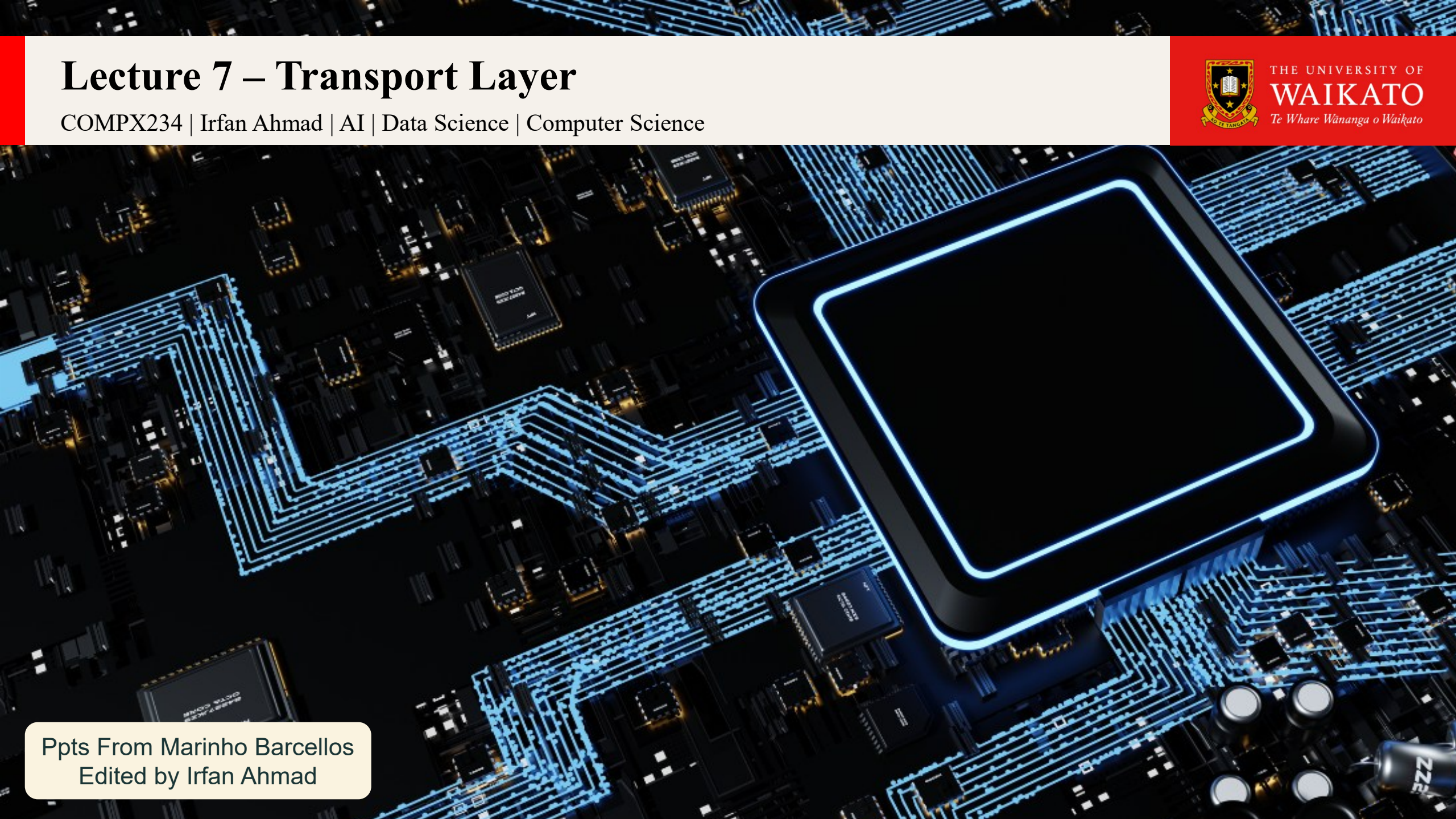
Lecture 7 – Transport Layer

COMPX234 | Irfan Ahmad | AI | Data Science | Computer Science

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato



Ppts From Marinho Barcellos
Edited by Irfan Ahmad

Transport layer: overview

Our goal:

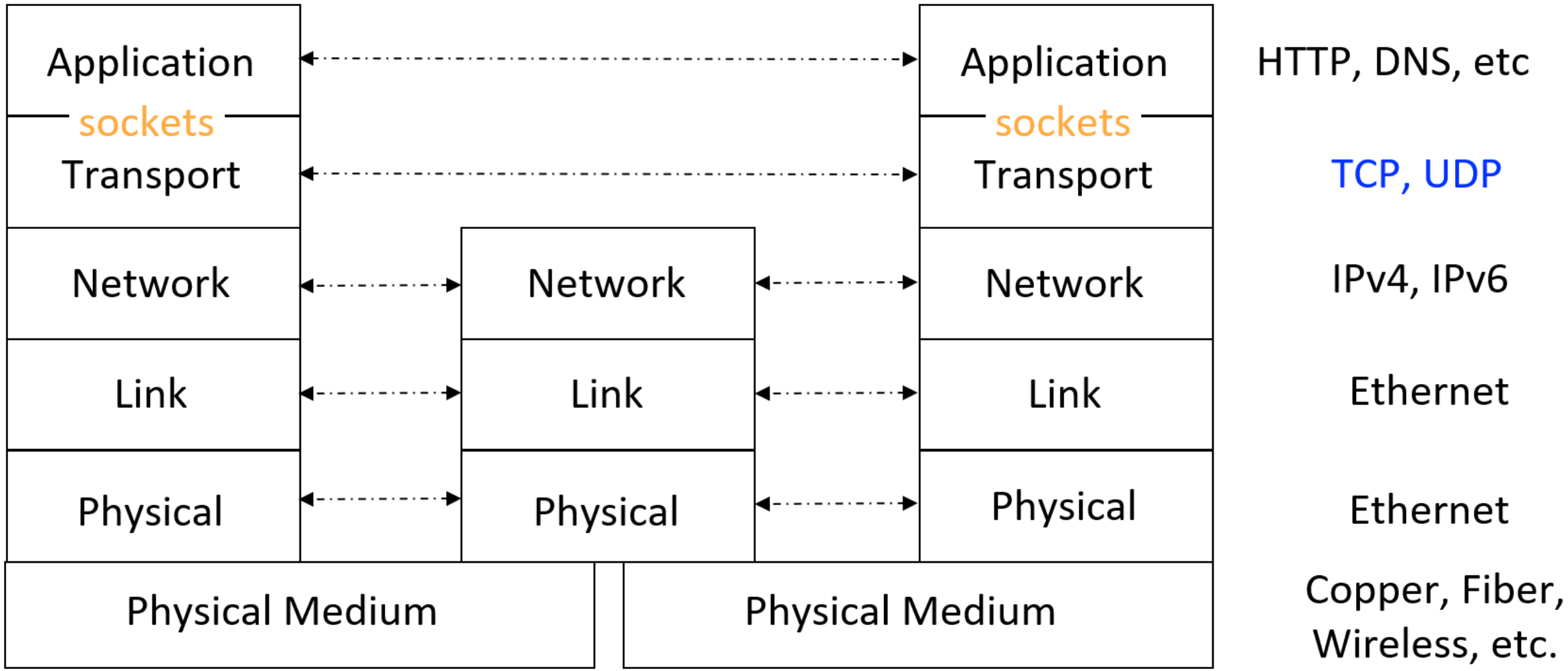
- understand principles behind transport layer services:
  - multiplexing, demultiplexing
  - reliable data transfer
  - flow control
  - congestion control
- learn about Internet transport layer protocols:
  - UDP: connectionless transport
  - TCP: connection-oriented reliable transport
  - TCP congestion control

Outlines

- Transport-layer services
- Multiplexing and demultiplexing
- Connectionless transport: UDP
- Principles of reliable data transfer
- Stop-and-wait Protocols
- Sliding Windows

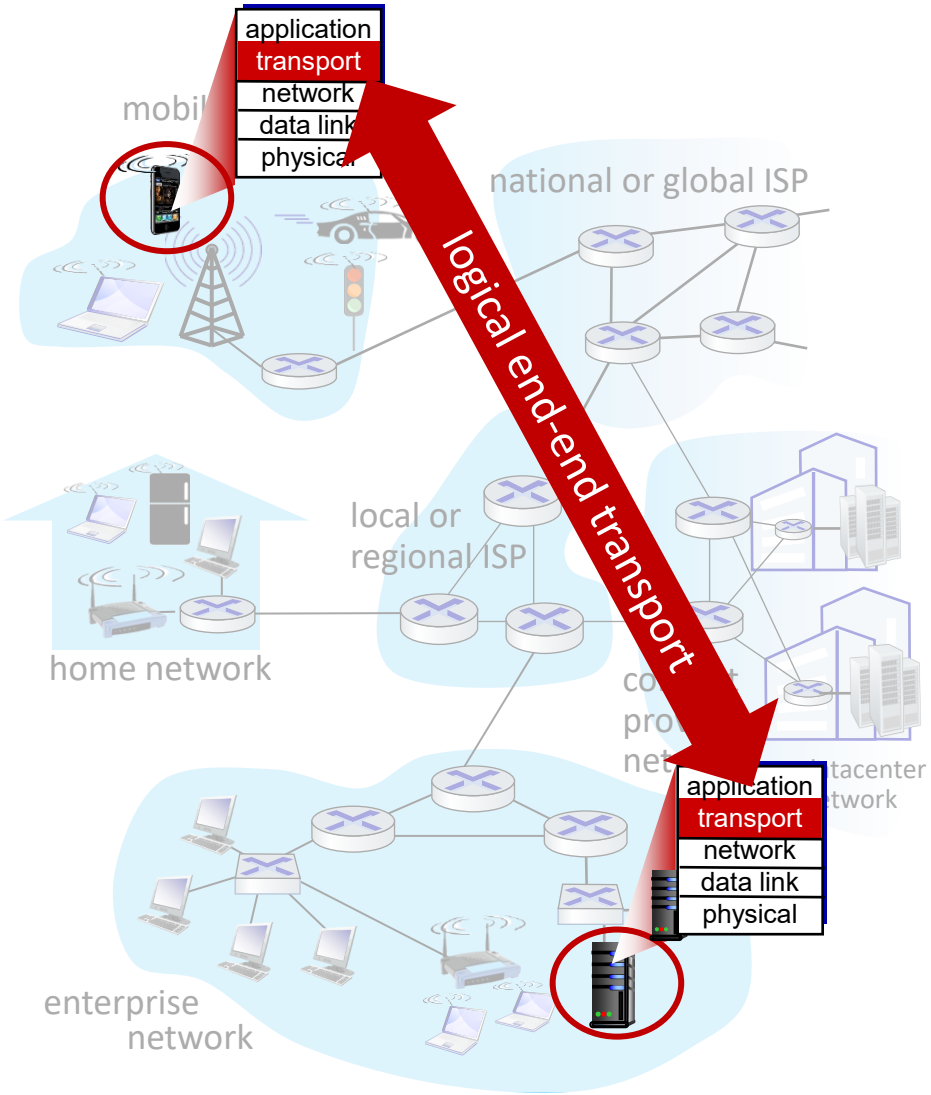
Transport-layer Services

Recall the internet protocol layers



Transport services and protocols

- provide *logical communication* between application processes running on different hosts
- transport protocols actions in end systems:
  - sender: breaks application messages into *segments*, passes to network layer
  - receiver: reassembles segments into messages, passes to application layer
- two transport protocols available to Internet applications
  - TCP, UDP



Transport vs. network layer services and protocols

household analogy:

12 kids in Ann's house sending letters to 12 kids in Bill's house:

- hosts = houses
- processes = kids
- app messages = letters in envelopes

Transport vs. network layer services and protocols

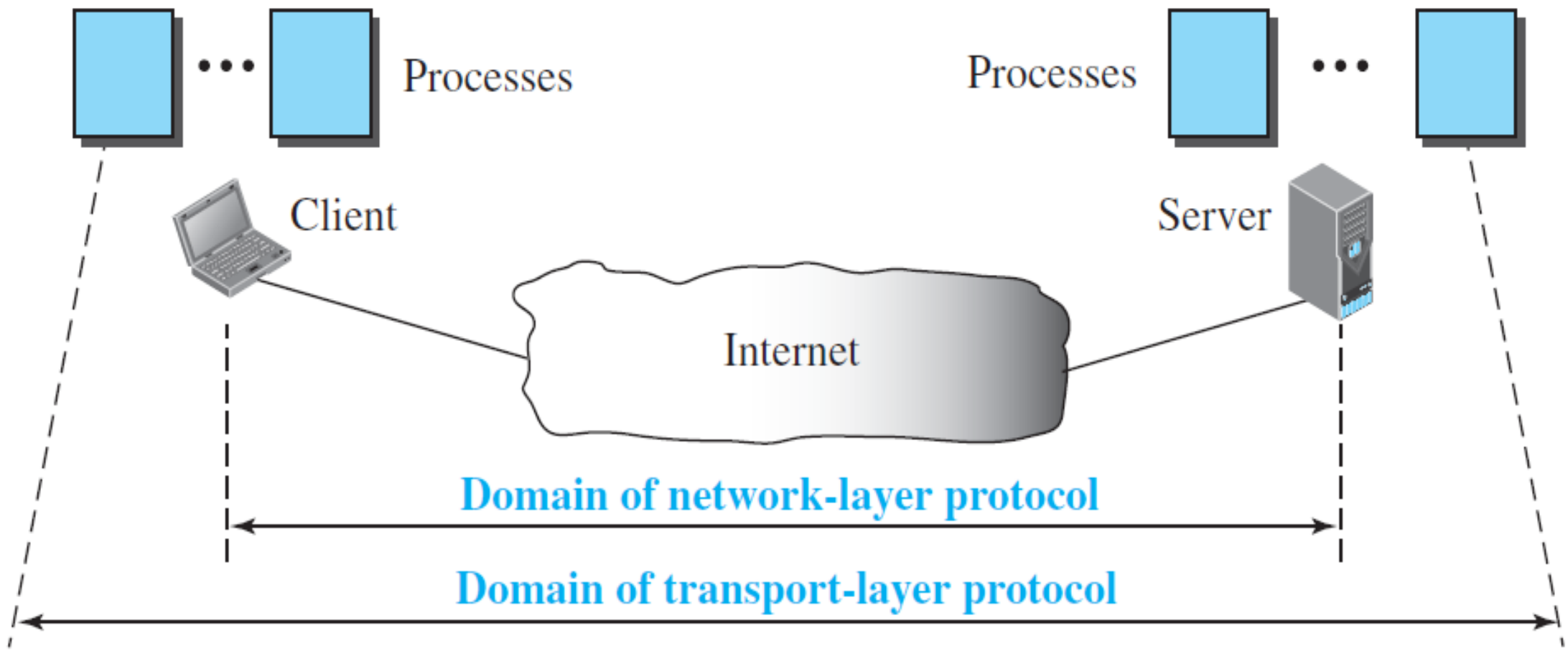
- transport layer: communication between *processes*
  - relies on, enhances, network layer services
- network layer: communication between *hosts*

household analogy:

12 kids in Ann's house sending letters to 12 kids in Bill's house:

- hosts = houses
- processes = kids
- app messages = letters in envelopes

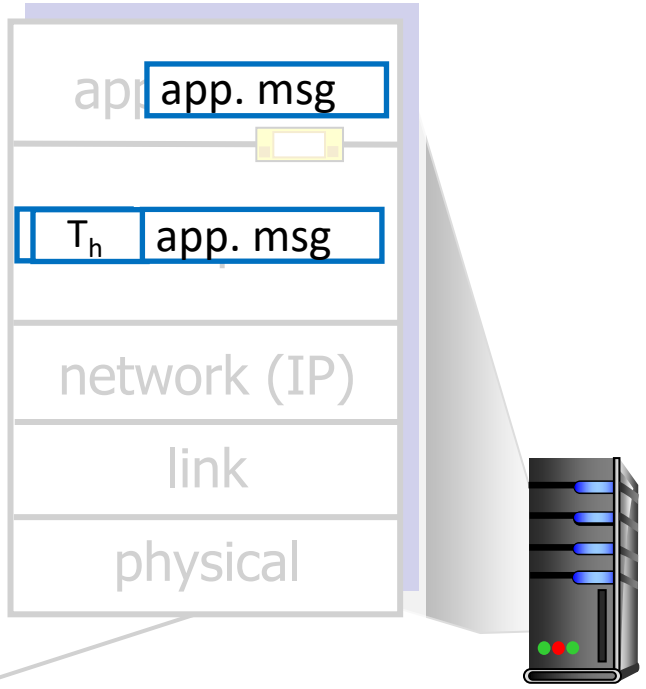
Transport vs. network layer services and protocols

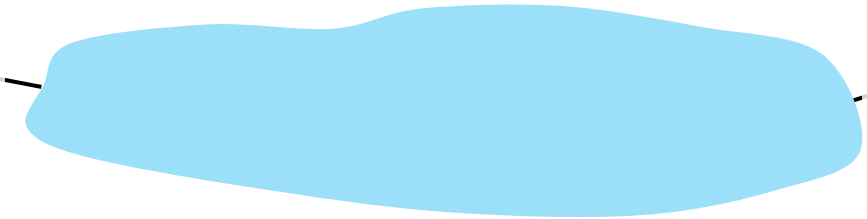


Transport Layer Actions

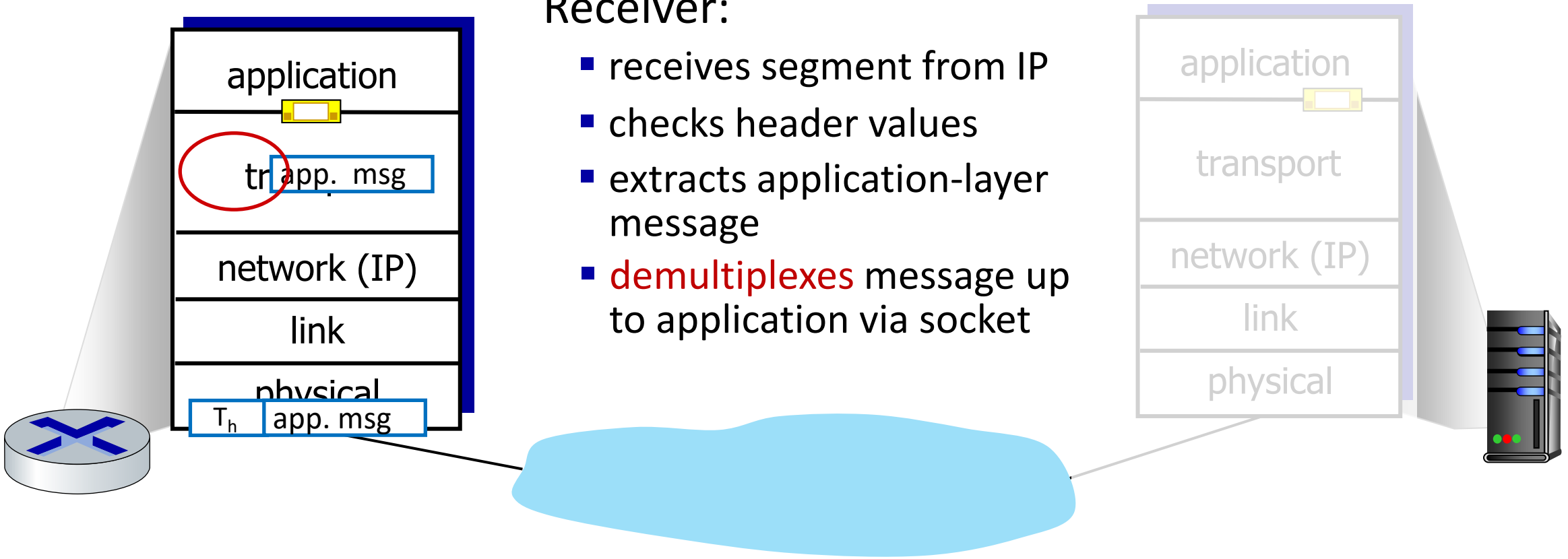
Sender:

- is given an application-layer message
- determines segment header fields values
- creates segment
- passes segment to IP





Transport Layer Actions

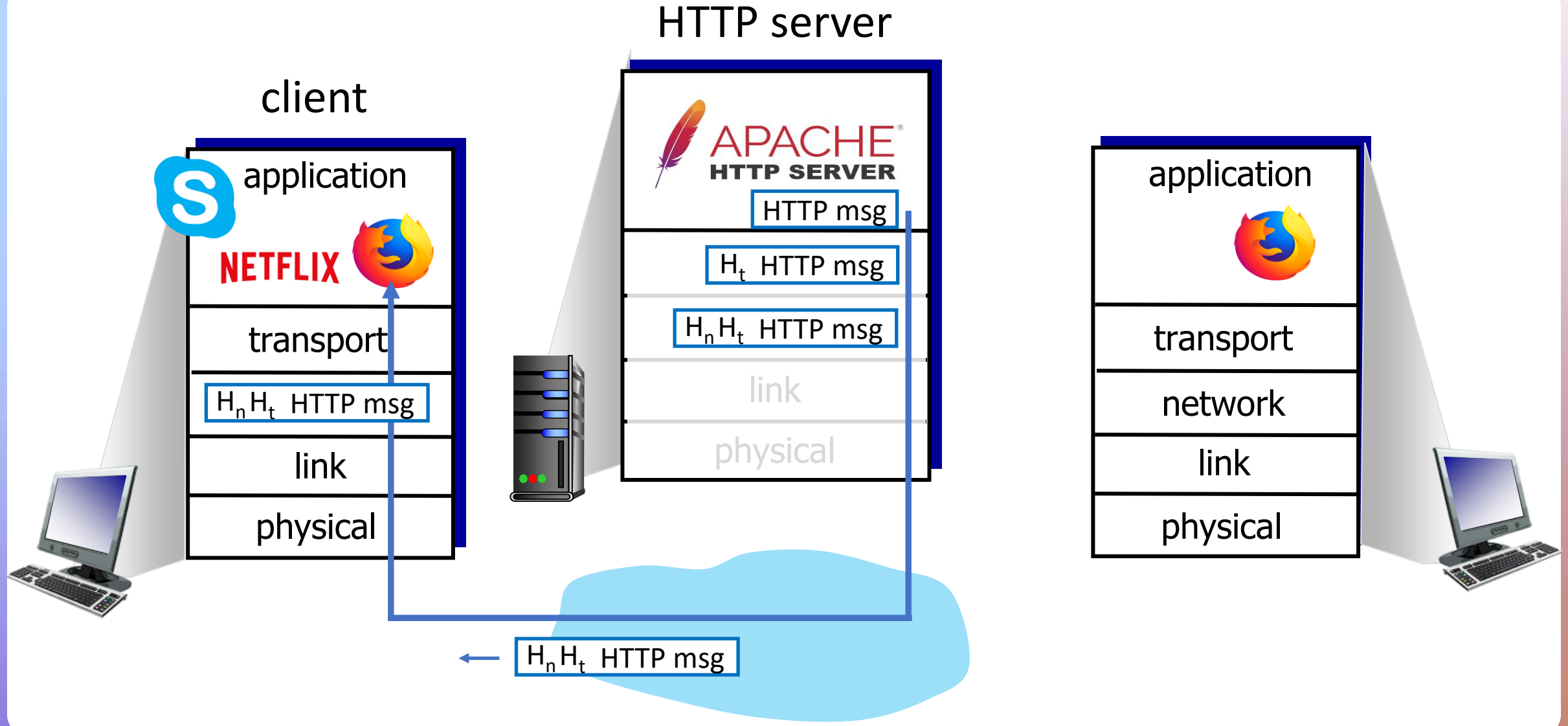


Receiver:

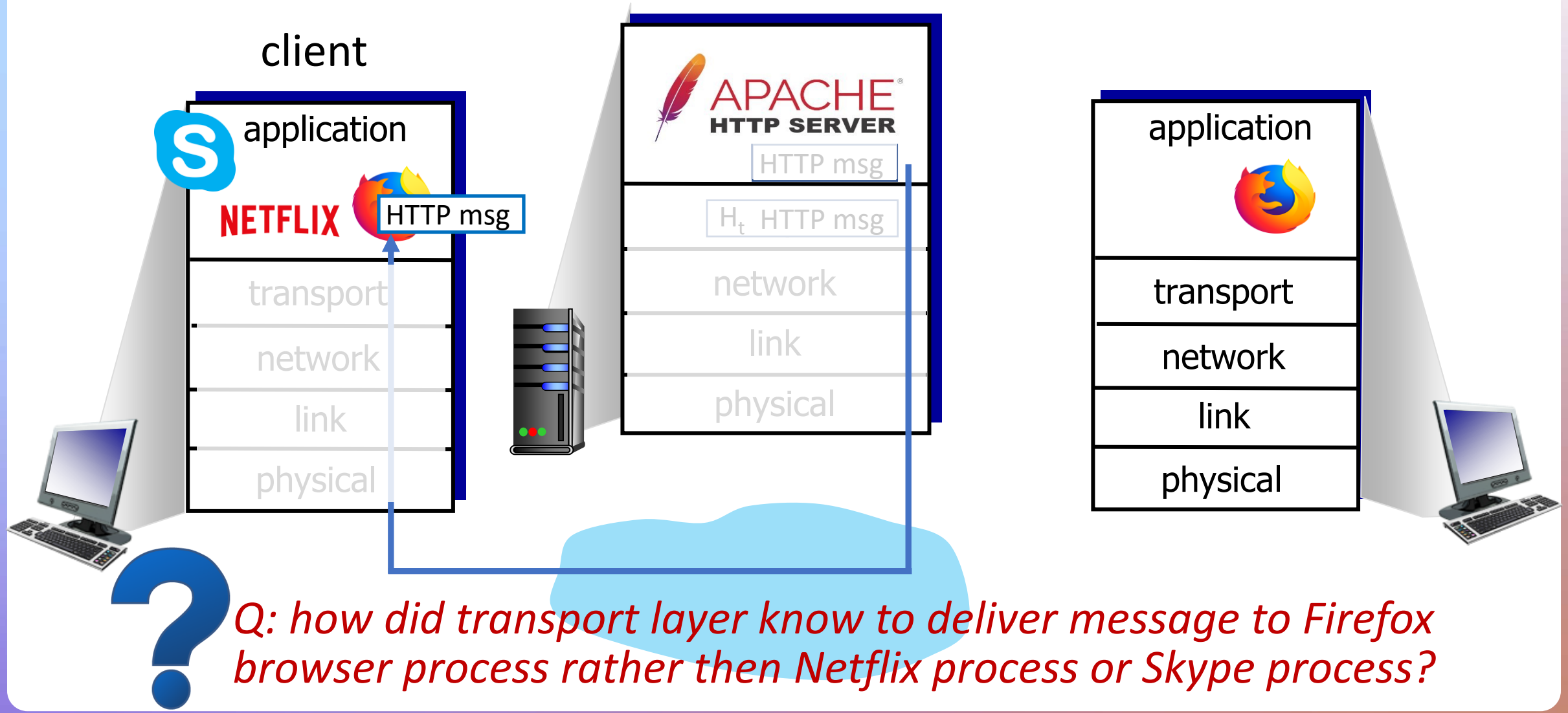
- receives segment from IP
- checks header values
- extracts application-layer message
- demultiplexes message up to application via socket

Multiplexing and Demultiplexing

Suppose Firefox requests a page from Apache server



Suppose Firefox requests a page from Apache server



Q: how did transport layer know to deliver message to Firefox browser process rather then Netflix process or Skype process?

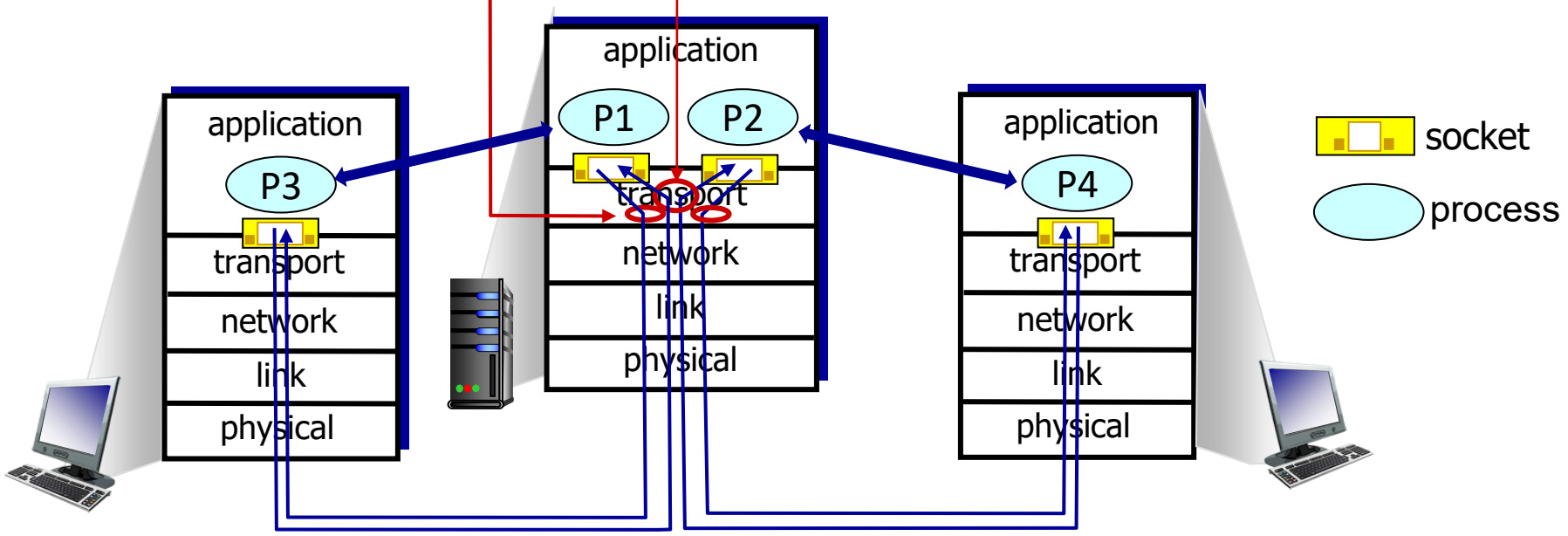
Multiplexing/demultiplexing

multiplexing as sender:

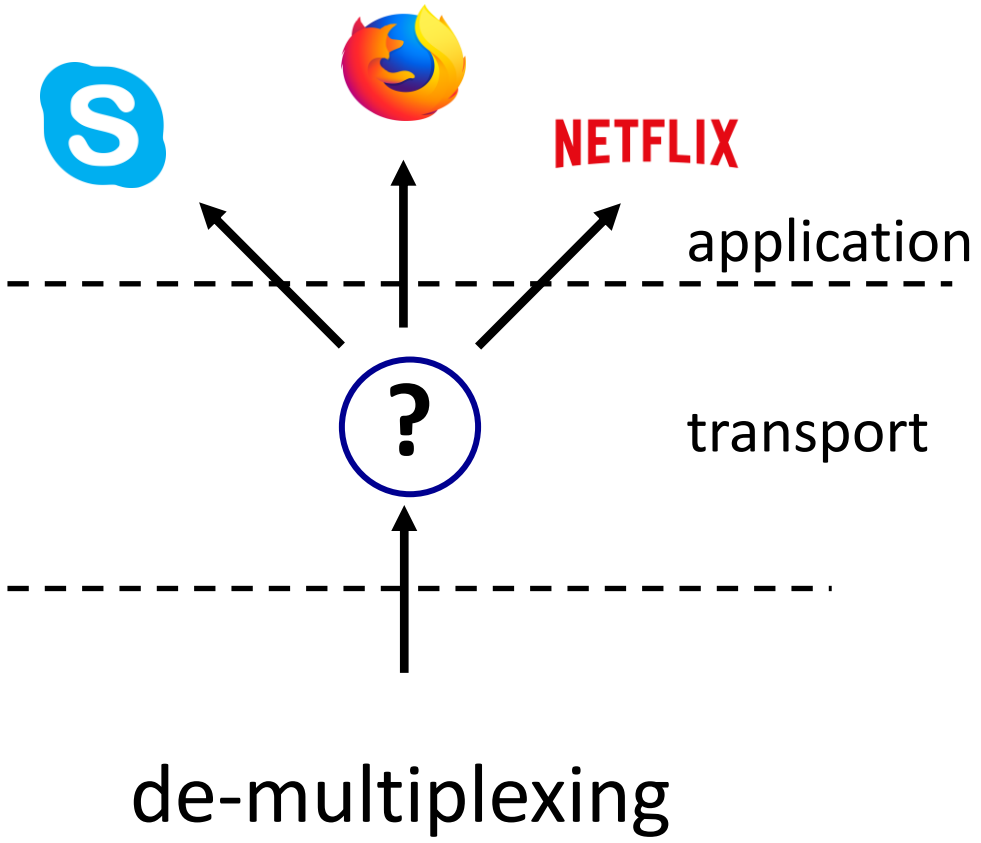
handle data from multiple sockets, add transport header (later used for demultiplexing)

demultiplexing as receiver:

use header info to deliver received segments to correct socket

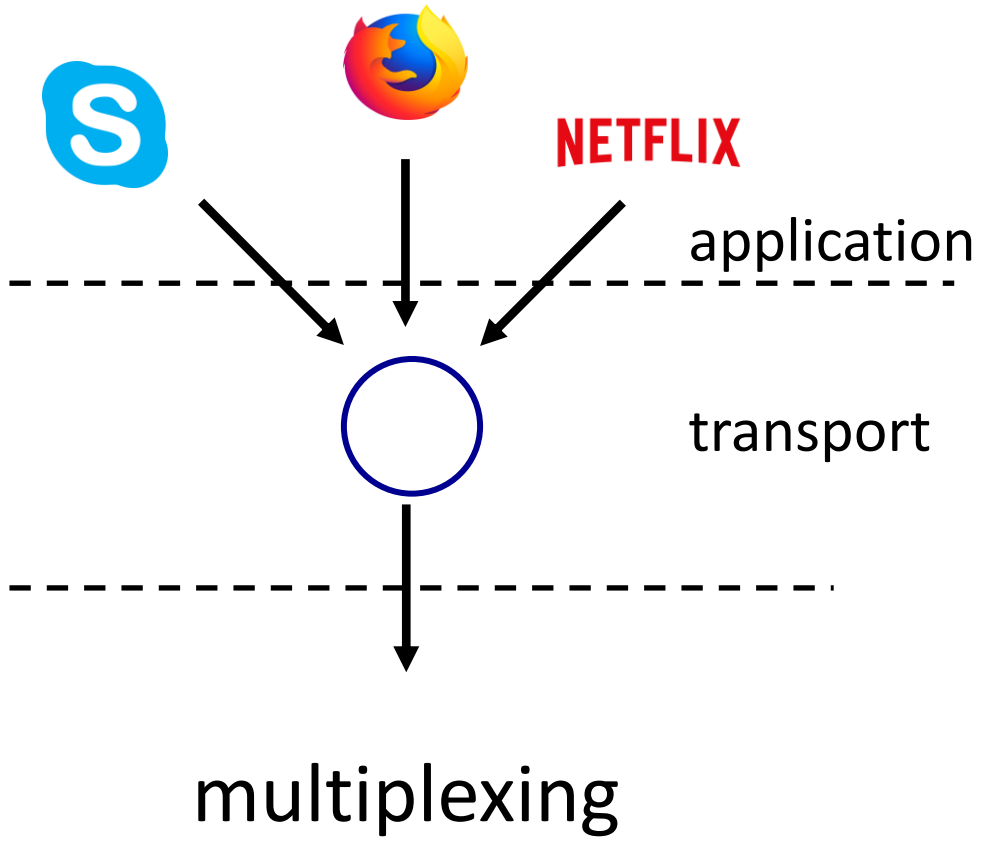


Multiplexing/demultiplexing



Demultiplexing

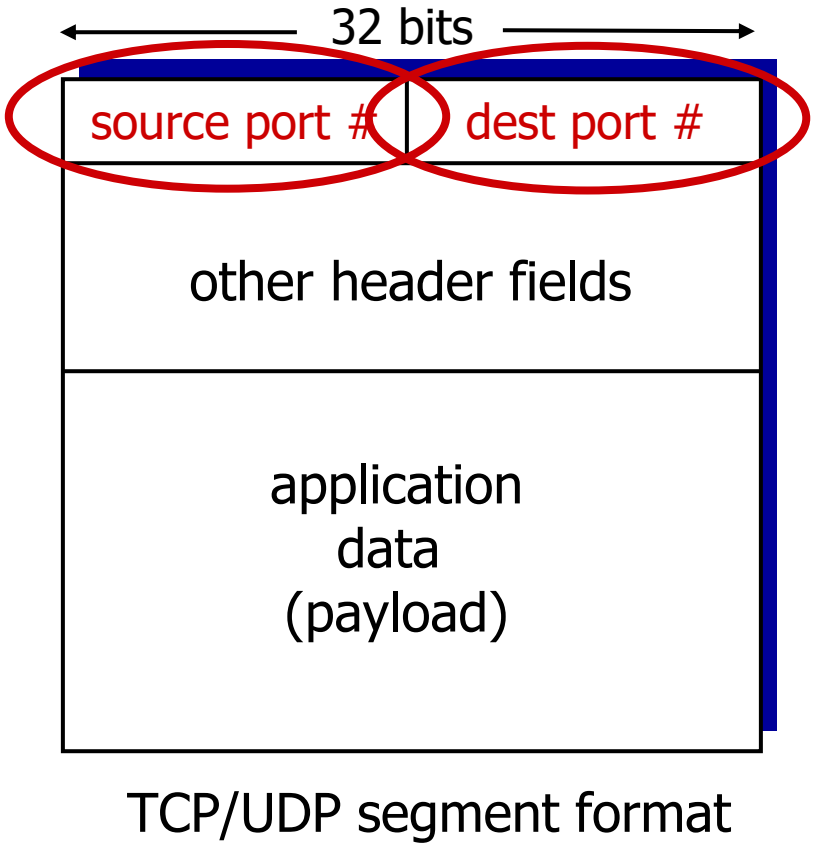
Multiplexing/demultiplexing




Multiplexing

How demultiplexing works

- host receives IP datagrams
  - each datagram has source IP address, destination IP address
  - each datagram carries one transport-layer segment
  - each segment has source, destination port number
- host uses *IP addresses & port numbers* to direct segment to appropriate socket



TCP/UDP segment format

Connectionless demultiplexing

Recall:

- when creating socket, must specify *host-local* port #:

```
DatagramSocket mySocket1
 = new DatagramSocket(12534);
```

- when creating datagram to send into UDP socket, must specify
  - destination IP address
  - destination port #

when receiving host receives *UDP* segment:

- checks destination port # in segment
- directs UDP segment to socket with that port #

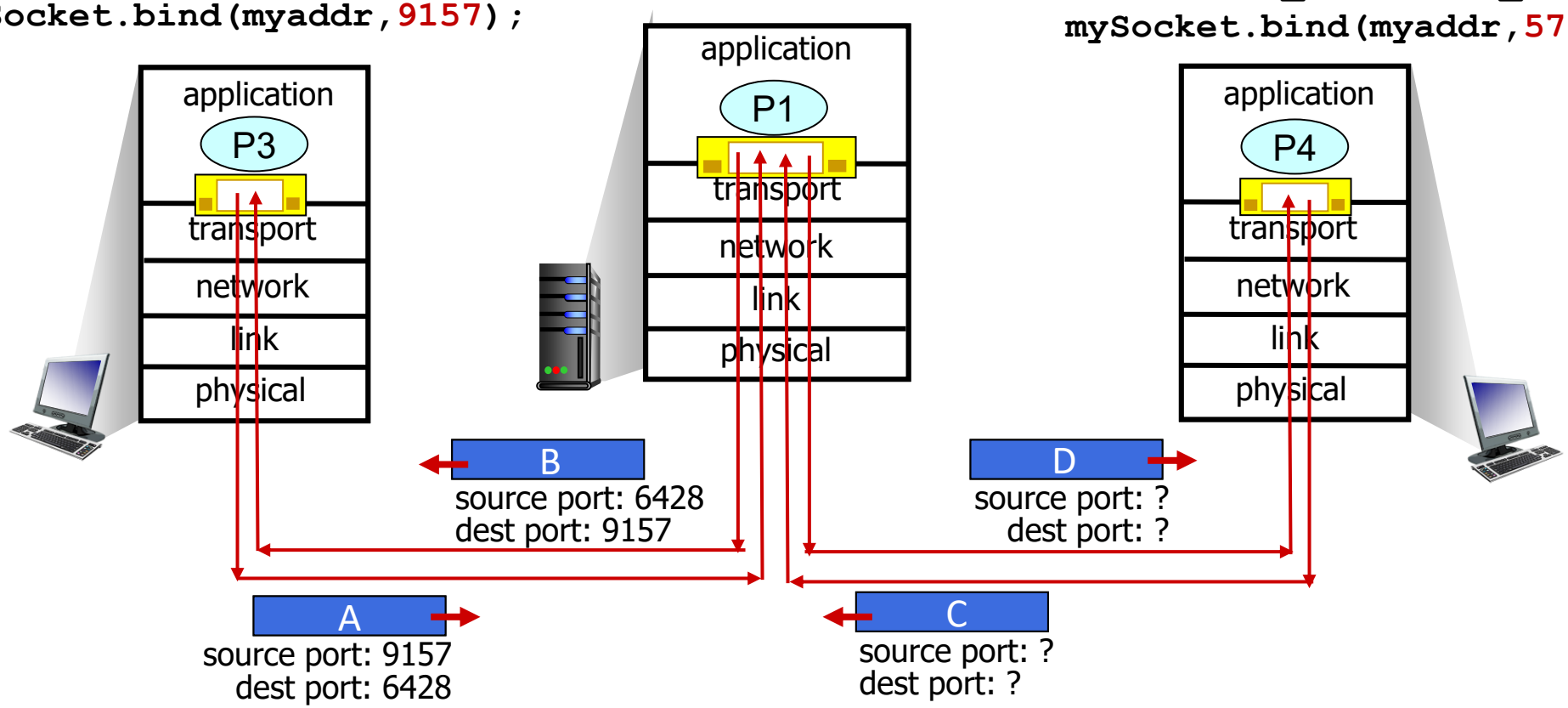
IP/UDP datagrams with *same dest. port #,* but different source IP addresses and/or source port numbers will be directed to *same socket* at receiving host

Connectionless demultiplexing: an example

```
mySocket =
 socket(AF_INET,SOCK_DGRAM)
mySocket.bind(myaddr,6428);
```

```
mySocket =
 socket(AF_INET,SOCK_STREAM)
mySocket.bind(myaddr,9157);
```

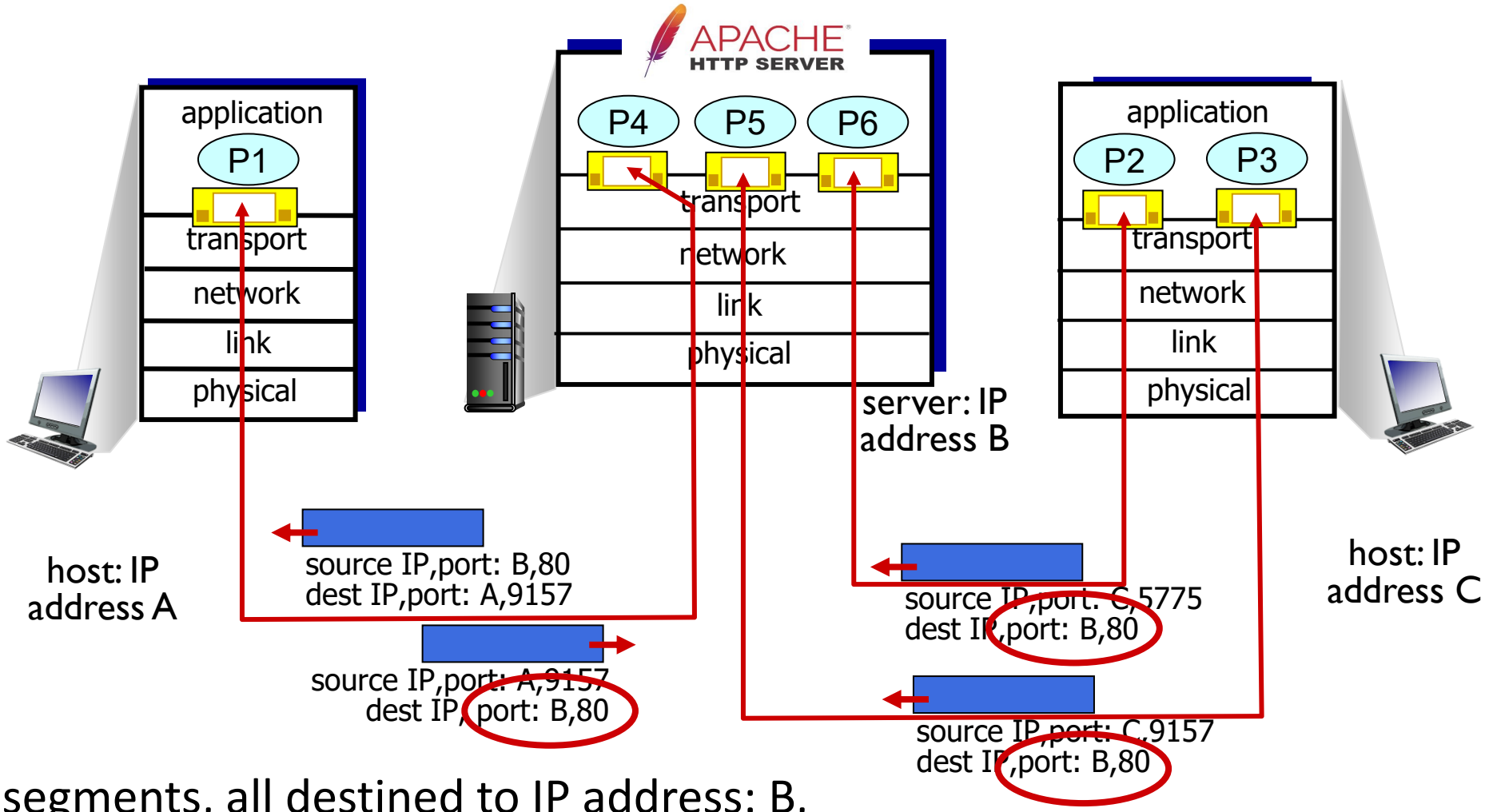
```
mySocket =
 socket(AF_INET,SOCK_STREAM)
mySocket.bind(myaddr,5775);
```



Connection-oriented demultiplexing

- TCP socket identified by 4-tuple:
  - source IP address
  - source port number
  - dest IP address
  - dest port number
- demux: receiver uses *all four values (4-tuple)* to direct segment to appropriate socket
- server may support many simultaneous TCP sockets:
  - each socket identified by its own 4-tuple
  - each socket associated with a different connecting client

Connection-oriented demultiplexing: example



Three segments, all destined to IP address: B, dest port: 80 are demultiplexed to *different* sockets

Summary

- Multiplexing, demultiplexing: based on segment, datagram header field values
- **UDP:** demultiplexing using destination port number (only)
- **TCP:** demultiplexing using 4-tuple: source and destination IP addresses, and port numbers
- Multiplexing/demultiplexing happen at *all* layers

Connectionless Transport: UDP

UDP: User Datagram Protocol

- "no frills," "bare bones" Internet transport protocol
- "best effort" service, UDP segments may be:
  - lost
  - delivered out-of-order to app
- *connectionless:*
  - no handshaking between UDP sender, receiver
  - each UDP segment handled independently of others

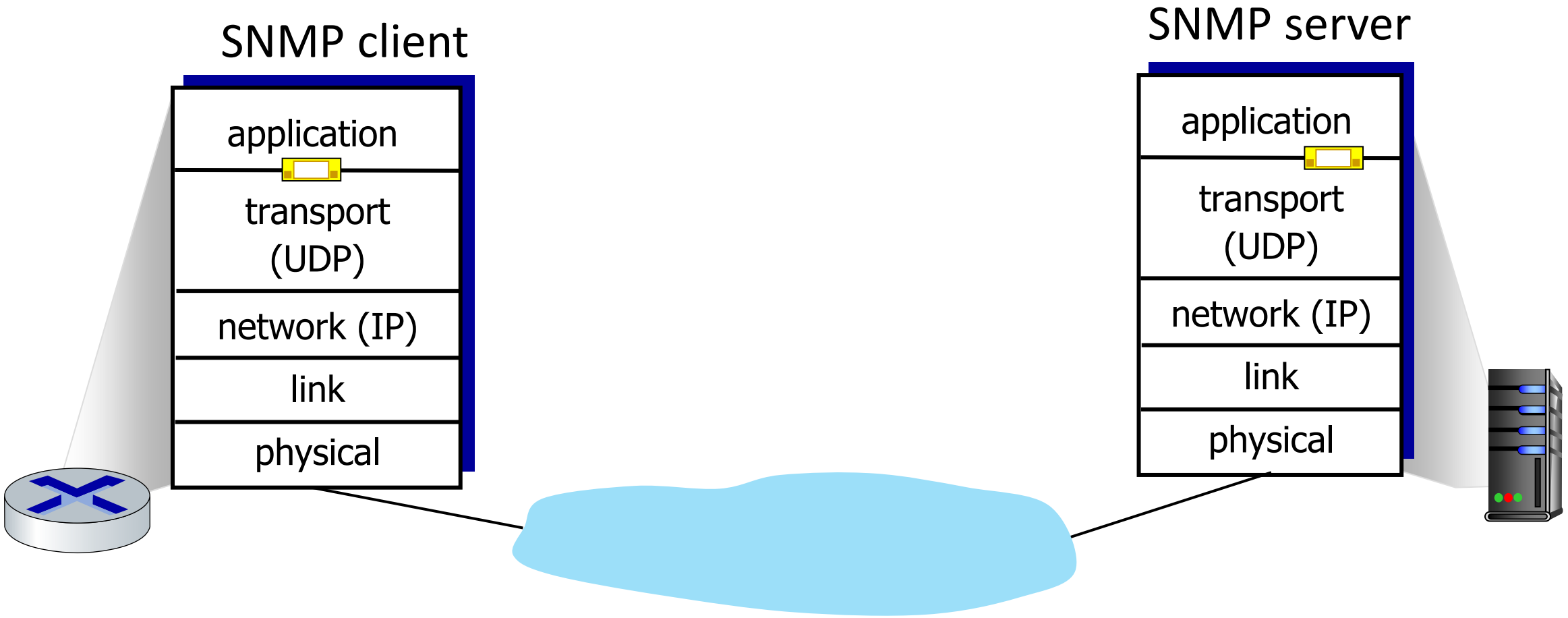
Why is there a UDP?

- no connection establishment (which can add RTT delay)
- simple: no connection state at sender, receiver
- small header size
- no congestion control
  - UDP can blast away as fast as desired!
  - can function in the face of congestion

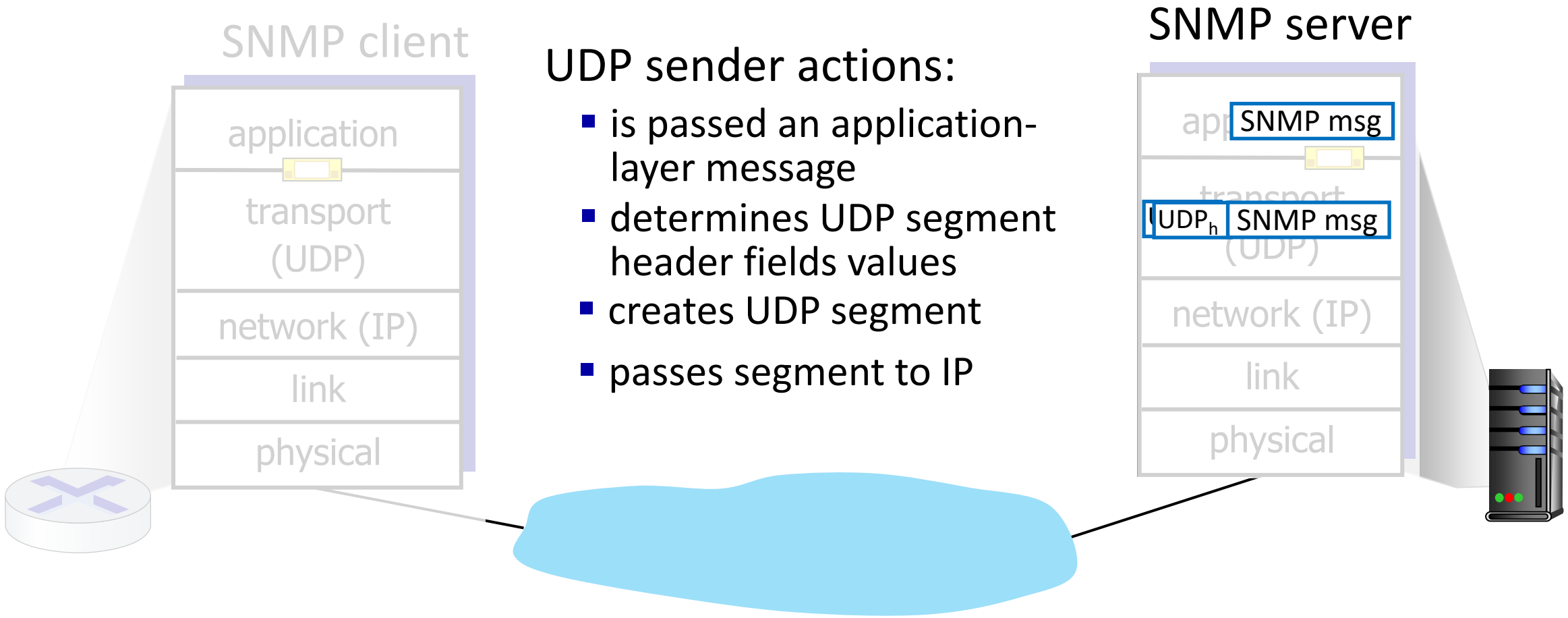
UDP: User Datagram Protocol

- UDP use:
  - streaming multimedia apps (loss tolerant, rate sensitive)
  - DNS
  - SNMP
  - HTTP/3
- if reliable transfer needed over UDP (e.g., HTTP/3):
  - add needed reliability at application layer
  - add congestion control at application layer

UDP: Transport Layer Actions



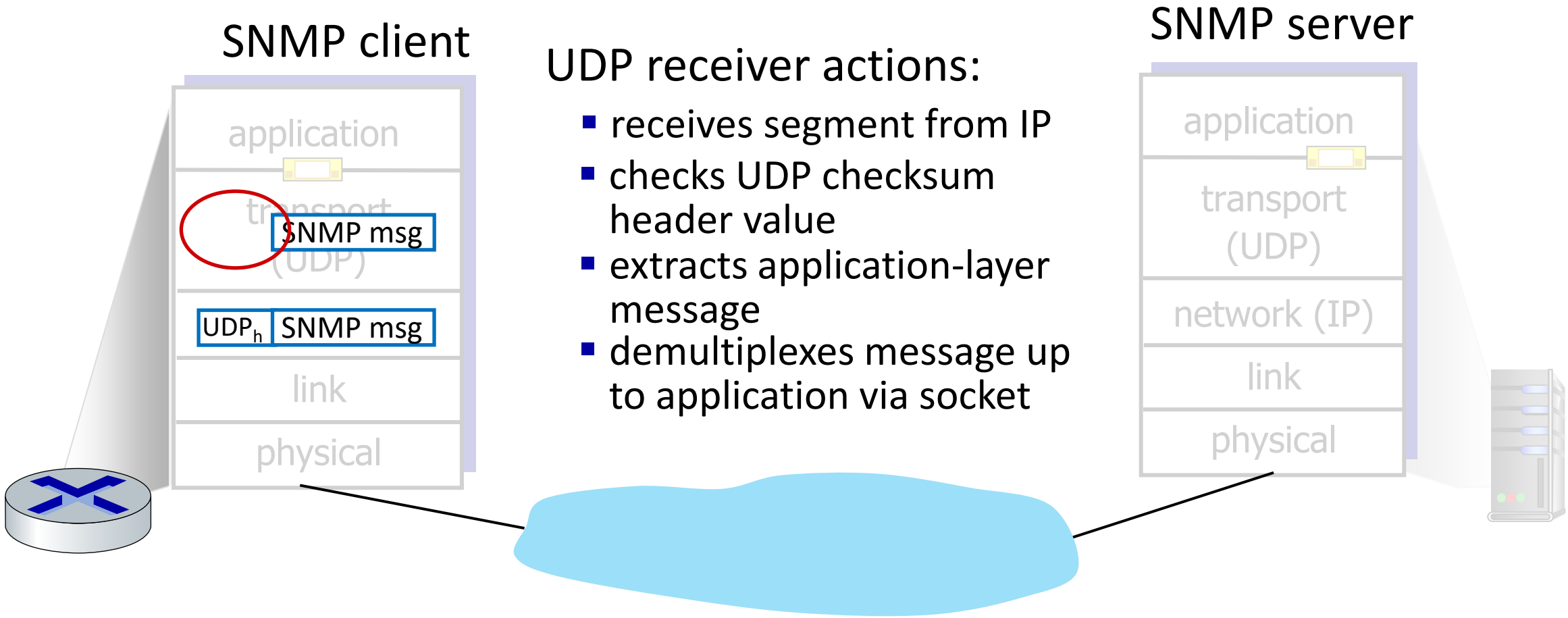
UDP: Transport Layer Actions



UDP sender actions:

- is passed an application-layer message
- determines UDP segment header fields values
- creates UDP segment
- passes segment to IP

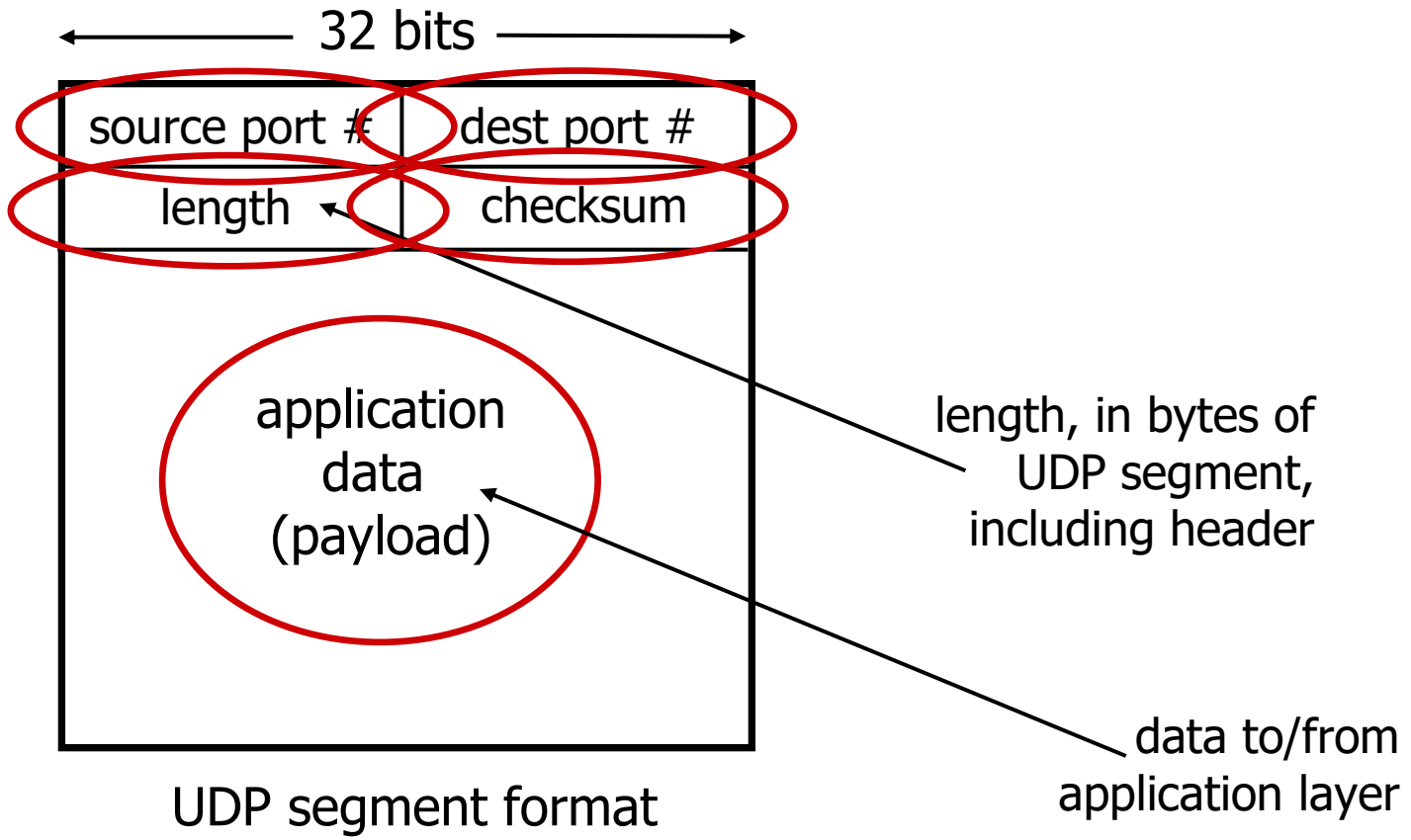
UDP: Transport Layer Actions



UDP receiver actions:

- receives segment from IP
- checks UDP checksum header value
- extracts application-layer message
- demultiplexes message up to application via socket

UDP segment header

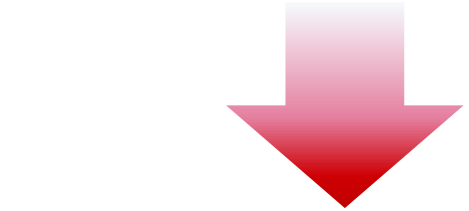


UDP segment format

UDP checksum

Goal: detect errors (*i.e.,* flipped bits) in transmitted segment

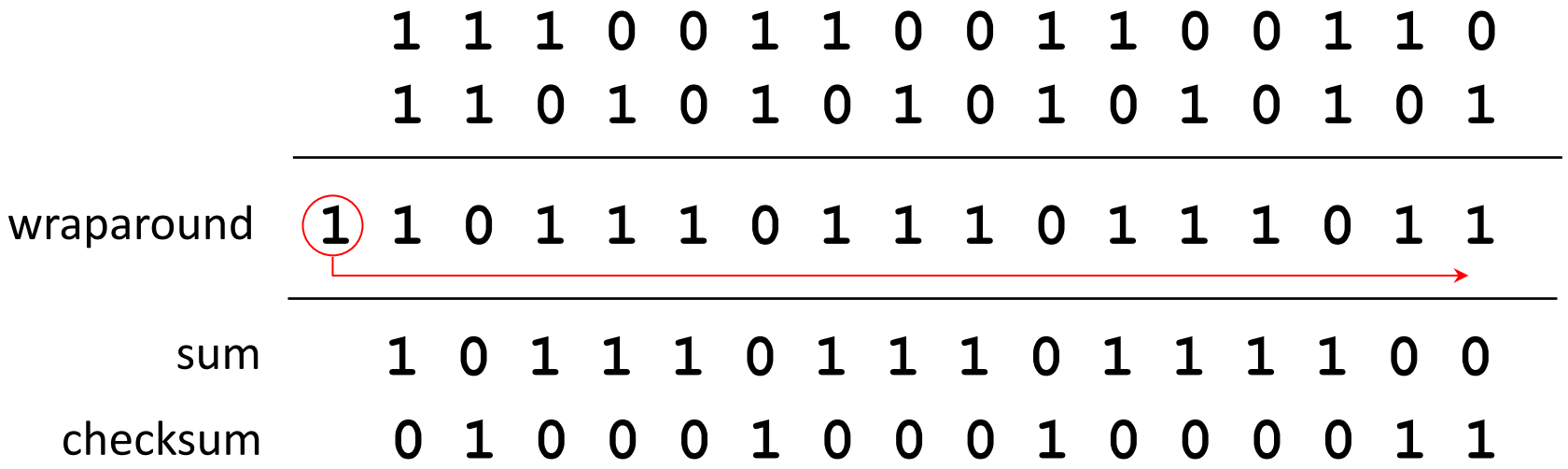
	1st number	2nd number	sum
Transmitted:	5	6	11
Received:	4	6	11



receiver-computed checksum ≠ sender-computed checksum (as received)

Internet checksum: an example

example: add two 16-bit integers

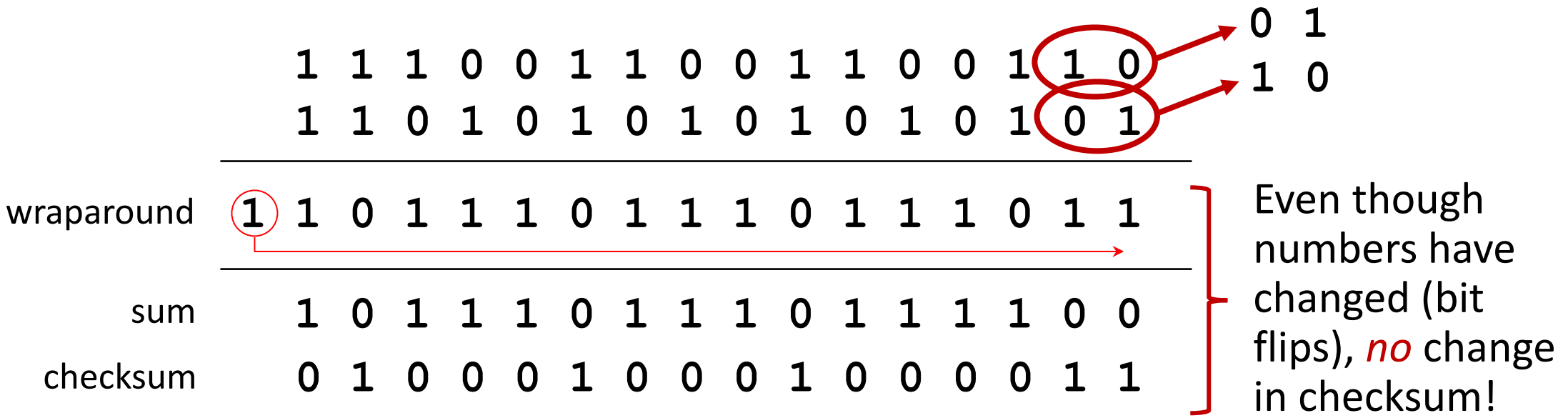


Note: when adding numbers, a carryout from the most significant bit needs to be added to the result

* Check out the online interactive exercises for more examples: http://gaia.cs.umass.edu/kurose_ross/interactive/

Internet checksum: weak protection!

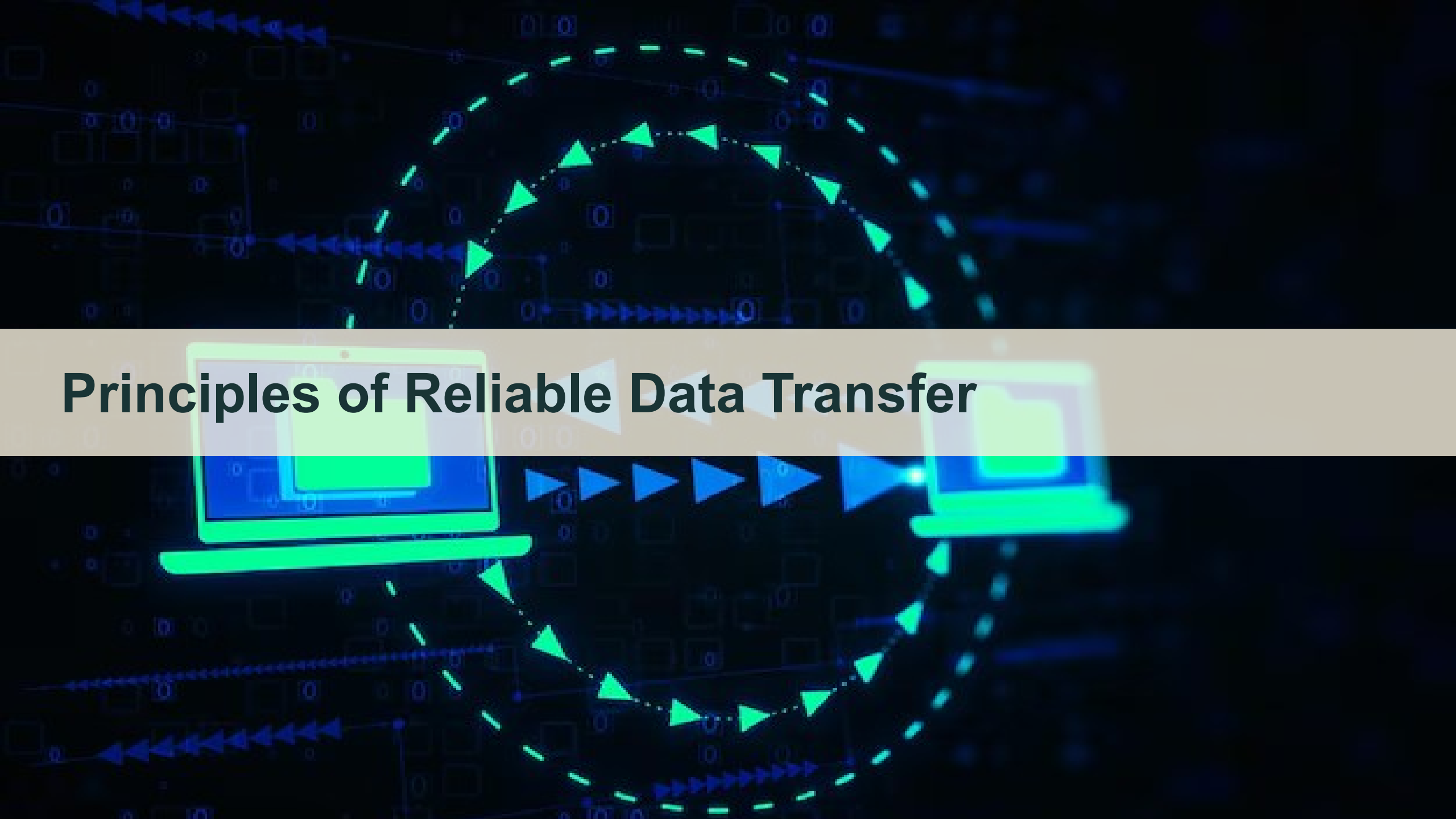
example: add two 16-bit integers



Even though numbers have changed (bit flips), *no* change in checksum!

Summary: UDP

- "no frills" protocol:
  - segments may be lost, delivered out of order
  - best effort service: "send and hope for the best"
- UDP has its plusses:
  - no setup/handshaking needed (no RTT incurred)
  - can function when network service is compromised
  - helps with reliability (checksum)
- build additional functionality on top of UDP in application layer (e.g., HTTP/3)

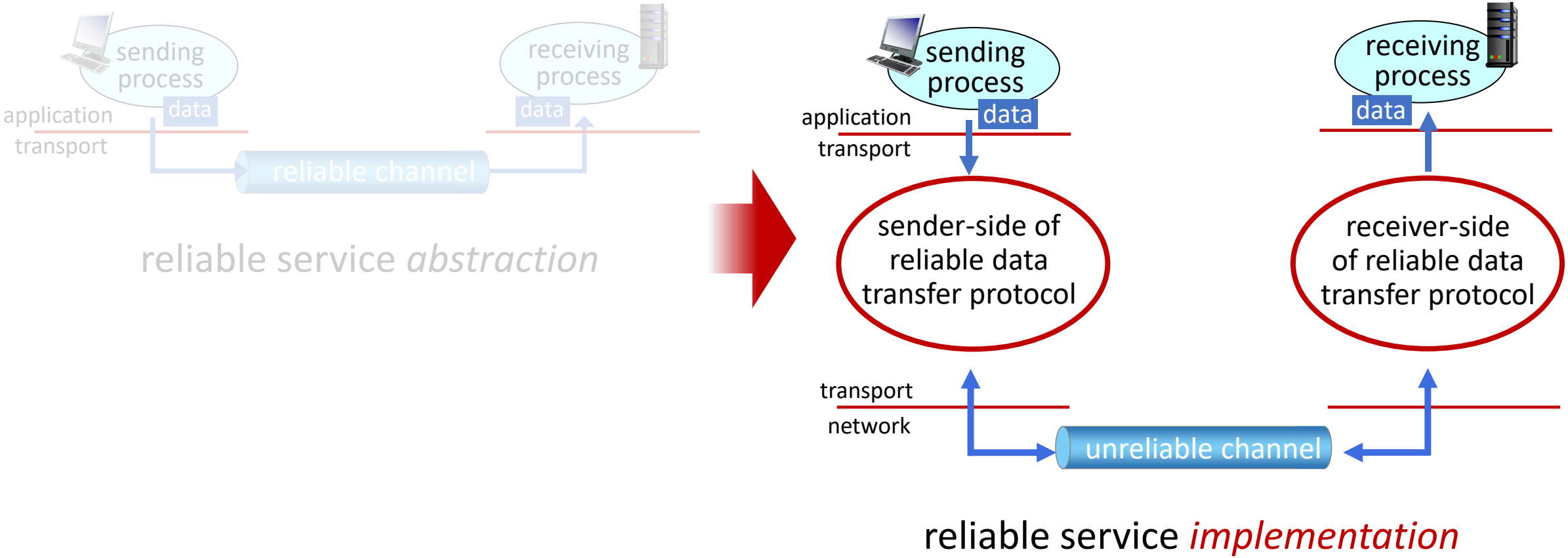


Principles of Reliable Data Transfer

Principles of reliable data transfer

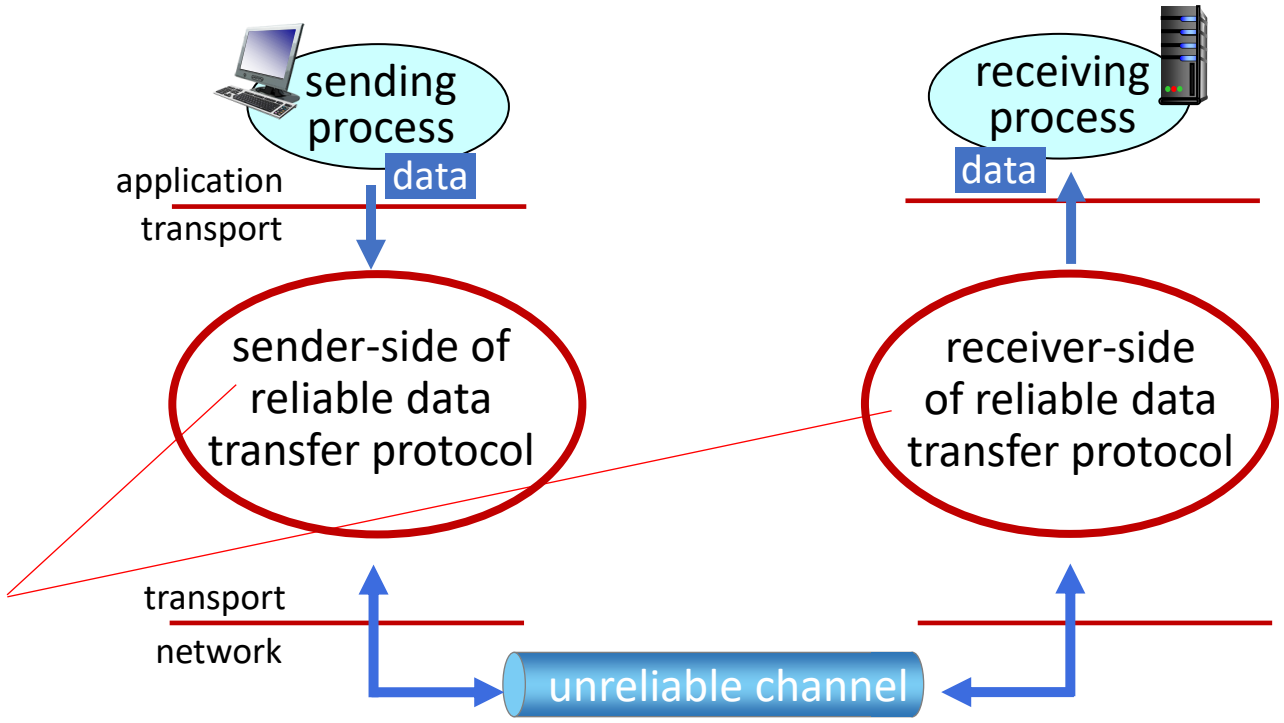
reliable service *abstraction*

Principles of reliable data transfer



Principles of reliable data transfer

Complexity of reliable data transfer protocol will depend (strongly) on characteristics of unreliable channel (lose, corrupt, reorder data?)

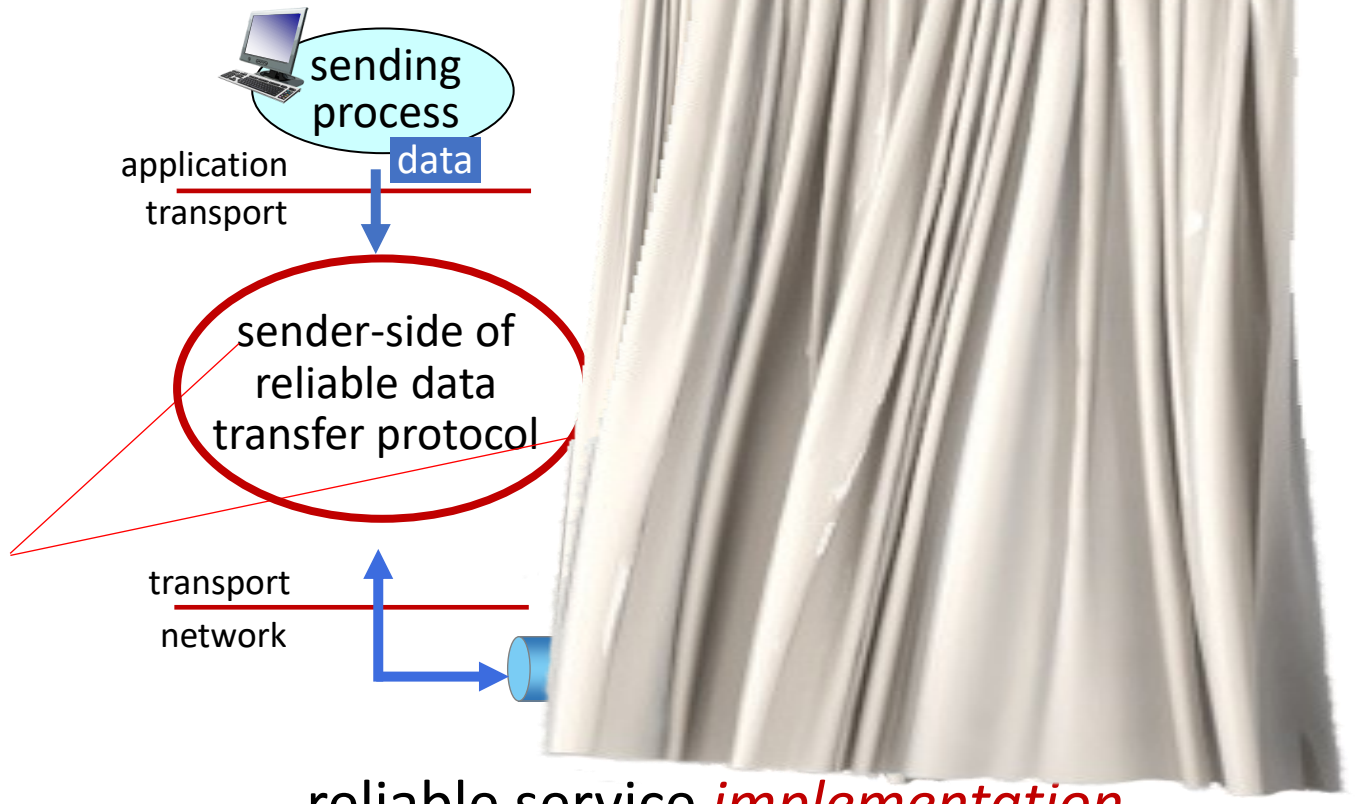


reliable service *implementation*

Principles of reliable data transfer

Sender, receiver do *not* know the "state" of each other, e.g., was a message received?

- unless communicated via a message



reliable service *implementation*

Principles of reliable data transfer

Let’s incrementally add reliability to protocol, so that it deals with the problems

Reliable transfer over a reliable channel

- Sender just transmits a packet, and it will get at the receiver
- Packet is never corrupted
- Packet is never lost
- Packet arrives within a finite amount of time (eventually)
- There is never a problem 🙂
- If so, the reliable protocol does not need to do anything

- This is not so crazy… 🙂 it is possible that reliability is provided by some lower layer
- Like from Web HTTP point-of-view, this is like what TCP provides!

A channel with bit errors

- underlying channel may flip bits in packet
  - checksum (e.g., Internet checksum) to detect bit errors
- *the* question: how to recover from errors?

How do humans recover from "errors" during conversation?

A channel with bit errors

- underlying channel may flip bits in packet
  - checksum to detect bit errors
- *the* question: how to recover from errors?
  - *acknowledgements (ACKs):* receiver explicitly tells sender that pkt received OK
  - *negative acknowledgements (NAKs):* receiver explicitly tells sender that pkt had errors
  - sender *retransmits* pkt on receipt of NAK

stop and wait

sender sends one packet, then waits for receiver response

A channel with bit errors

what happens if ACK/NAK corrupted?

- sender doesn't know what happened at receiver!
- can't just retransmit: possible duplicate

handling duplicates:

- sender retransmits current pkt if ACK/NAK corrupted
- sender adds *sequence number* to each pkt
- receiver discards (doesn't deliver up) duplicate pkt

stop and wait

sender sends one packet, then waits for receiver response

A channel with bit errors and losses (rdt3.0)

New channel assumption: underlying channel can also *lose* packets (data, ACKs)

- checksum, sequence #s, ACKs, retransmissions will be of help … but not quite enough

Q: How do *humans* handle lost sender-to-receiver words in conversation?

A channel with bit errors and losses (rdt3.0)

Approach: sender waits "reasonable" amount of time for ACK

- retransmits if no ACK received in this time
- if pkt (or ACK) just delayed (not lost):
  - retransmission will be duplicate, but seq #s already handles this!
  - receiver must specify seq # of packet being ACKed
- use countdown timer to interrupt after "reasonable" amount of time

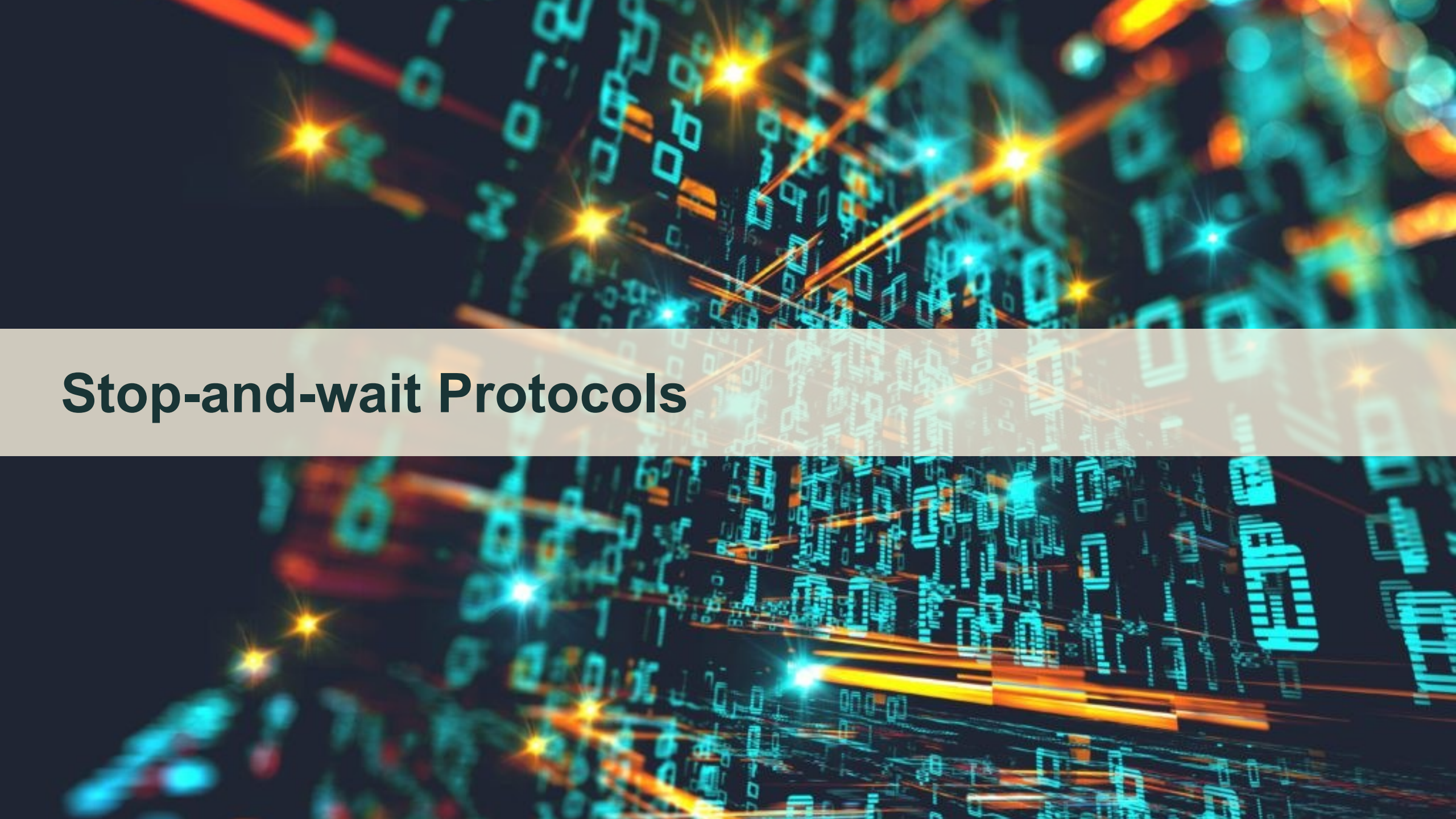
timeout

Reliability mechanisms - (rdt3.0) summary

- Checksums to detect packet corruption, in both directions!
- Negative confirmations (NACKs) sent by receiver to inform sender that packet was received with error
- Retransmissions to recover from errors
- Positive confirmations (ACKs) sent by receiver to inform the sender that packet was received correctly
- Sequence numbers to identify data and deal with duplicates

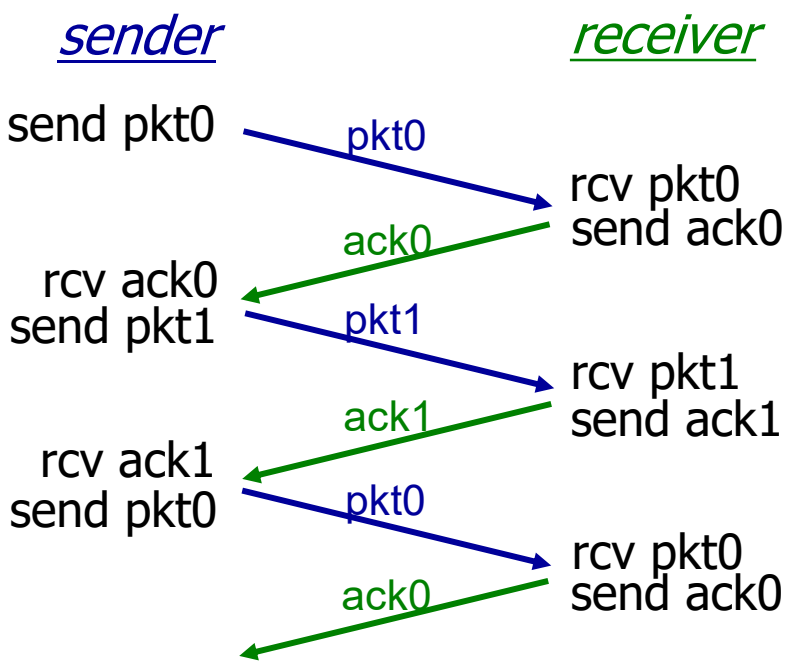
Reliability mechanisms - (rdt3.0) summary

- Timeouts at the sender to deal with loss of packets transmitted to the receiver
- How about lost ACKs/NACKs, what happens?
- Stop and wait!
- Can the protocol use only ACKs, and silently discard any corrupted packets?
- What if there is no feedback channel? Can the sender simply send several times the same data?

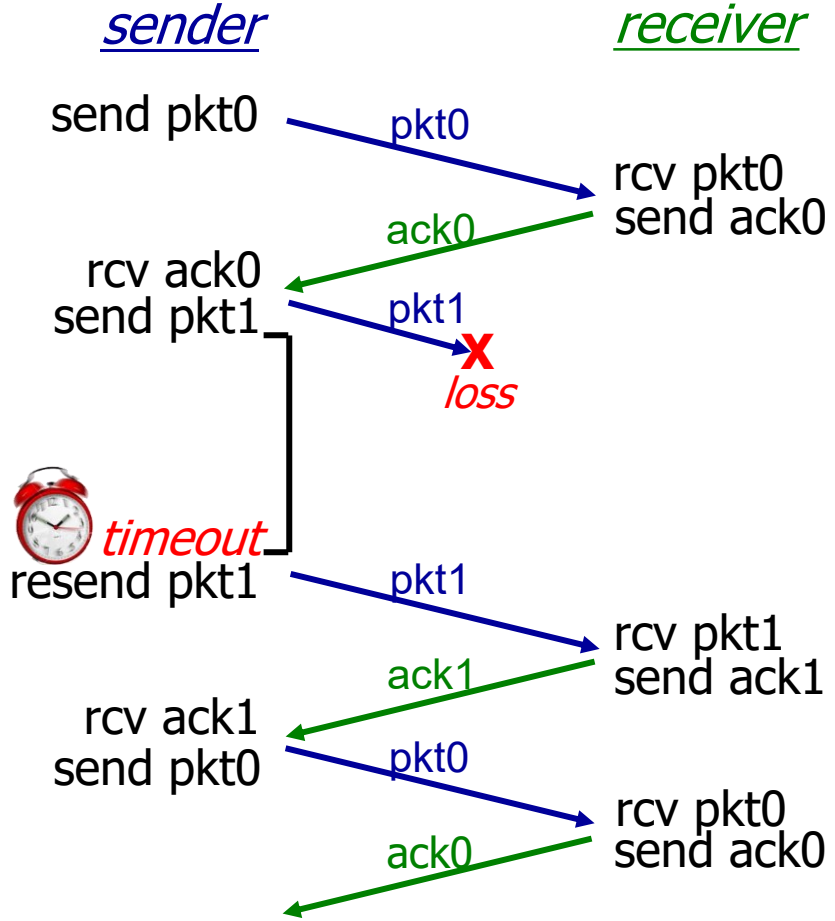


Stop-and-wait Protocols

rdt3.0 in action: Stop-and-wait protocol operation

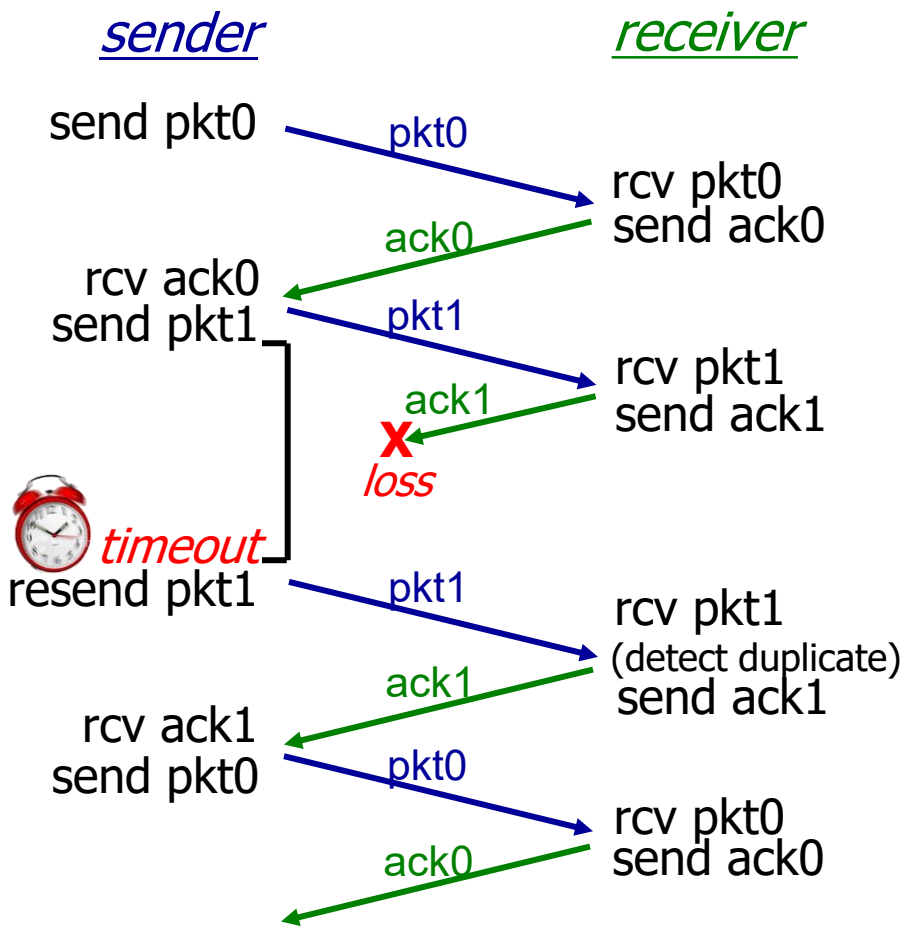


(a) no loss

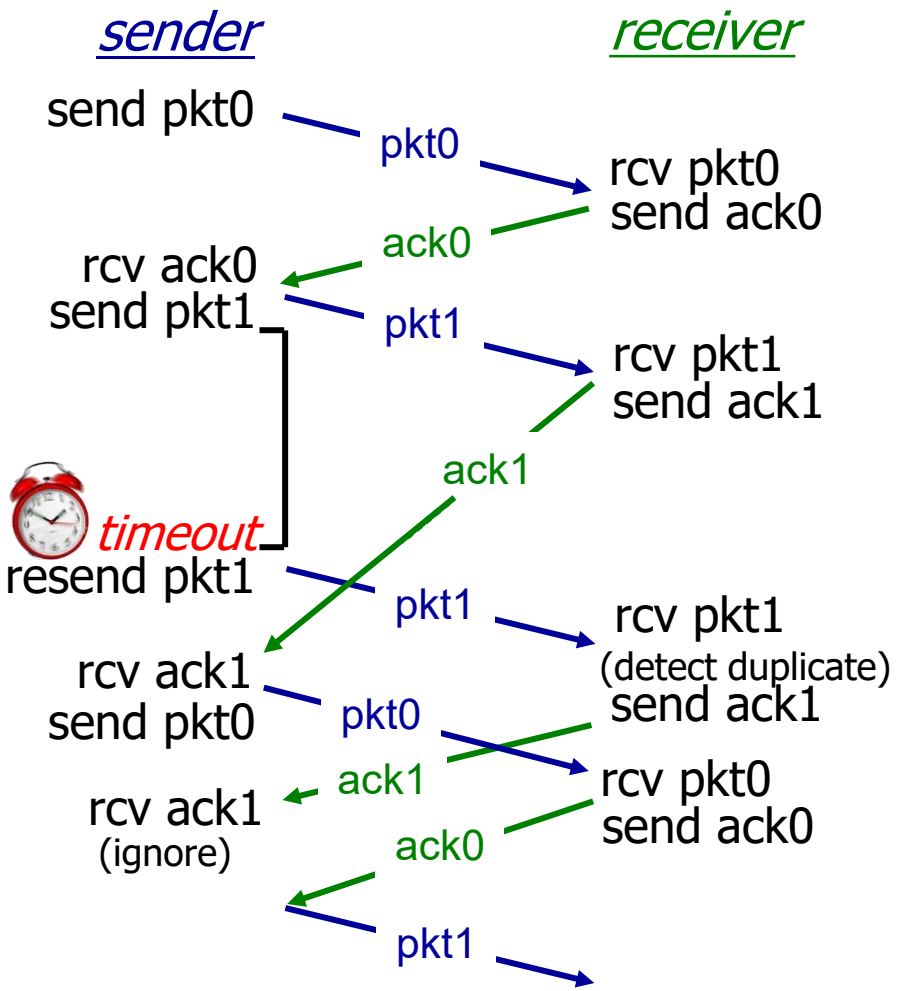


(b) packet loss

rdt3.0 in action: stop-and-wait protocol operation



(c) ACK loss



(d) premature timeout/ delayed ACK

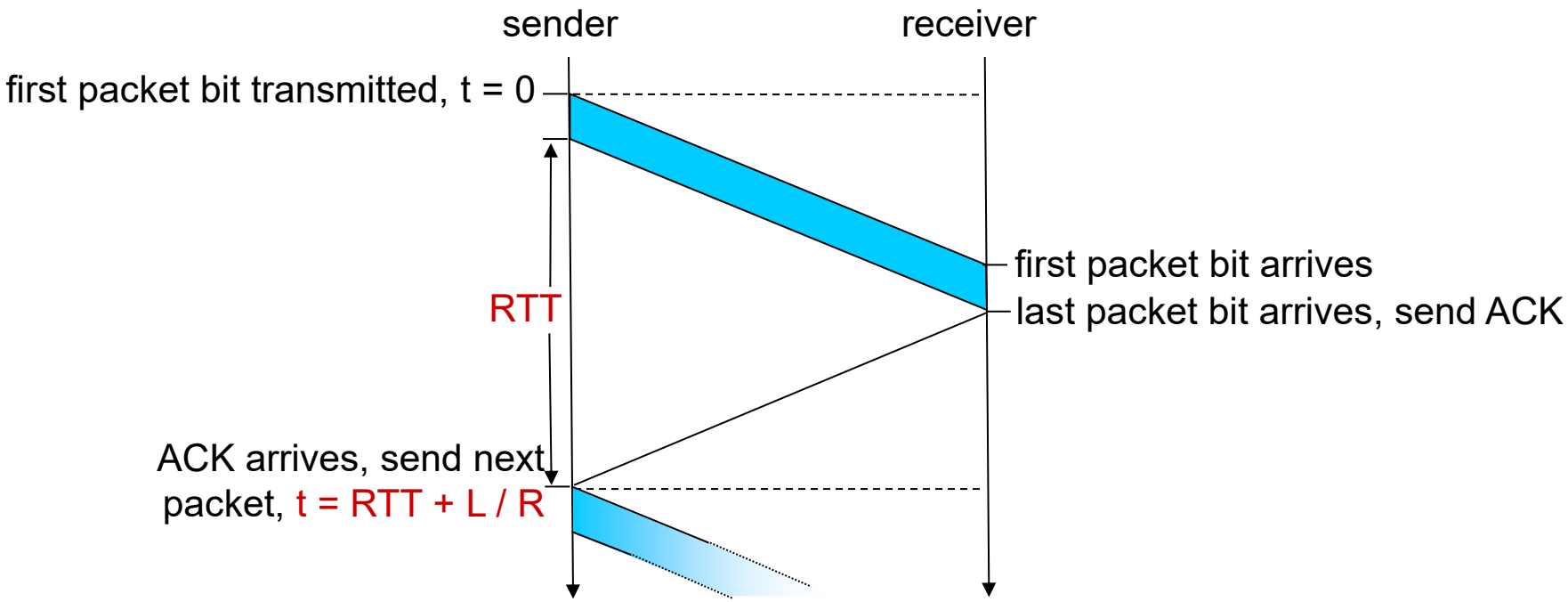
rdt3.0: stop-and-wait performance

- U_{sender}: *utilization* – fraction of time sender busy sending

- example: 1 Gbps link, 15 ms prop. delay, 8000 bit packet
  - time to transmit packet into channel:

$$D_{trans} = \frac{L}{R} = \frac{8000 \text{ bits}}{10^9 \text{ bits/sec}} = 8 \text{ microsecs}$$

rdt3.0: stop-and-wait operation

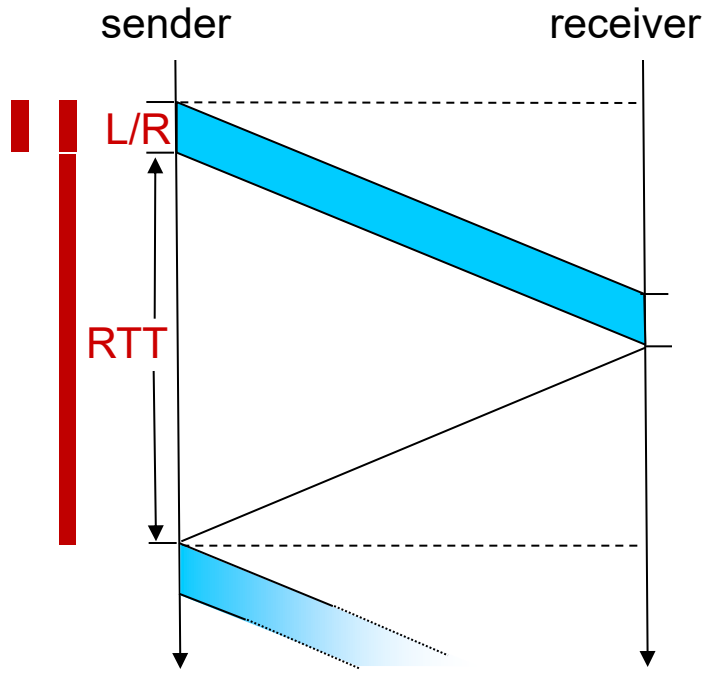


rdt3.0: stop-and-wait operation

$$U_{sender} = \frac{L / R}{RTT + L / R}$$

$$= \frac{.008}{30.008}$$

$$= 0.00027$$



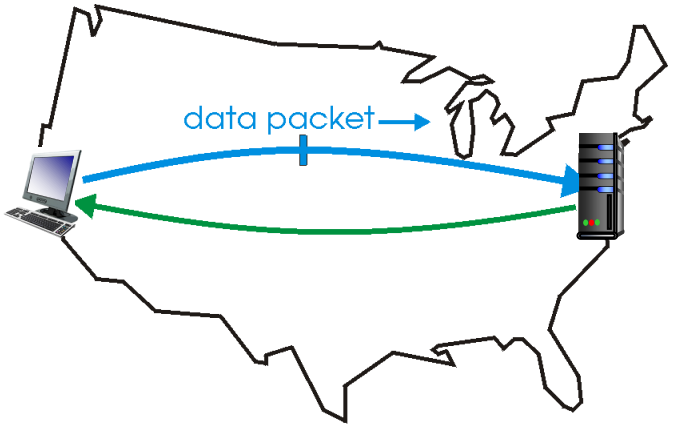
- rdt 3.0 (stop-and-wait operation) protocol performance stinks!
- Protocol limits performance of underlying infrastructure (channel)

Sliding Windows

rdt3.0: From stop-and-wait to sliding window protocols

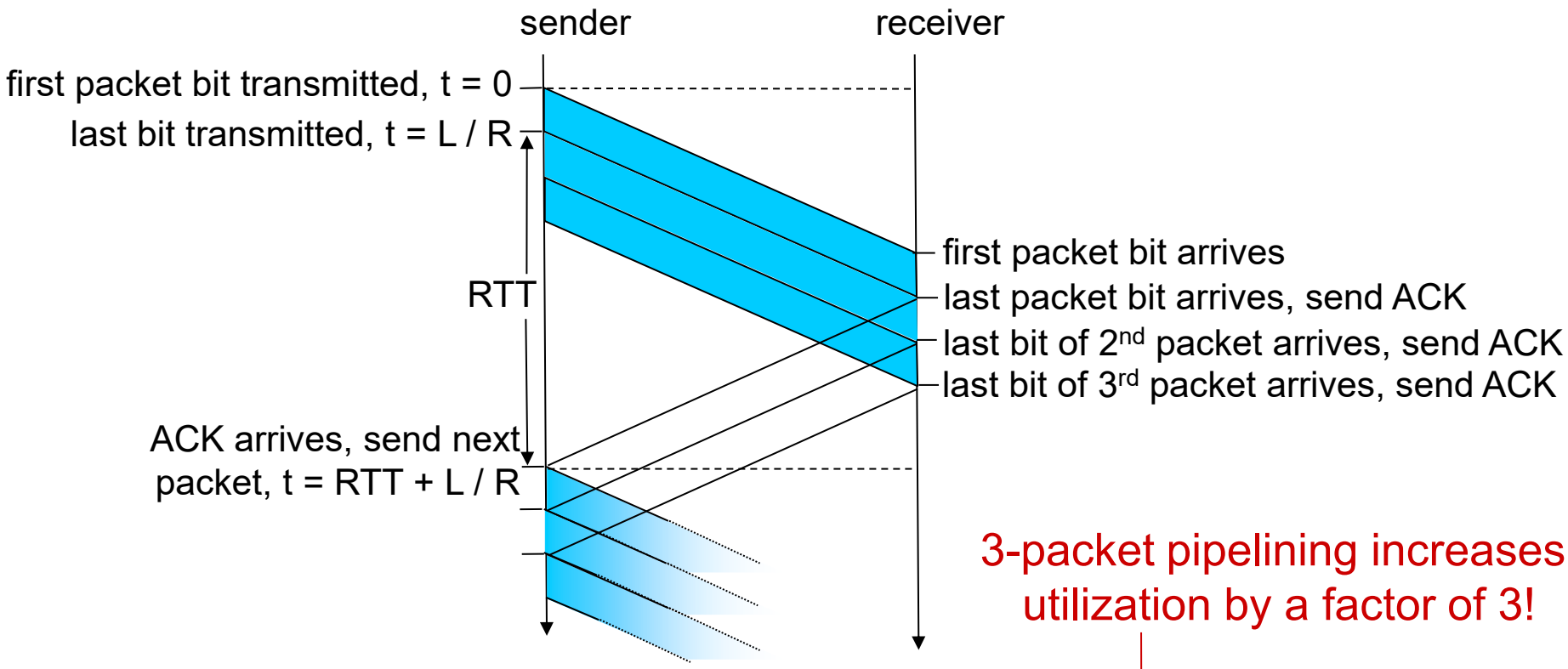
Sliding window/pipelining: sender allows multiple, “in-flight”, yet-to-be-acknowledged packets

- range of sequence numbers must be increased
- buffering at sender and/or receiver



(a) a stop-and-wait protocol in operation

Sliding window protocols: increased utilization



3-packet pipelining increases utilization by a factor of 3!

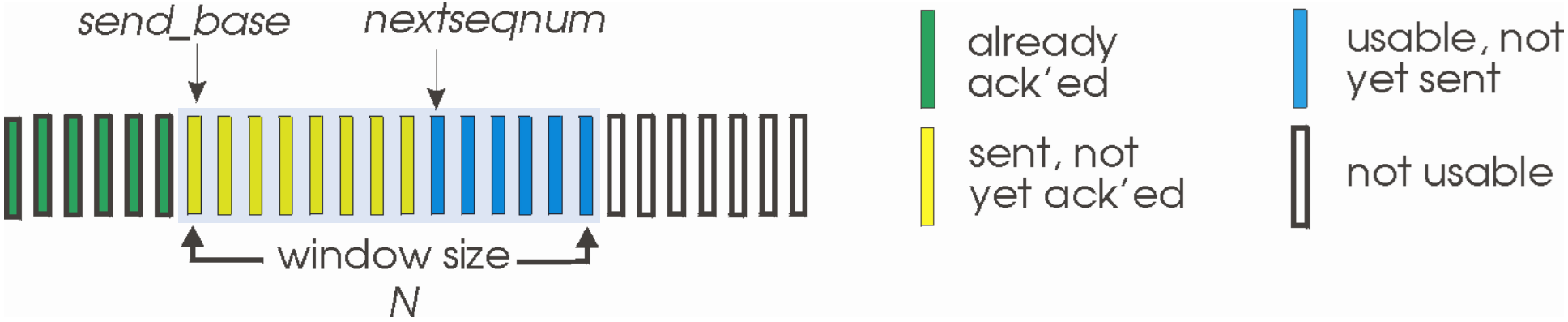
$$U_{sender} = \frac{3L/R}{RTT + L/R} = \frac{.0024}{30.008} = 0.00081$$

Sliding window protocols: two types

Go-back-N	Selective Repeat
Sender can have up to N unACKed packets in flight	Sender can have up to N unACKed packets in flight
Receiver only sends **"cumulative" ACKs**, and does not send ACK if there's a gap in sequence	Receiver sends **individual** ACK for each packet
Sender has **(single) timer**, for oldest unacked packet	Sender maintains one **timer for each unACKed packet**
When timer expires, sender retransmits **all** unACKed packets	When timer expires, retransmit **only** that unACKed packet

Go-Back-N: sender

- sender: “window” of up to N, consecutive transmitted but unACKed pkts
  - Sender places a seq # in packet header to identify the data

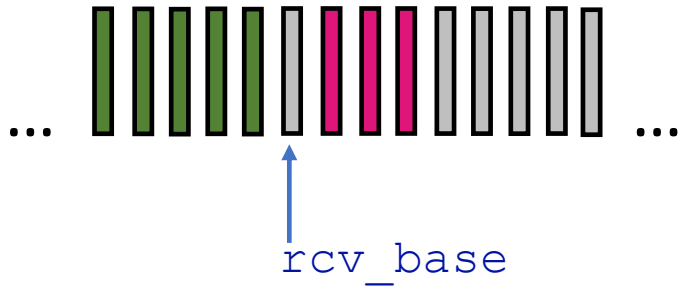


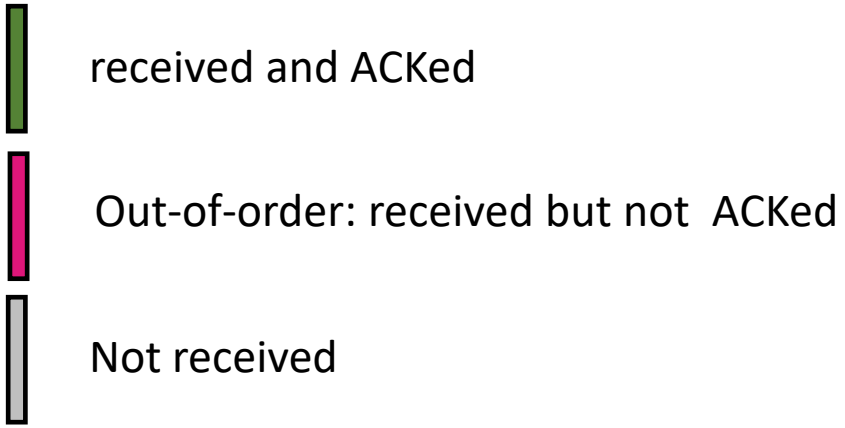
- *cumulative ACK:* ACK(*n*): ACKs all packets up to, including seq # *n*
  - On receiving packet ACK(n): sender slides window forward to start at n+1.
- Sender keeps a timer for oldest "in-flight packet”, i.e. unACKed
- When there is a timeout (at the sender), the sender retransmits oldest unconfirmed packet and all other subsequent packets (higher seq) in window

Go-Back-N: receiver

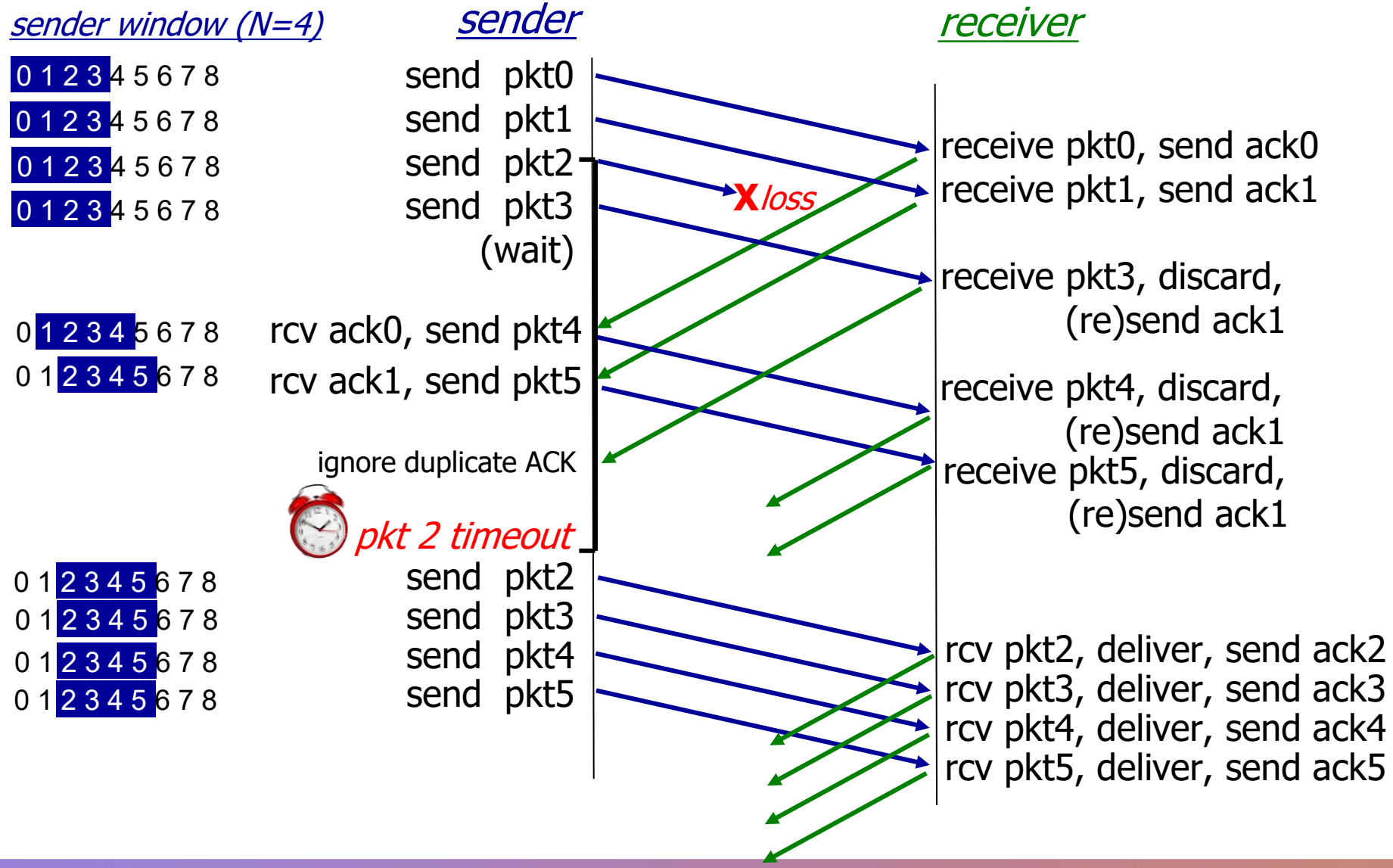
- ACK-only: always send ACK for correctly-received packet so far, with highest in-order packet
  - may generate duplicate ACKs
  - need only remember `rcv_base`
- on receipt of out-of-order packet:
  - can discard (don't buffer) or buffer: an implementation decision
  - re-ACK pkt with highest in-order seq #

Receiver view of sequence number space:

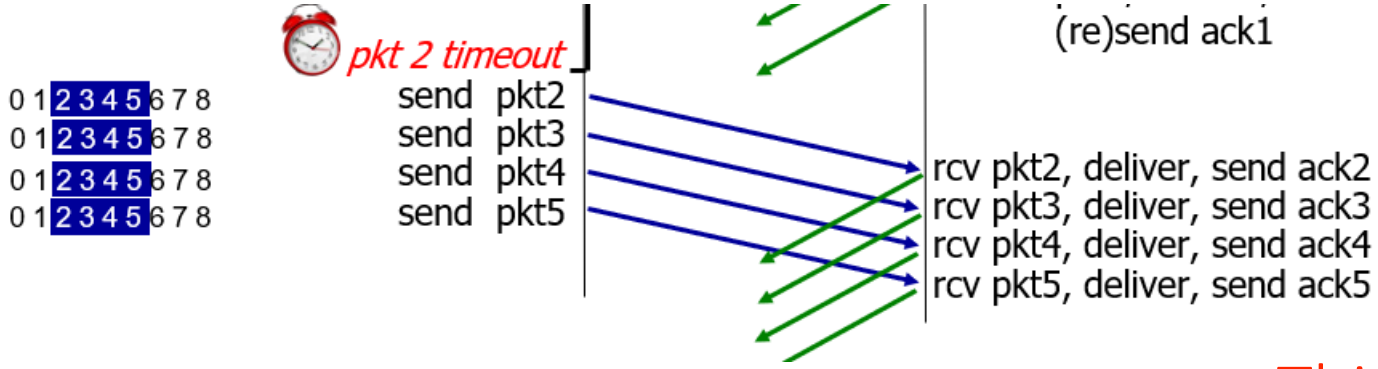




Go-Back-N in action



Go-Back-N in action

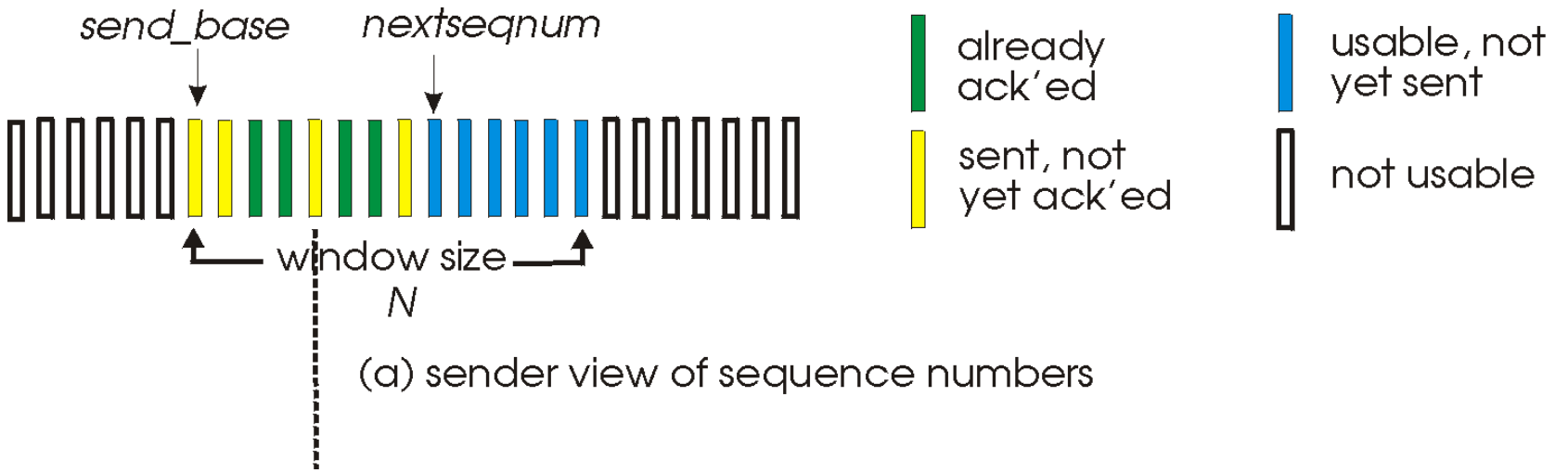


This looks inefficient, but if the receiver discards packets, then it's necessary

Selective repeat: the approach

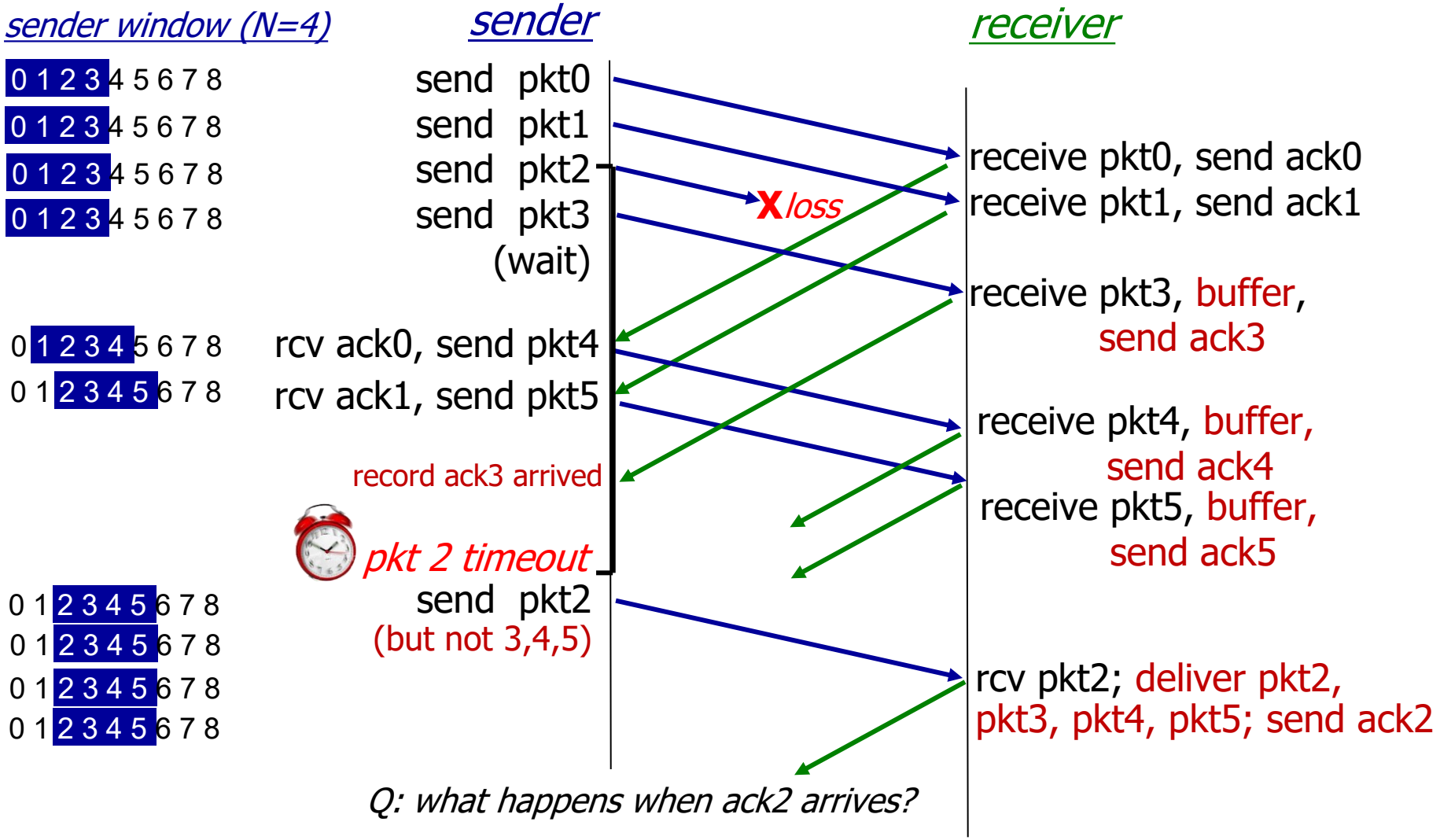
- *Window/pipelining*: *multiple* packets in flight
- *receiver individually ACKs* all correctly received packets
  - buffers packets, as needed, for in-order delivery to application layer
- sender:
  - maintains (conceptually) a timer for each unACKed pkt
    - timeout: retransmits single unACKed packet associated with timeout
  - maintains (conceptually) "window" over *N* consecutive seq #s
    - Window limits the sequence numbers of packets sent, according to unACKed packets

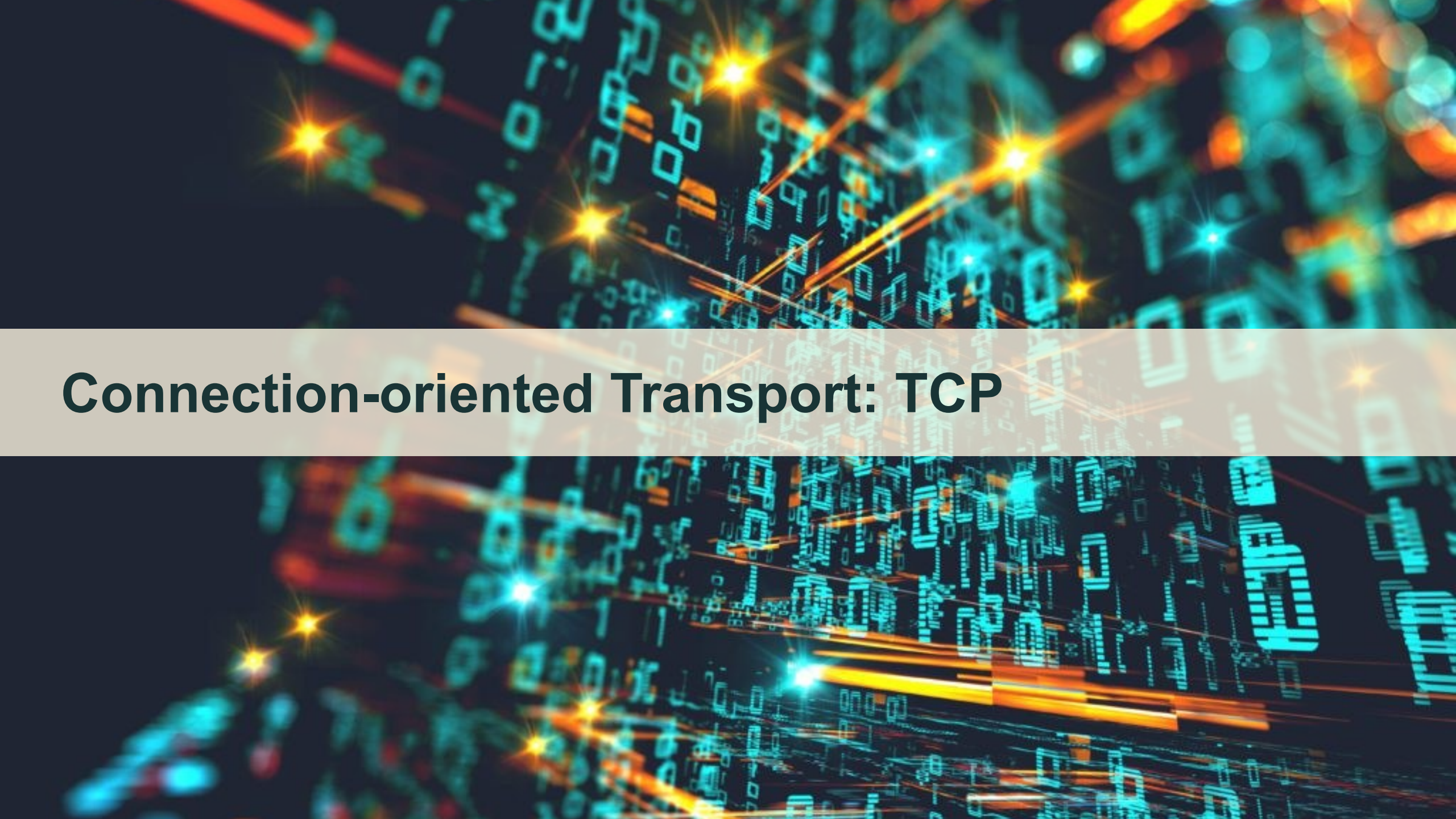
Selective repeat: sender, receiver windows



(a) sender view of sequence numbers

Selective Repeat in action



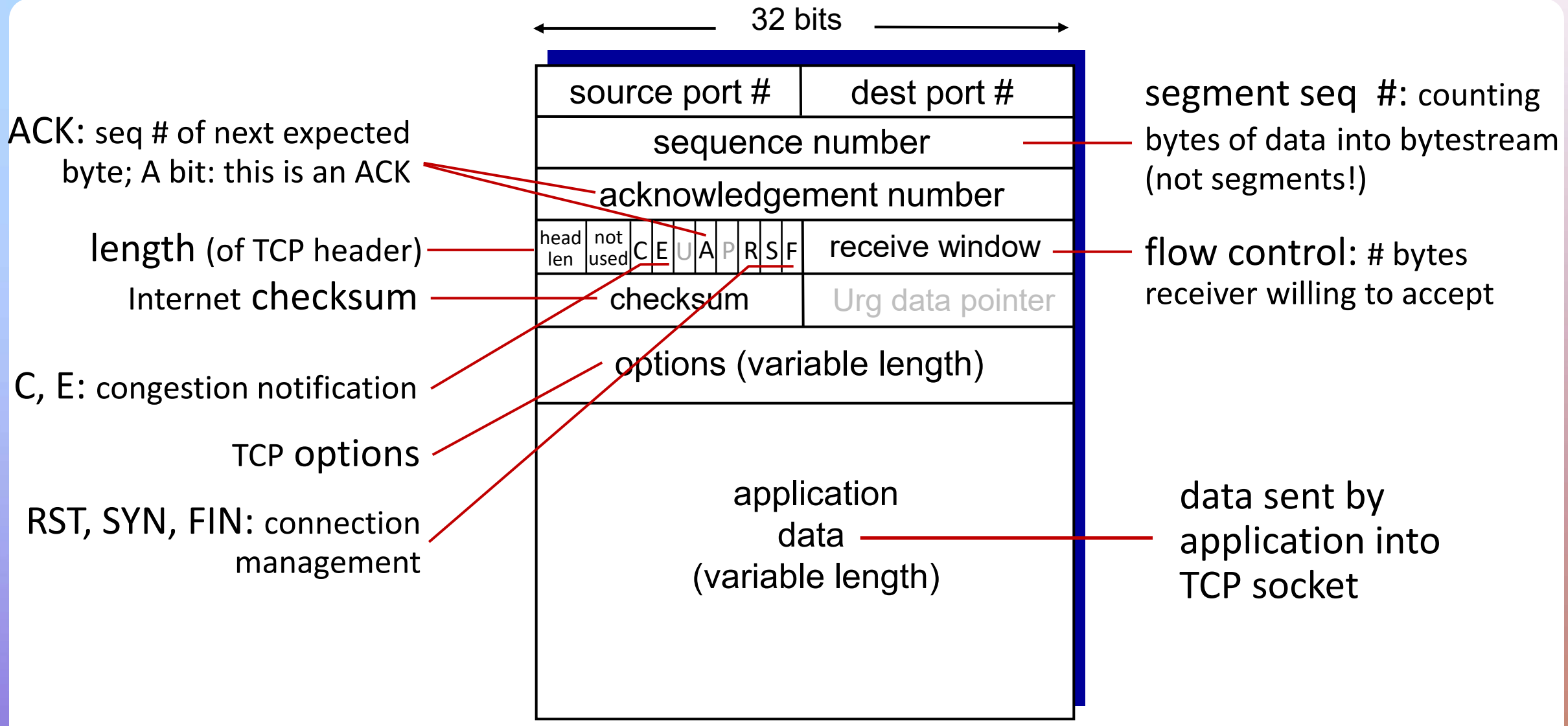


Connection-oriented Transport: TCP

TCP: overview

- point-to-point:
  - one sender, one receiver
- reliable, in-order *byte steam:*
  - no “message boundaries"
- full duplex data:
  - bi-directional data flow in same connection
  - MSS: maximum segment size
- cumulative ACKs
- pipelining:
  - TCP congestion and flow control set window size
- connection-oriented:
  - client and sender do handshake before exchanging data
- flow controlled:
  - sender will not overwhelm receiver

TCP segment structure



TCP sequence numbers, ACKs

Sequence numbers:

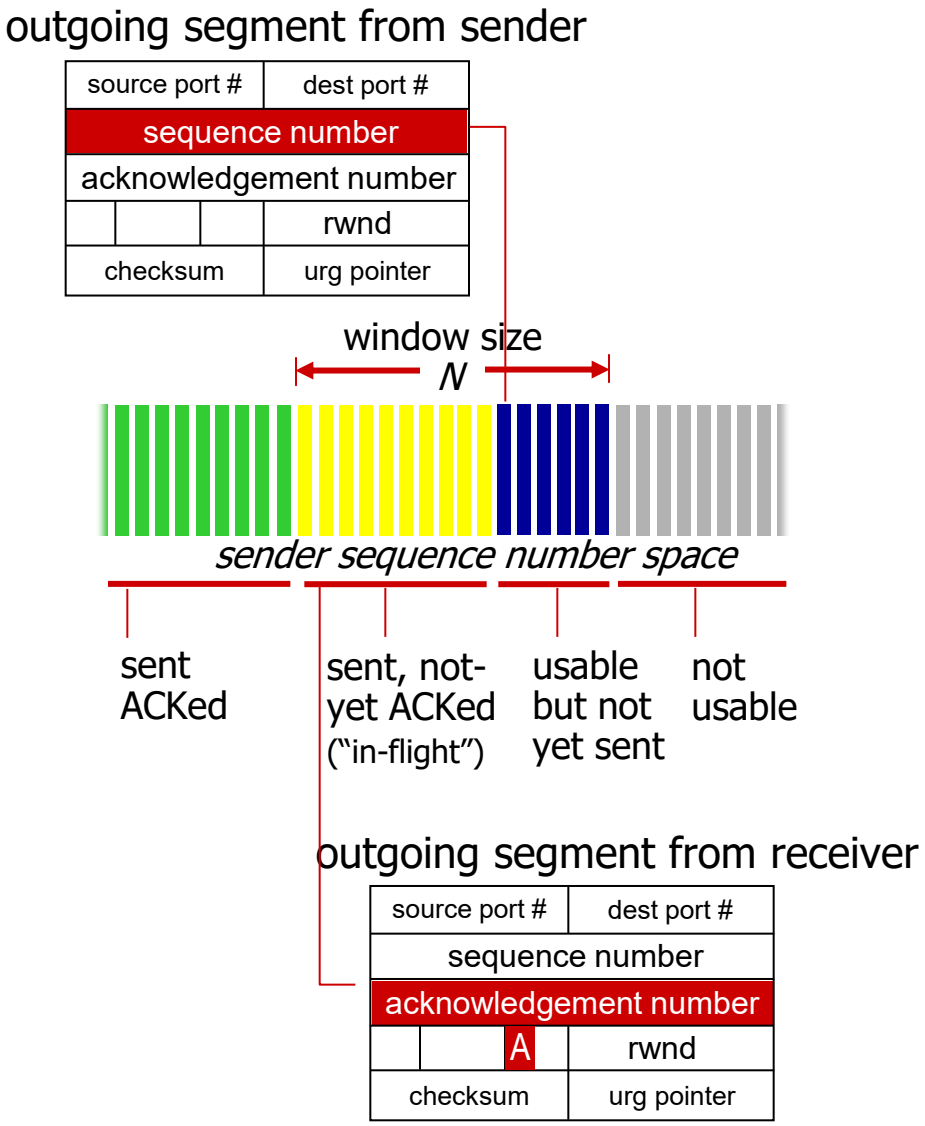
- byte stream "number" of first byte in segment's data

Acknowledgements:

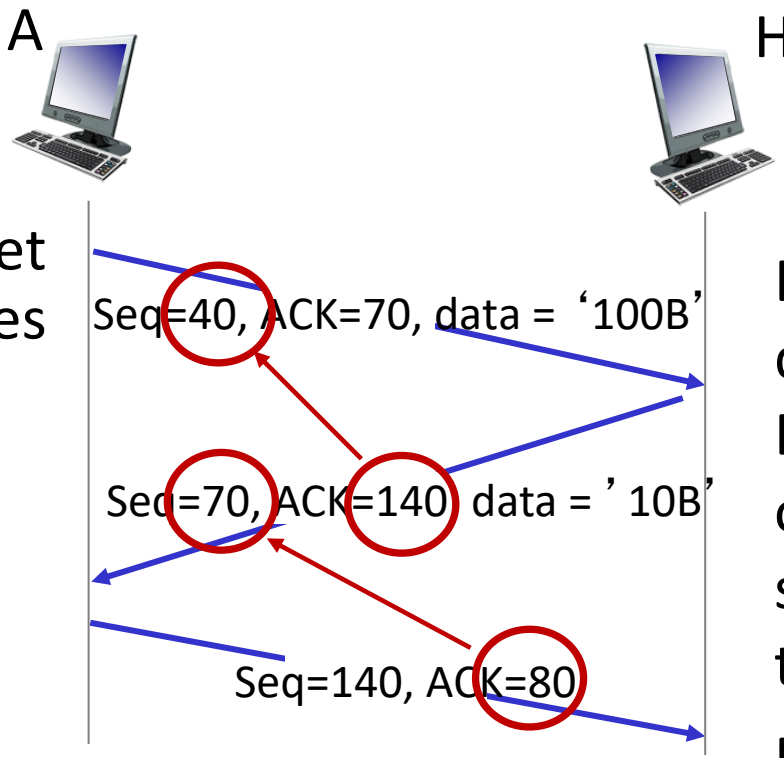
- seq # of next byte expected from other side
- cumulative ACK

Q: how receiver handles out-of-order segments

- *A:* TCP spec doesn't say, - up to implementor



TCP sequence numbers, ACKs



A transmits packet with 100 bytes

A gets confirmation that the 100bytes were received by B, and A also receives 10 bytes of data from B

B gets packet with 100 bytes of data from A.
B sends confirmation of receipt of these 100bytes and in the same packet, sends 10 bytes to A

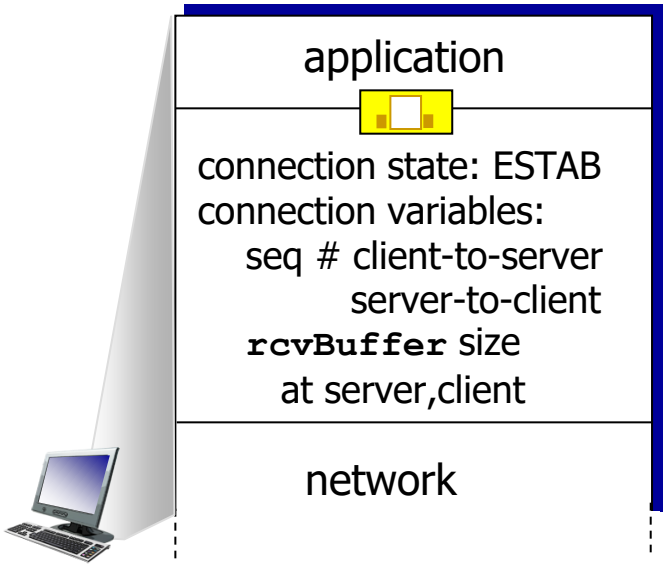
Now B knows that A has received up to 79 and is expecting 80 next

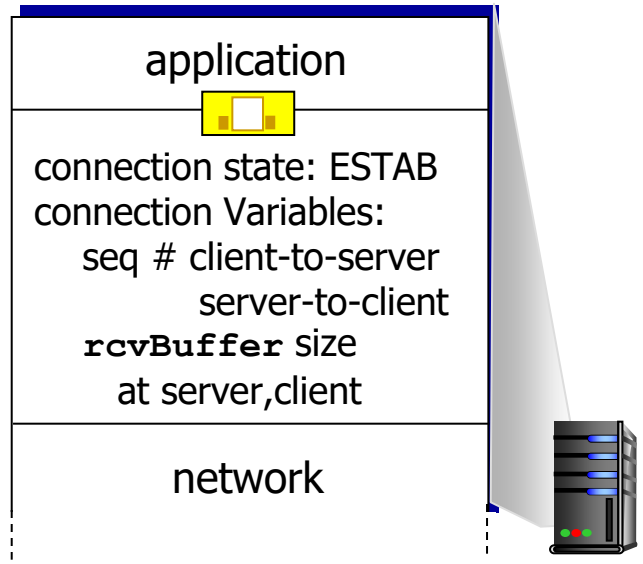
Bi-directional!

TCP connection management

before exchanging data, sender/receiver “handshake”:

- agree to establish connection (each knowing the other willing to establish connection)
- agree on connection parameters (e.g., starting seq #s)






```
Socket clientSocket =
  newSocket("hostname","port number");
```

```
Socket connectionSocket =
  welcomeSocket.accept();
```

TCP 3-way handshake

Client state

```
clientSocket = socket(AF_INET, SOCK_STREAM)
```

LISTEN

```
clientSocket.connect((serverName,serverPort))
```

choose init seq num, x
send TCP SYN msg

SYNSENT

SYNbit=1, Seq=x

Server state

```
serverSocket = socket(AF_INET,SOCK_STREAM)
serverSocket.bind(('',serverPort))
serverSocket.listen(1)
connectionSocket, addr = serverSocket.accept()
```

LISTEN

choose init seq num, y
send TCP SYNACK
msg, acking SYN

SYN RCVD

SYNbit=1, Seq=y
ACKbit=1; ACKnum=x+1

received SYNACK(x)
indicates server is live;
send ACK for SYNACK;
this segment may contain
client-to-server data

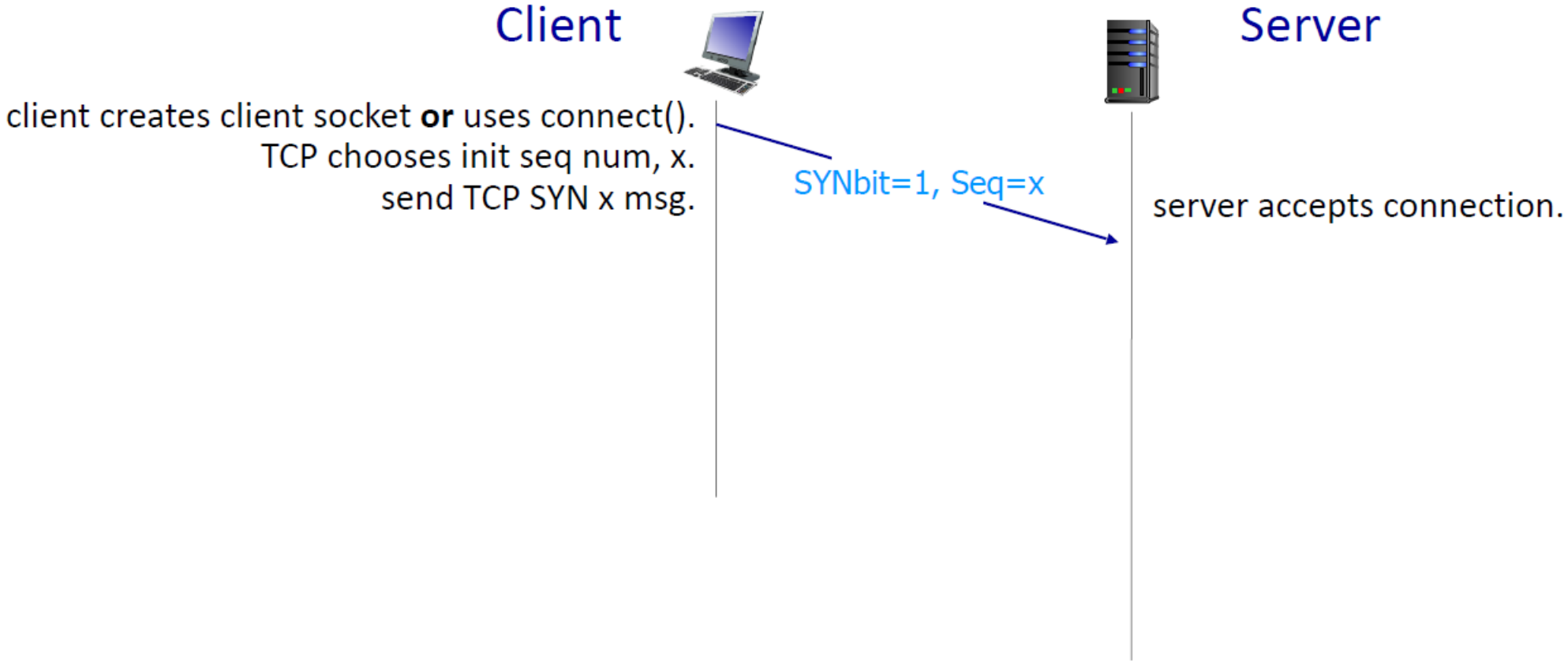
ESTAB

ACKbit=1, ACKnum=y+1

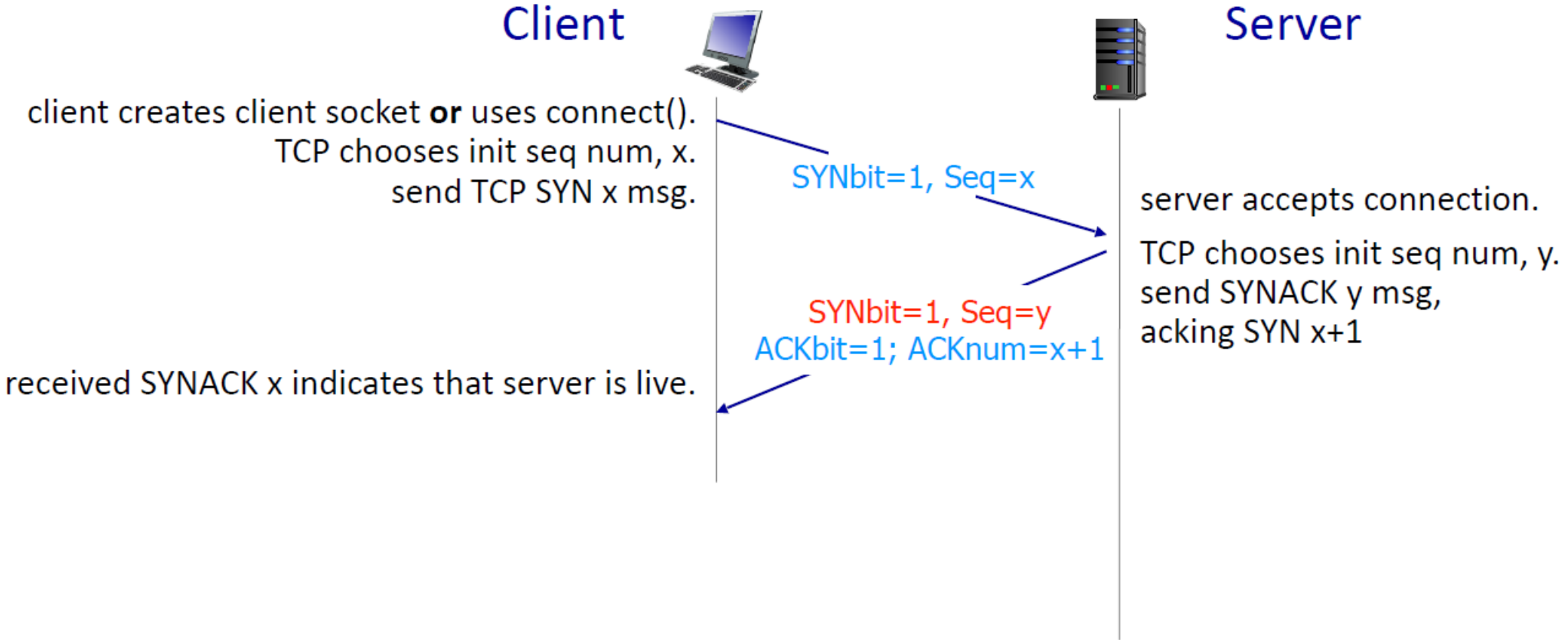
received ACK(y)
indicates client is live

ESTAB

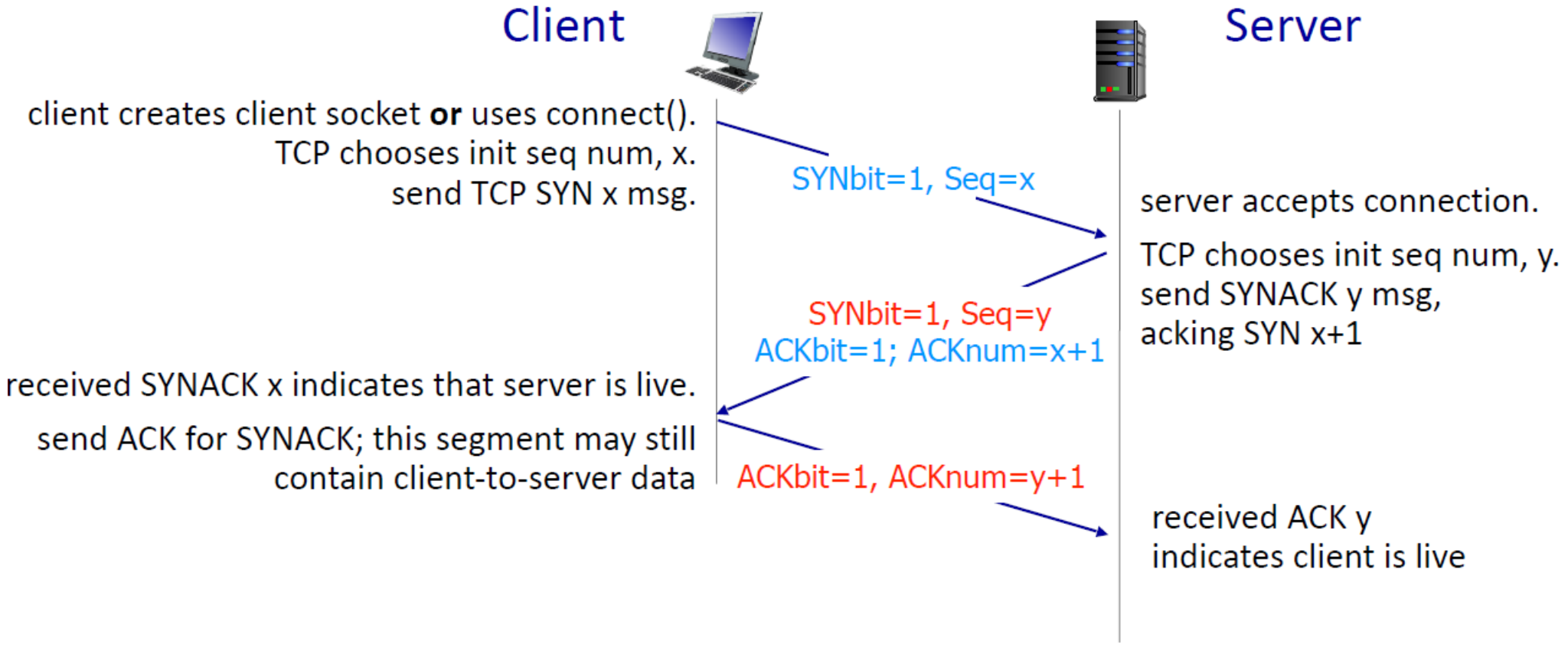
TCP 3-way handshake



TCP 3-way handshake



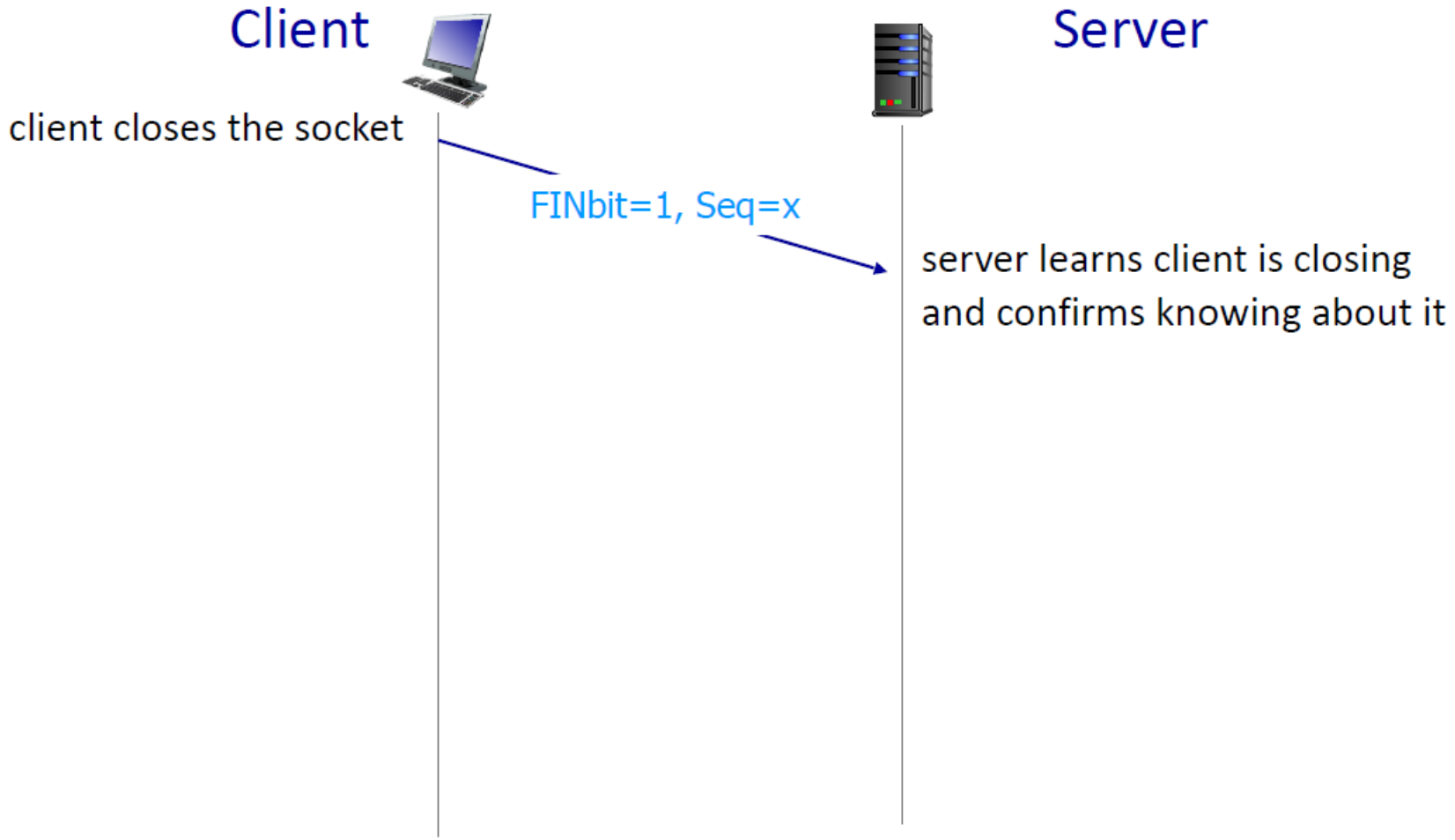
TCP 3-way handshake



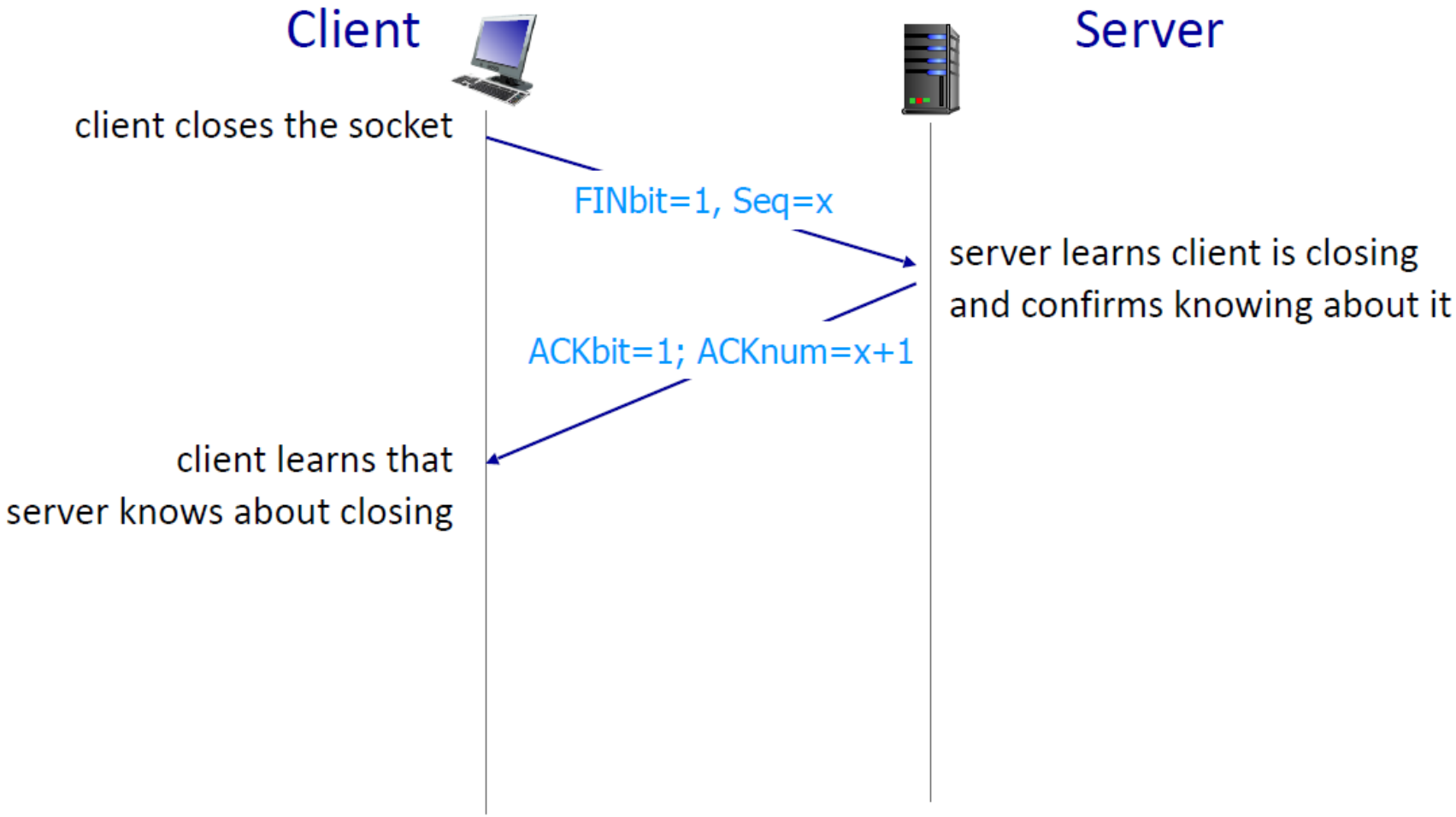
Closing a TCP connection

- client, server each close their side of connection
  - send TCP segment with FIN bit = 1
- respond to received FIN with ACK
  - on receiving FIN, ACK can be combined with own FIN
- simultaneous FIN exchanges can be handled

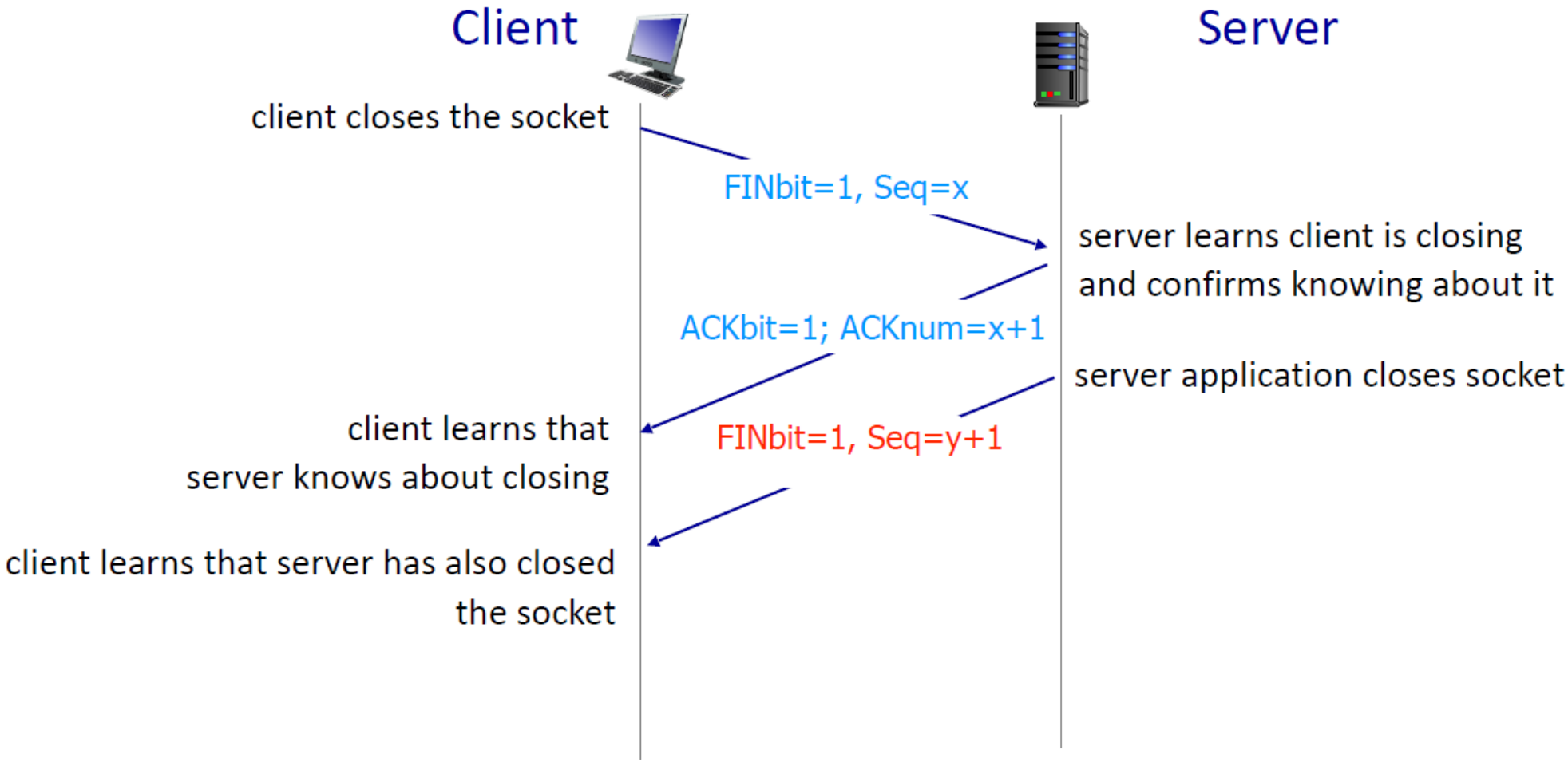
Closing a TCP connection



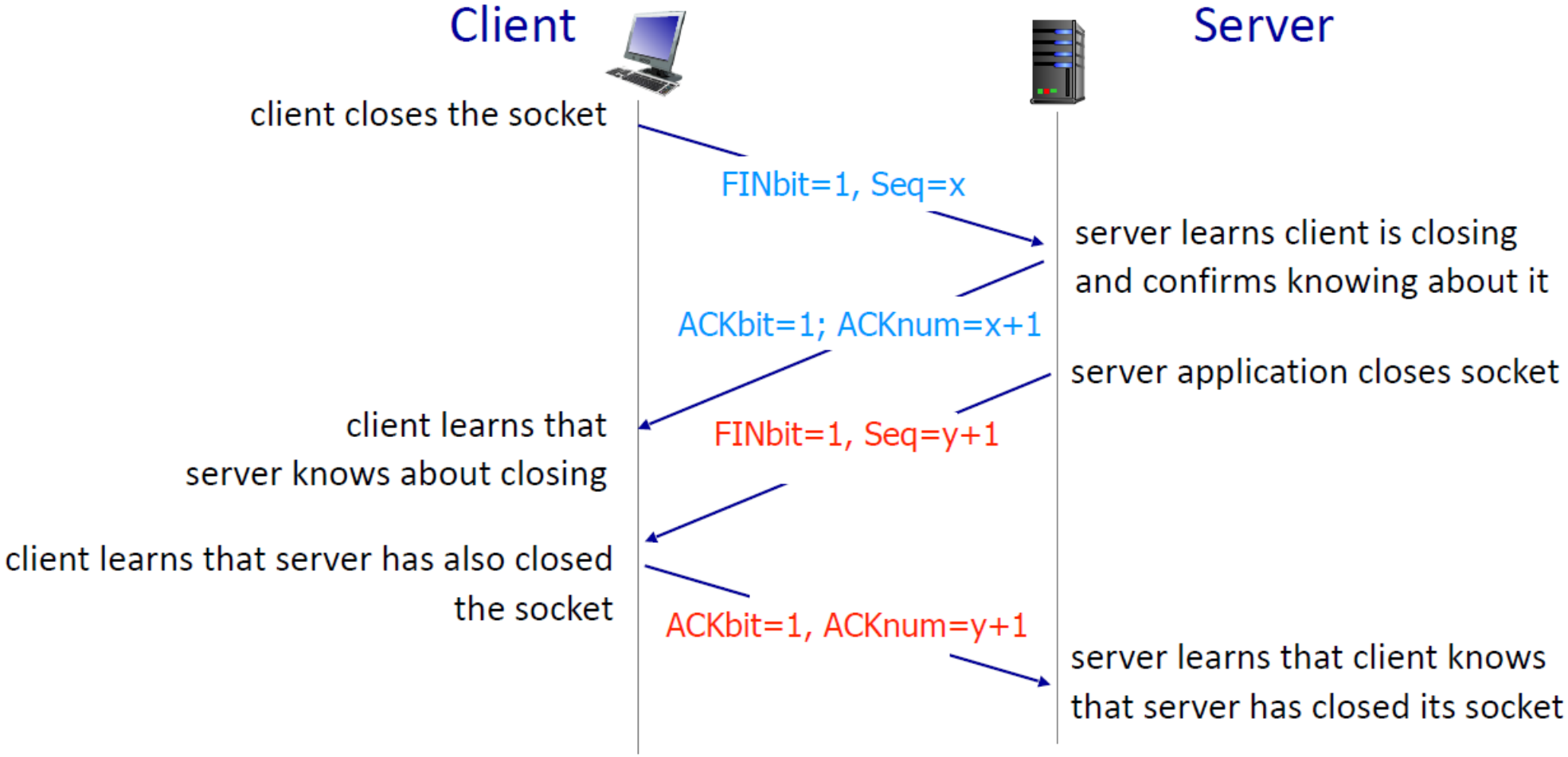
Closing a TCP connection



Closing a TCP connection

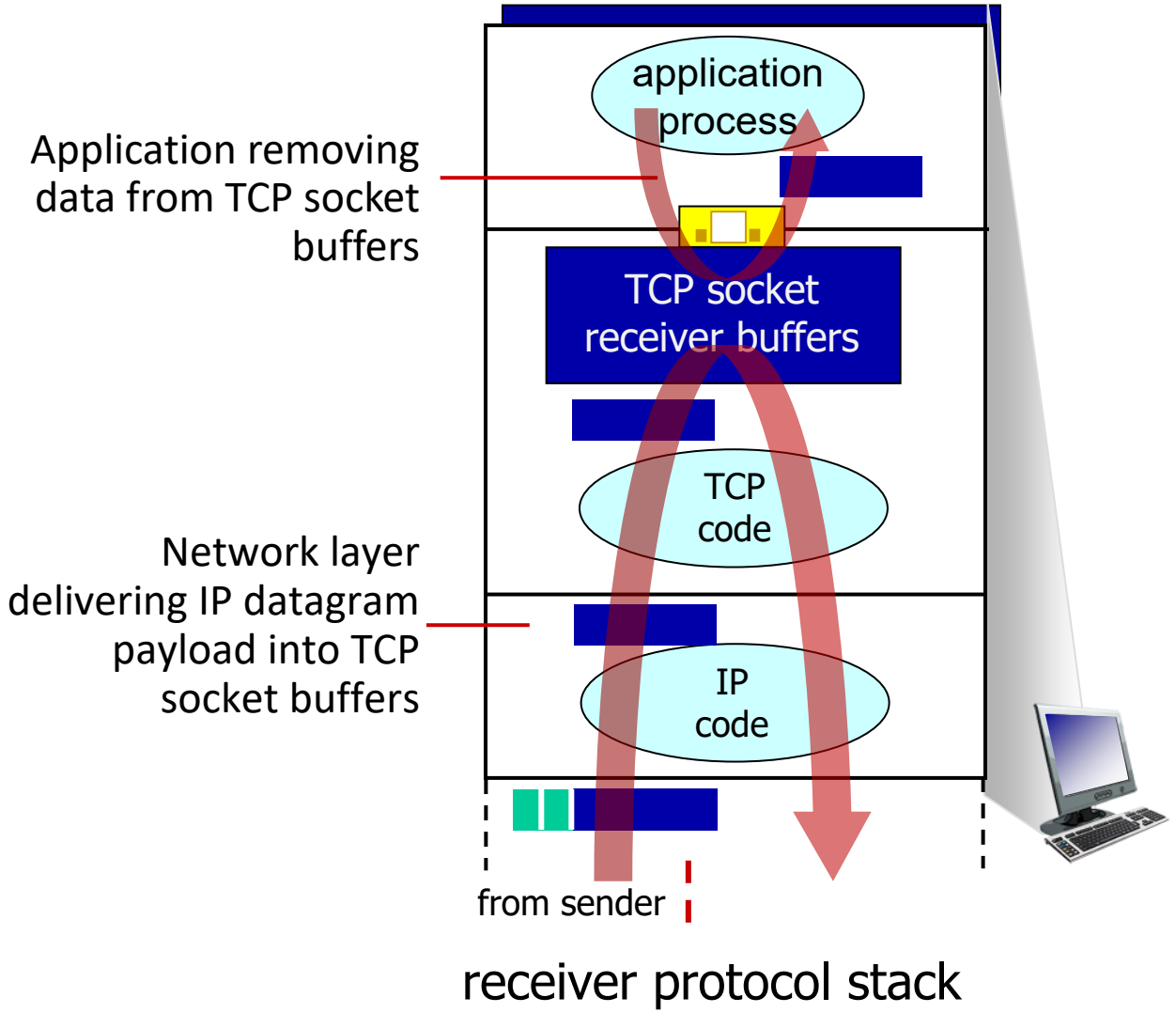


Closing a TCP connection



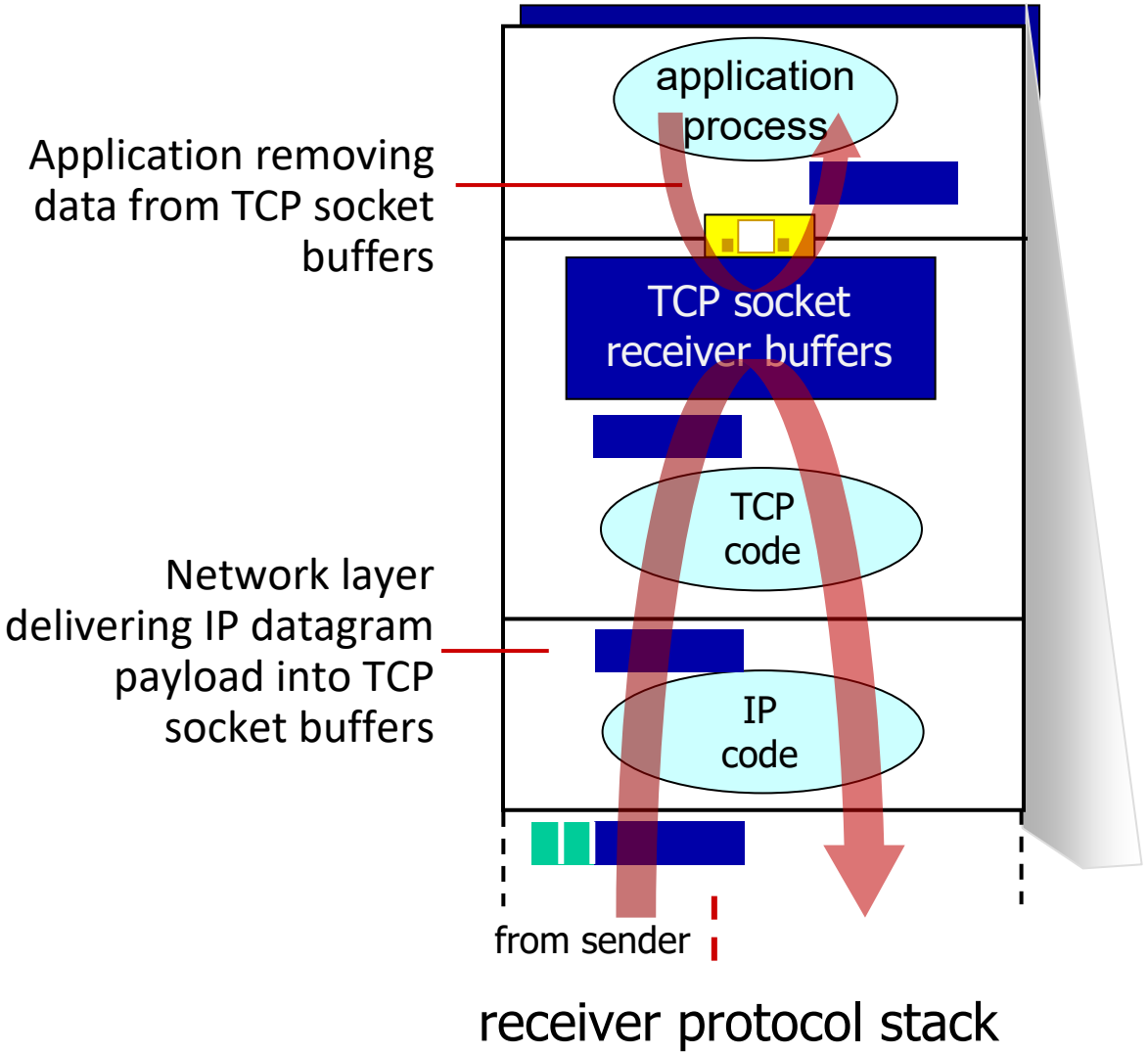
TCP flow control

Q: What happens if network layer delivers data faster than application layer removes data from socket buffers?



TCP flow control

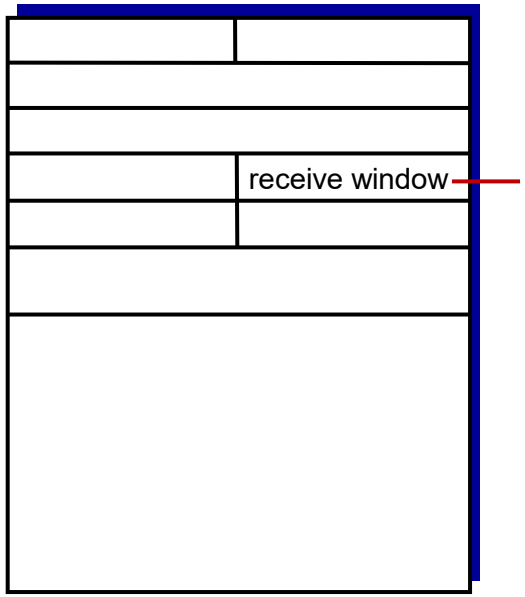
Q: What happens if network layer delivers data faster than application layer removes data from socket buffers?



receiver protocol stack

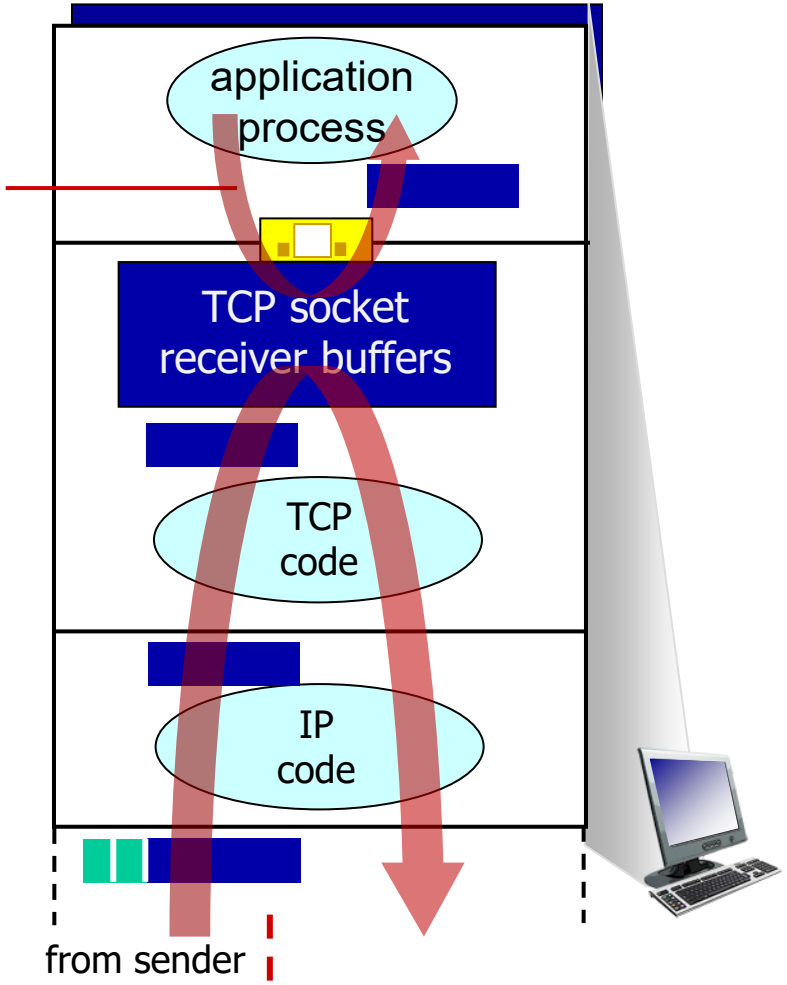
TCP flow control

Q: What happens if network layer delivers data faster than application layer removes data from socket buffers?



flow control: # bytes receiver willing to accept

Application removing data from TCP socket buffers



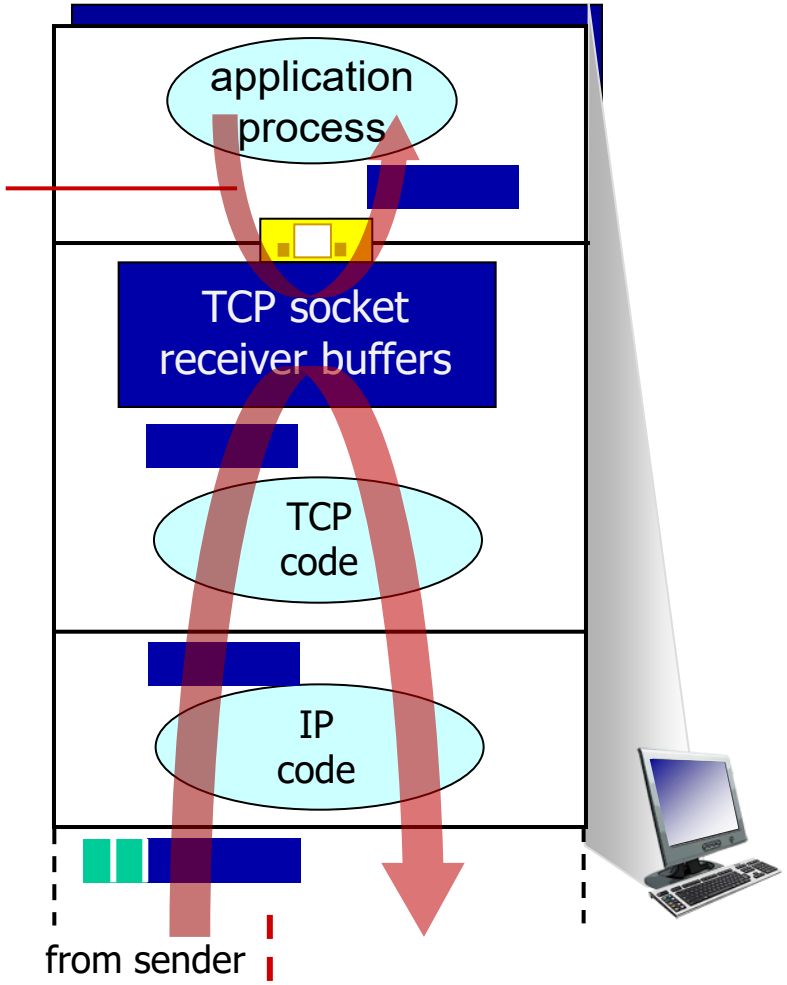
receiver protocol stack

TCP flow control

Q: What happens if network layer delivers data faster than application layer removes data from socket buffers?

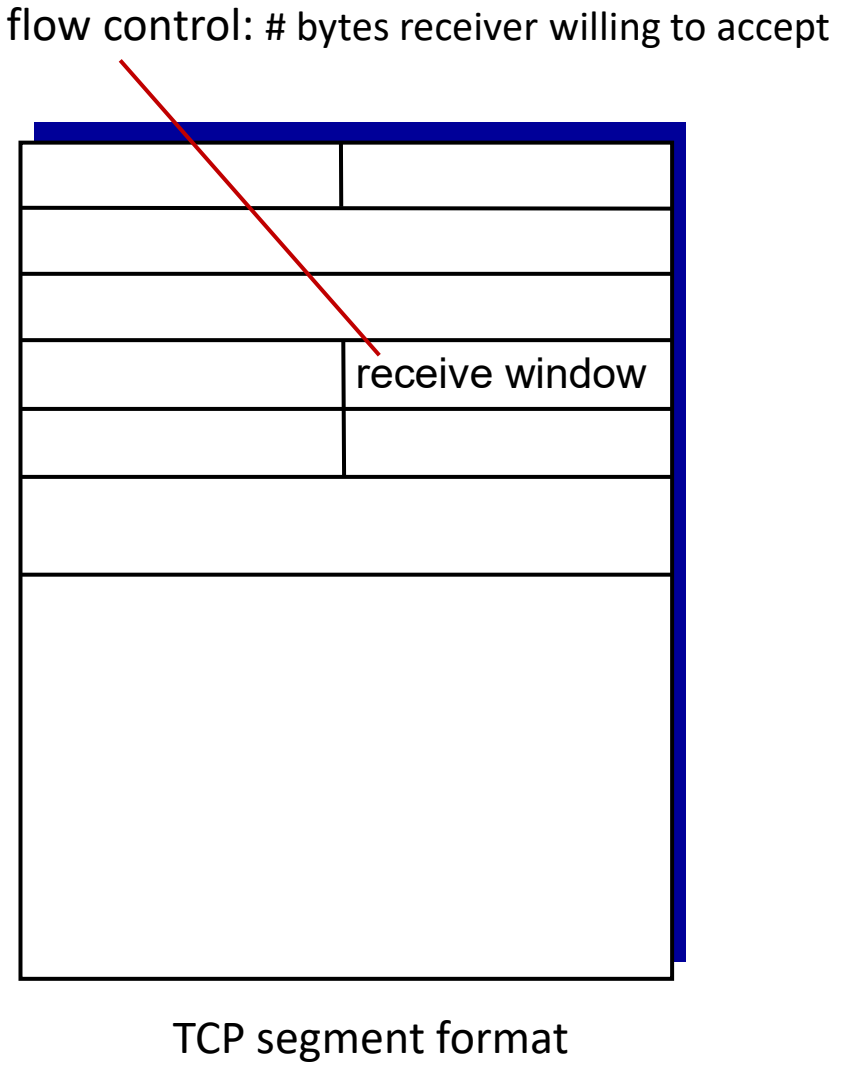
flow control

receiver controls sender, so sender won't overflow receiver's buffer by transmitting too much, too fast



TCP flow control

- TCP receiver “advertises” free buffer space in **`rwnd`** field in TCP header
  - **`RcvBuffer`** size set via socket options (typical default is 4096 bytes)
  - many operating systems auto-adjust **`RcvBuffer`**
- sender limits amount of unACKed (“in-flight”) data to received **`rwnd`**
- guarantees receive buffer will not overflow



Principles of congestion control

Congestion:

- informally: “too many sources sending too much data too fast for *network* to handle”
- manifestations:
  - long delays (queueing in router buffers)
  - packet loss (buffer overflow at routers)
- different from flow control!
- a top-10 problem!

congestion control:
too many senders,
sending too fast

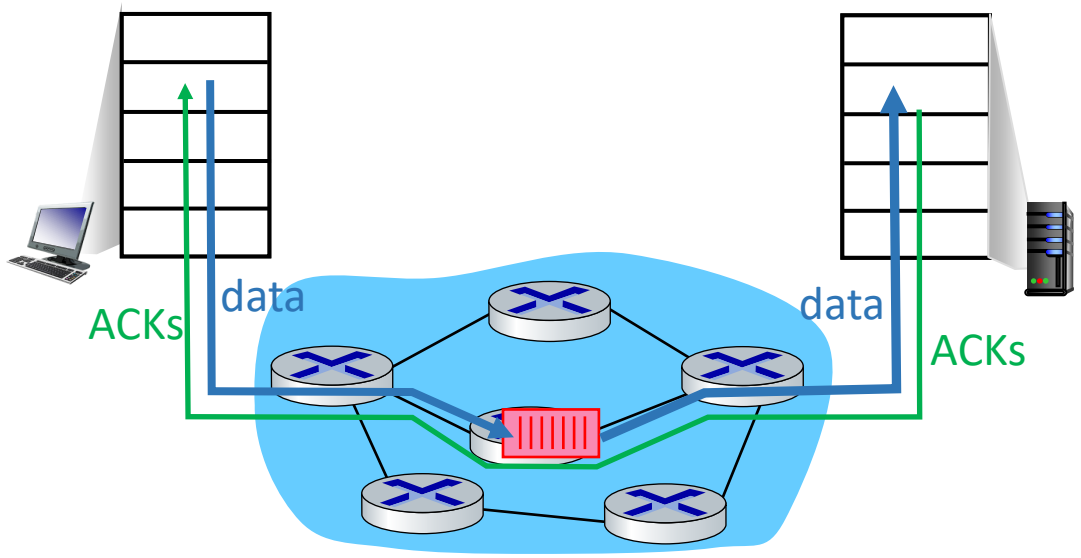
flow control: one sender
too fast for one receiver

Approaches towards congestion control

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

End-end congestion control:

- no explicit feedback from network
- congestion *inferred* from observed loss, delay
  - approach taken by TCP
  - No help from inside the network



TCP congestion control: AIMD
(Additive Increase, Multiplicative Decrease)

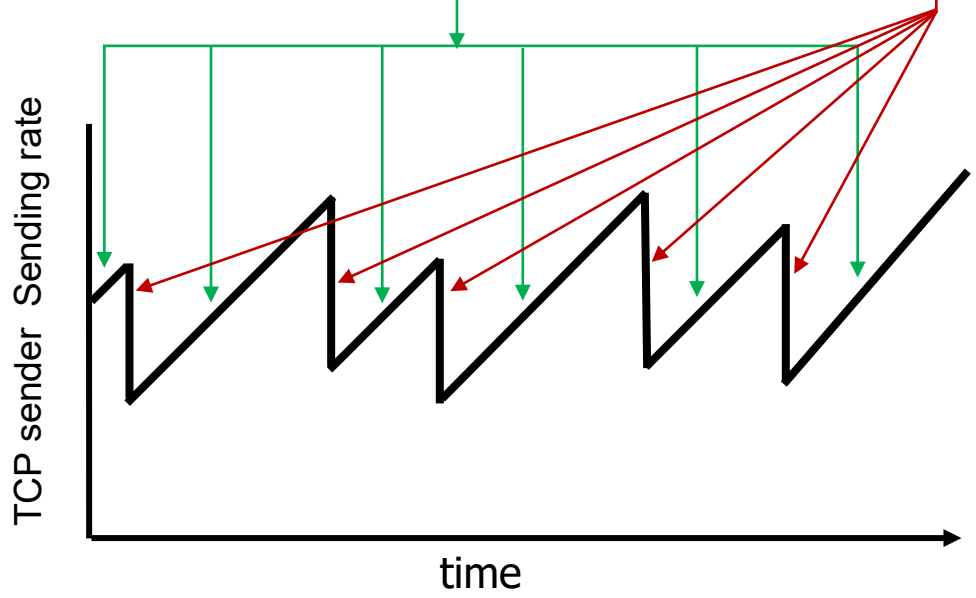
- *approach:* senders can increase sending rate until packet loss (congestion) occurs, then decrease sending rate on loss event

Additive Increase

increase sending rate by 1 maximum segment size every RTT until loss detected

Multiplicative Decrease

cut sending rate in half at each loss event

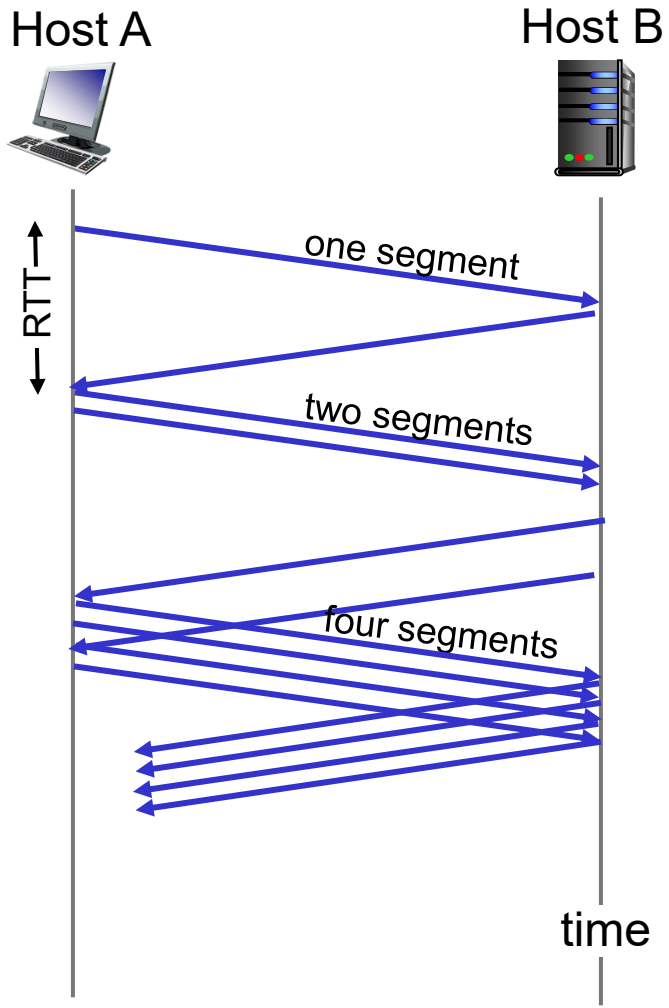


AIMD sawtooth behavior: *probing* for bandwidth

TCP slow start

- when connection begins, increase rate exponentially until first loss event:
  - initially **`cwnd`** = 1 MSS
  - double **`cwnd`** every RTT
  - done by incrementing **`cwnd`** for every ACK received
  - cwnd stands for Congestion Window and MSS = Maximum Segment Size → size of one data packet)

summary: initial rate is slow, but ramps up exponentially fast



TCP: from slow start to congestion avoidance

Q: when should the exponential increase switch to linear?

A: when **`cwnd`** gets to 1/2 of its value before timeout.

Implementation:

- variable **`ssthresh`**
- on loss event, **`ssthresh`** is set to 1/2 of **`cwnd`** just before loss event