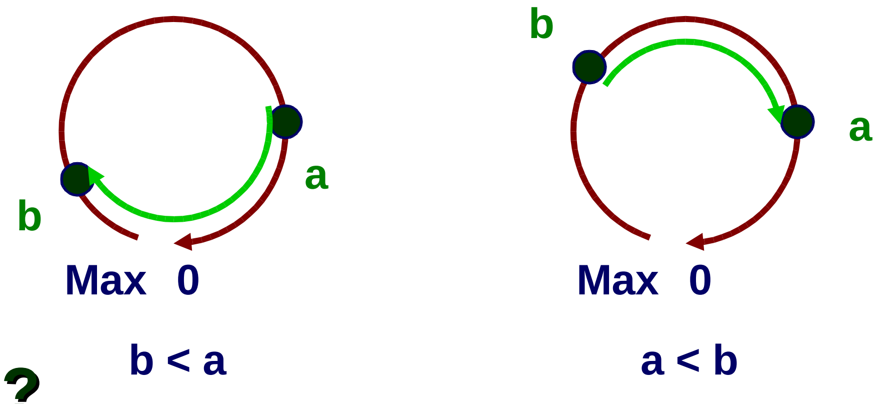
Sequence number can wrap around

- What is the problem?
- What is the solution?
  - Hint: not “crash the kernel”
  - Not even “crash the connection” or “connection full”

Sequence Numbers

32 Bits, Unsigned

Circular Comparison



Why So Big?

- For sliding window, must have
- |Sequence Space| > |Sending Window| + |Receiving Window|
  - No problem
- Also, want to guard against stray packets
  - With IP, packets have maximum lifetime of 120s
  - Sequence number would wrap around in this time at 286MB/s

Error Control Summary

Basic mechanisms

- CRC, checksum
- Timeout
- Acknowledgement
- Sequence numbers
- Window

Many variations and details

TCP Flow Control

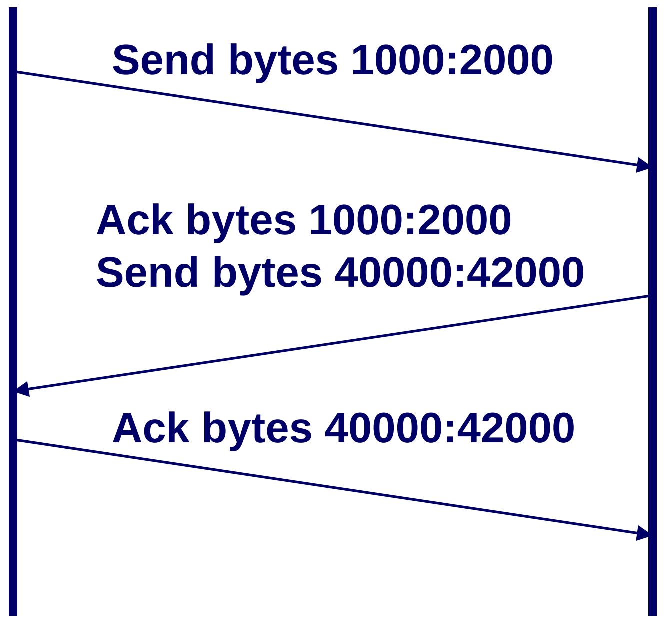
Recall sliding-window as used for error control

- For window size *n*, can send up to *n* bytes without receiving an acknowledgement
- When the data are acknowledged then the window slides forward

Achieve flow control via dynamically-sized window

- Sender naturally tracks outstanding packets versus max
  - Sending one packet decreases budget by one
- Receiver updates “open window” in every response
  - Packet B ⇒ A contains Ack_A and Window_A
  - Sender can send bytes up through $(\text{Ack}_A + \text{Window}_A)$
  - Receiver can increase or decrease window at any time
- Original TCP always sent entire window
  - Congestion control now limits this

Bidirectional Communication



Send bytes 1000:2000

Ack bytes 1000:2000
Send bytes 40000:42000

Ack bytes 40000:42000

Each Side of Connection can Send *and* Receive

What this Means

- Maintain different sequence numbers for each direction
- Single segment can contain new data for one direction, plus acknowledgement for other
  - But some contain only data & others only acknowledgement

Ongoing Communication

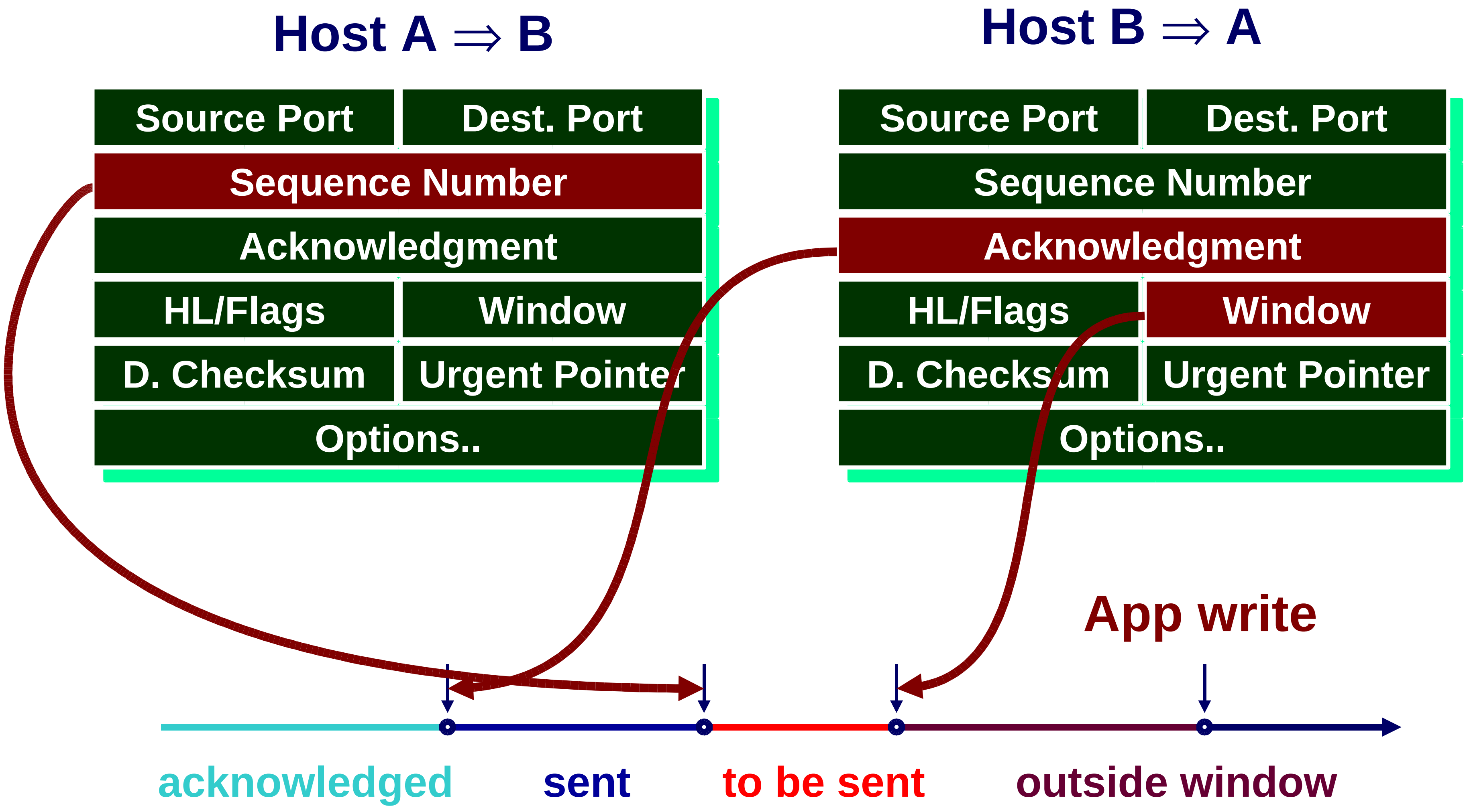
Bidirectional Communication

- Each side acts as sender & receiver
- Every message contains acknowledgement of received sequence
  - Even if no new data have been received
- Every message advertises window size
  - Size of its receiving window
- Every message contains sent sequence number
  - Even if no new data being sent

When Does Sender Actually Send Message?

- When a maximal-sized segment worth of bytes is available
- When application tells it
  - Set PUSH flag for last segment sent
- When timer expires

Window Flow Control: Send Side

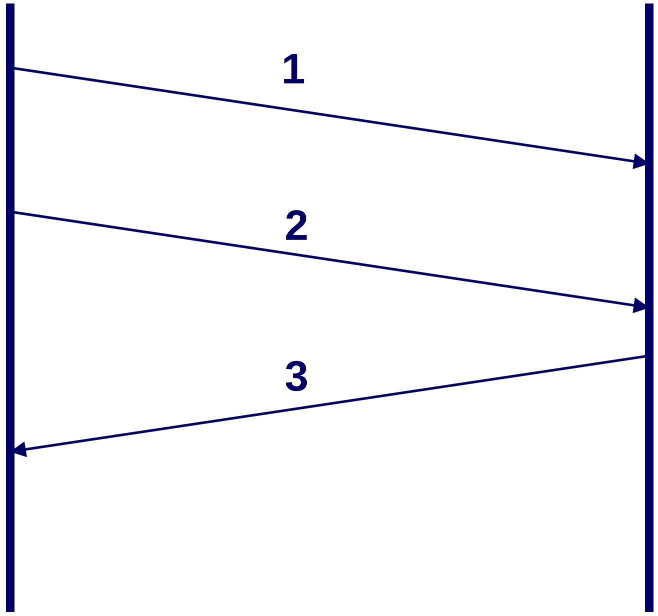


TCP Transmission

```
09:23:33.132509 IP 128.2.222.198.3123 > 192.216.219.96.80: P
  4019802005:4019802801(796) ack 3428951570 win 65535 (DF)

09:23:33.149875 IP 128.2.222.198.3123 > 192.216.219.96.80: .
  4019802801:4019804061(1260) ack 3428951570 win 65535 (DF)

09:23:33.212291 IP 192.216.219.96.80 > 128.2.222.198.3123: . ack
  4019802801 win 7164 (DF)
```



Client sends 796 bytes

Client sends 1260 more bytes

Server acknowledges 1996 bytes

Tearing Down Connection

Either Side Can Initiate Tear Down

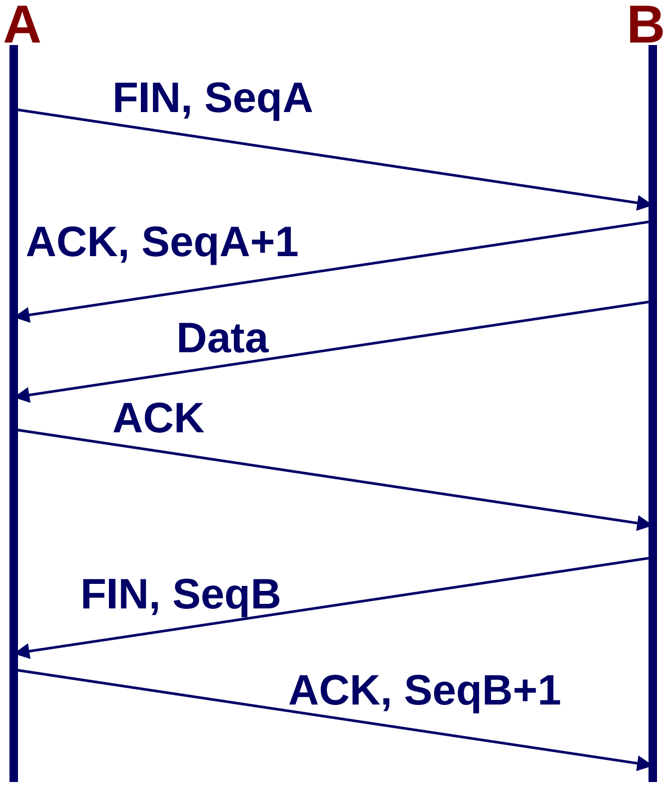
- Send FIN signal
- “I'm not going to send any more data”

Other Side Can Continue Sending Data

- Half-open connection
- Must continue to acknowledge

Acknowledging FIN

- Acknowledge last sequence number + 1



TCP Connection Teardown Example

```
09:54:17.585396 IP 128.2.222.198.4474 > 128.2.210.194.6616: F
  1489294581:1489294581(0) ack 1909787689 win 65434 (DF)

09:54:17.585732 IP 128.2.210.194.6616 > 128.2.222.198.4474: F
  1909787689:1909787689(0) ack 1489294582 win 5840 (DF)

09:54:17.585764 IP 128.2.222.198.4474 > 128.2.210.194.6616: . ack
  1909787690 win 65434 (DF)
```

Session

- Echo client on 128.2.222.198, server on 128.2.210.194

Client FIN

- SeqC: 1489294581

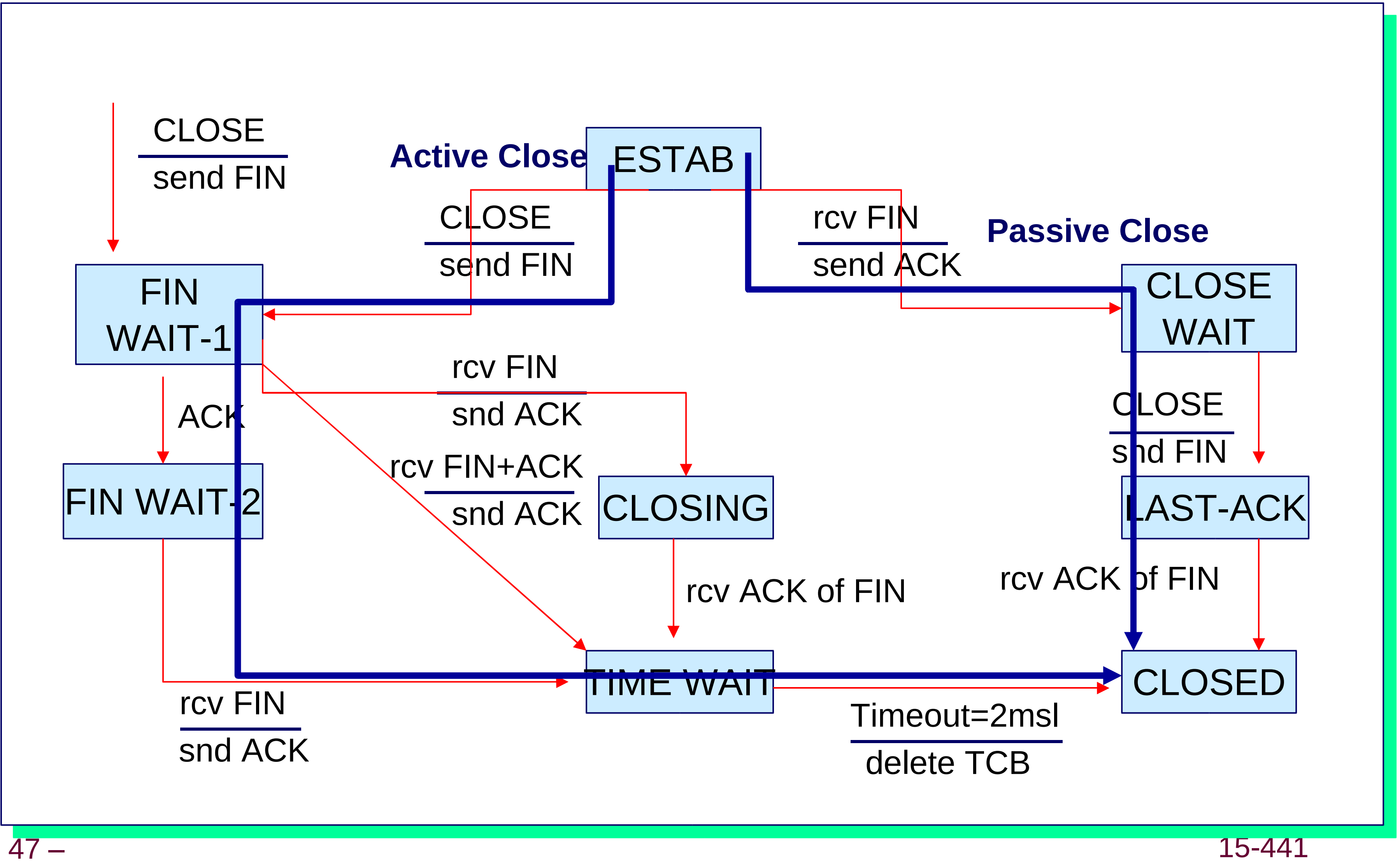
Server ACK + FIN

- Ack: 1489294582 (= SeqC+1)
- SeqS: 1909787689

Client ACK

- Ack: 1909787690 (= SeqS+1)

State Diagram: Connection Tear-down



Key TCP Design Decisions

Connection Oriented

- Explicit setup & teardown of connections

Byte-stream oriented

- vs. message-oriented
- Sometimes awkward for application to infer message boundaries

Sliding Window with Cumulative Acknowledgement

- Single acknowledgement covers range of bytes
- Single missing message may trigger series of retransmissions

No Negative Acknowledgements

- Any problem with transmission must be detected by timeout
- OK for IP to silently drop packets