Chapter 3
Transport Layer

A note on the use of these PowerPoint slides:

We're making these slides freely available to all (faculty, students, readers). They're in PowerPoint form so you see the animations; and can add, modify, and delete slides (including this one) and slide content to suit your needs. They obviously represent a *lot* of work on our part. In return for use, we only ask the following:

▪ If you use these slides (e.g., in a class) that you mention their source (after all, we'd like people to use our book!)
▪ If you post any slides on a www site, that you note that they are adapted from (or perhaps identical to) our slides, and note our copyright of this material.

For a revision history, see the slide note for this page.

Thanks and enjoy! JFK/KWR
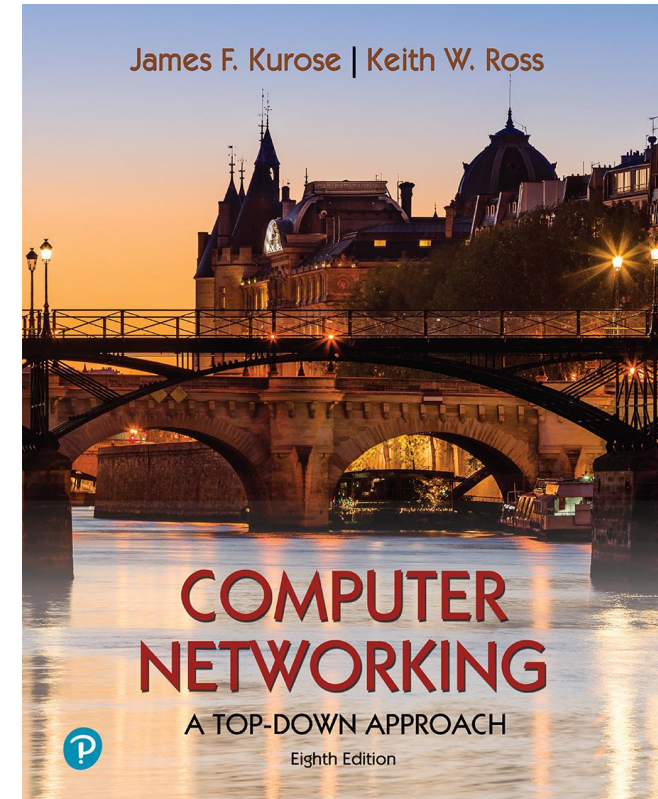
Computer Networking: A Top-Down Approach
8th edition
Jim Kurose, Keith Ross
Pearson, 2020

Internet protocol stack

- *application:* supporting network applications
  - HTTP, SMTP, IMAP
- *transport:* process-process data transfer
  - TCP, UDP
- *network:* routing of datagrams from source to destination
  - IP, routing protocols
- *link:* data transfer between neighboring network elements
  - Ethernet, 802.11 (WiFi), PPP
- *physical:* bits "on the wire"

application
transport
network
link
physical
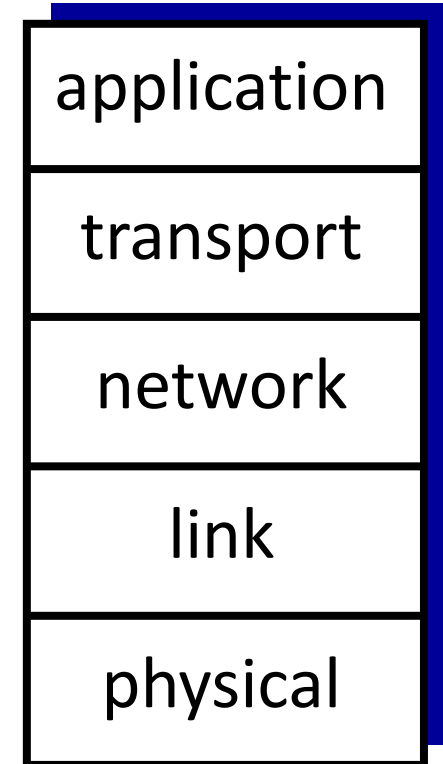
Transport layer: overview

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, best-effort service
  - TCP: reliable, flow- and congestion-controlled connection-oriented transport
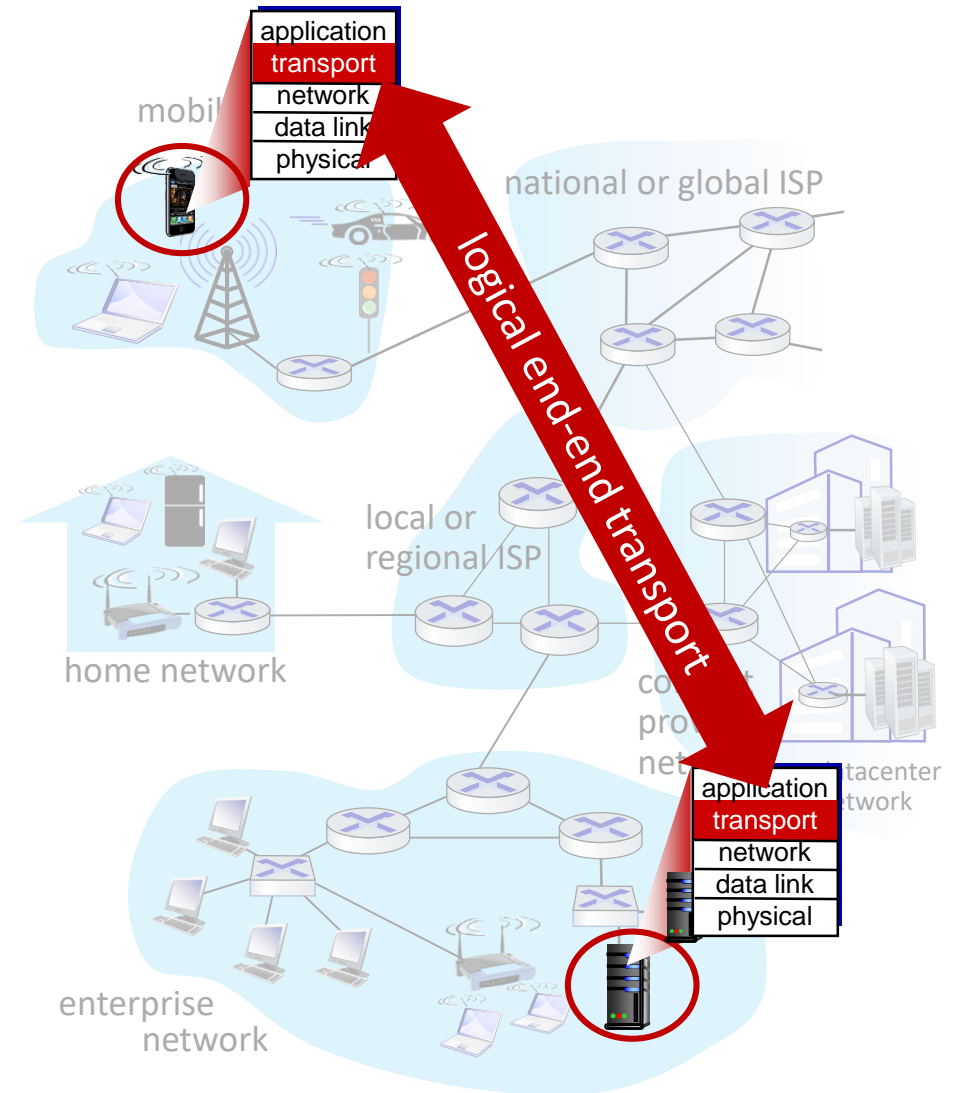
Transport layer: roadmap

- Transport-layer services
- Multiplexing and demultiplexing
- Connectionless transport: UDP
- Principles of reliable data transfer
- Connection-oriented transport: TCP
- Principles of congestion control
- TCP congestion control
- Evolution of transport-layer functionality

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

THERE was an old woman who lived in a shoe,
She had so many children, she didn't know what to do.
She gave them some milk and nice butter bread,
She kissed them all round and put them to bed.

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

- network layer: logical communication between *hosts*

Transport Layer Actions

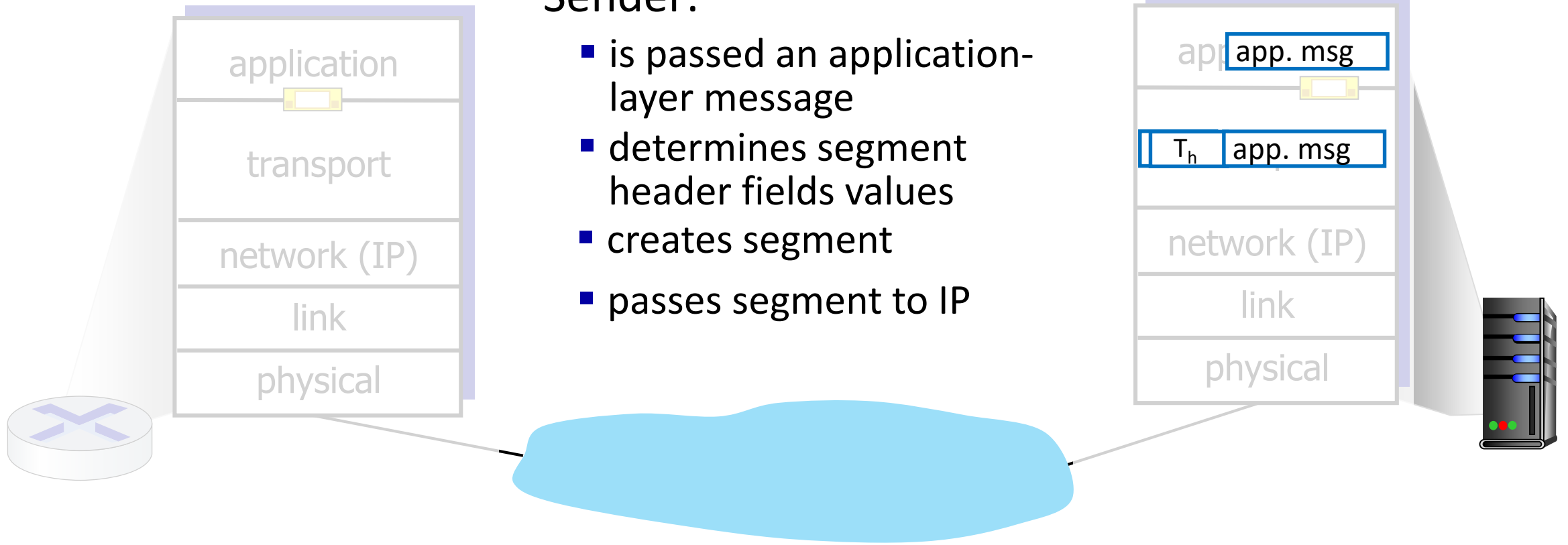**Sender:**

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

application

transport

network (IP)

link

physical

app

app. msg

T_h app. msg

network (IP)

link

physical

Transport Layer Actions

application

app. msg

network (IP)

link

physical

T_h app. msg

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

application

transport

network (IP)

link

physical
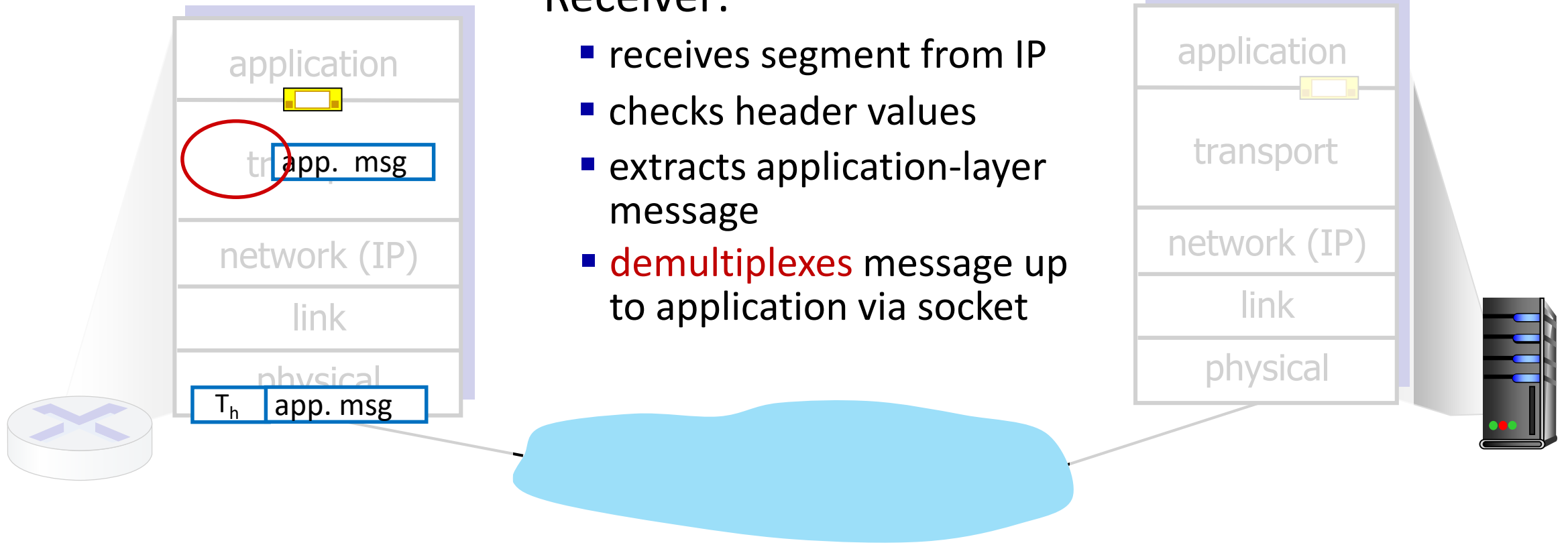
Two principal Internet transport protocols

- **TCP:** Transmission Control Protocol
  - reliable, in-order delivery
  - congestion control
  - flow control
  - connection setup

- **UDP:** User Datagram Protocol
  - unreliable, unordered delivery
  - no-frills extension of "best-effort" IP

- services *not* available:
  - delay guarantees
  - bandwidth guarantees
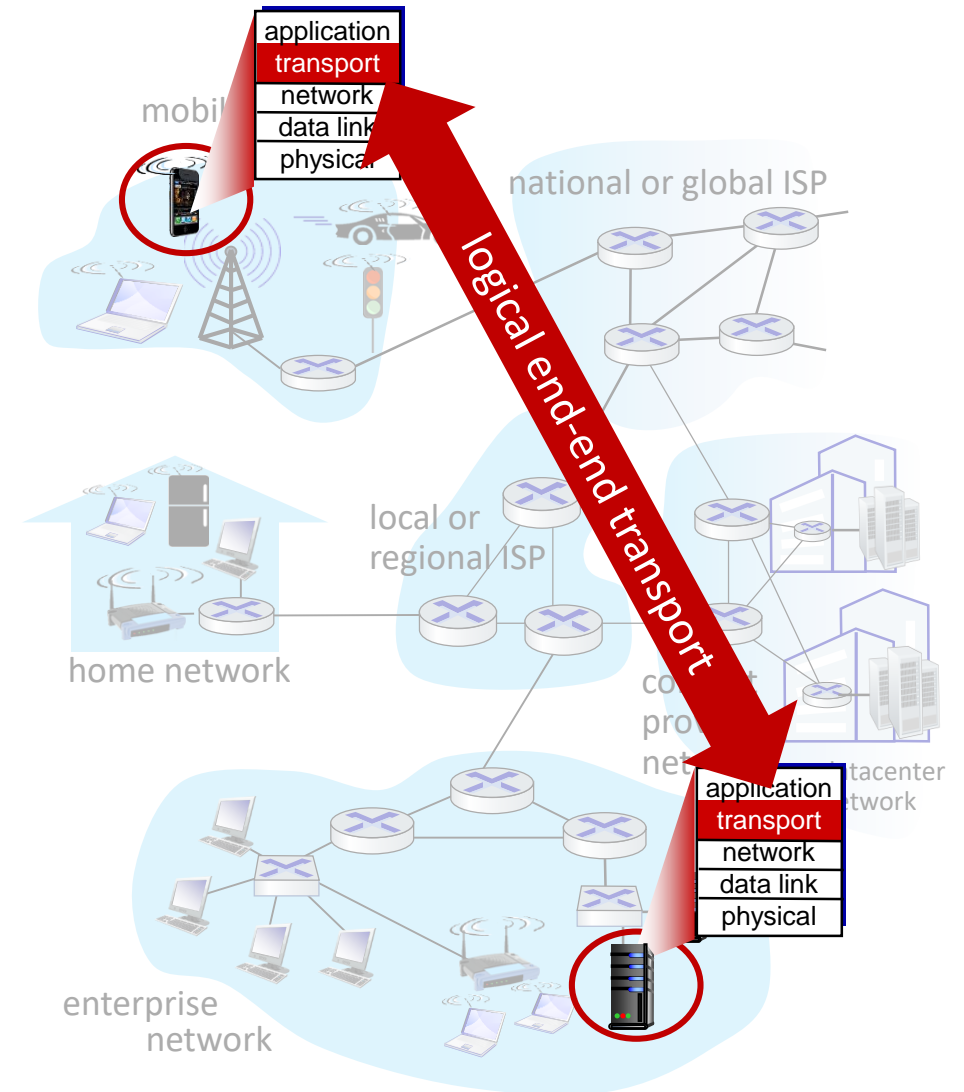
application
transport
network
data link
physical

mobile

national or global ISP

logical end-end transport

local or
regional ISP

home network

datacenter
network

application
transport
network
data link
physical

enterprise
network

Chapter 3: roadmap

- Transport-layer services
- **Multiplexing and demultiplexing**
- Connectionless transport: UDP
- Principles of reliable data transfer
- Connection-oriented transport: TCP
- Principles of congestion control
- TCP congestion control
- Evolution of transport-layer functionality

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

HTTP server

client

application
transport
$H_n H_t$ HTTP msg
link
physical

HTTP msg
H_t HTTP msg
$H_n H_t$ HTTP msg
link
physical

application
transport
network
link
physical

$H_n H_t$ HTTP msg
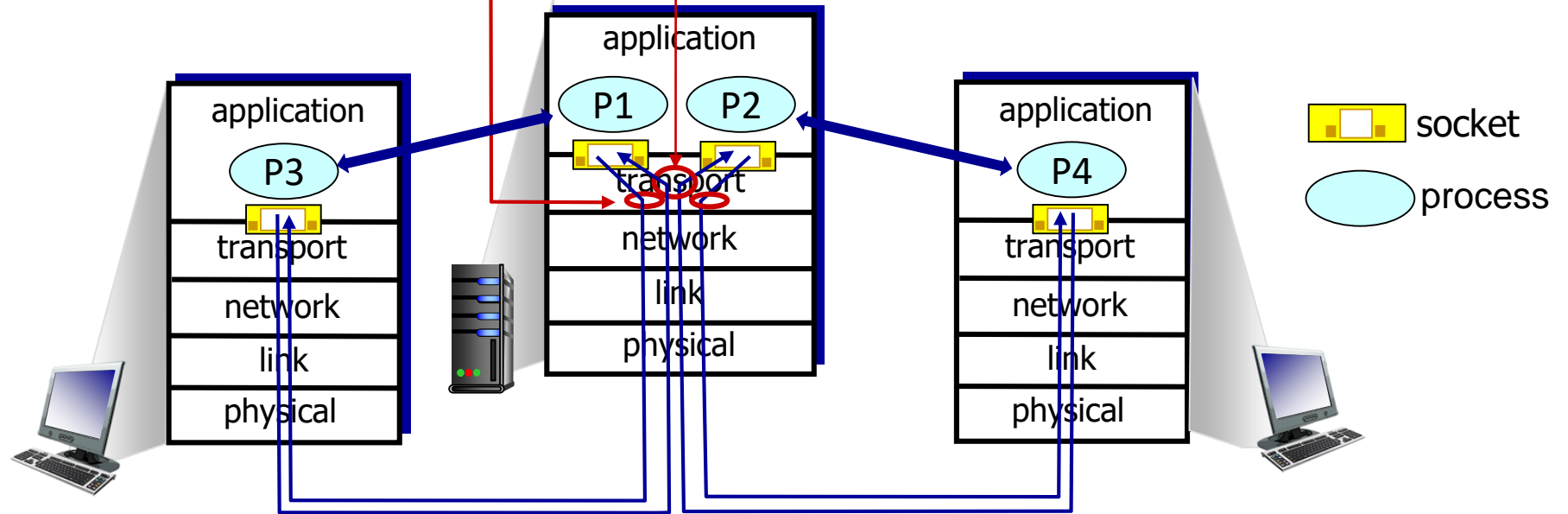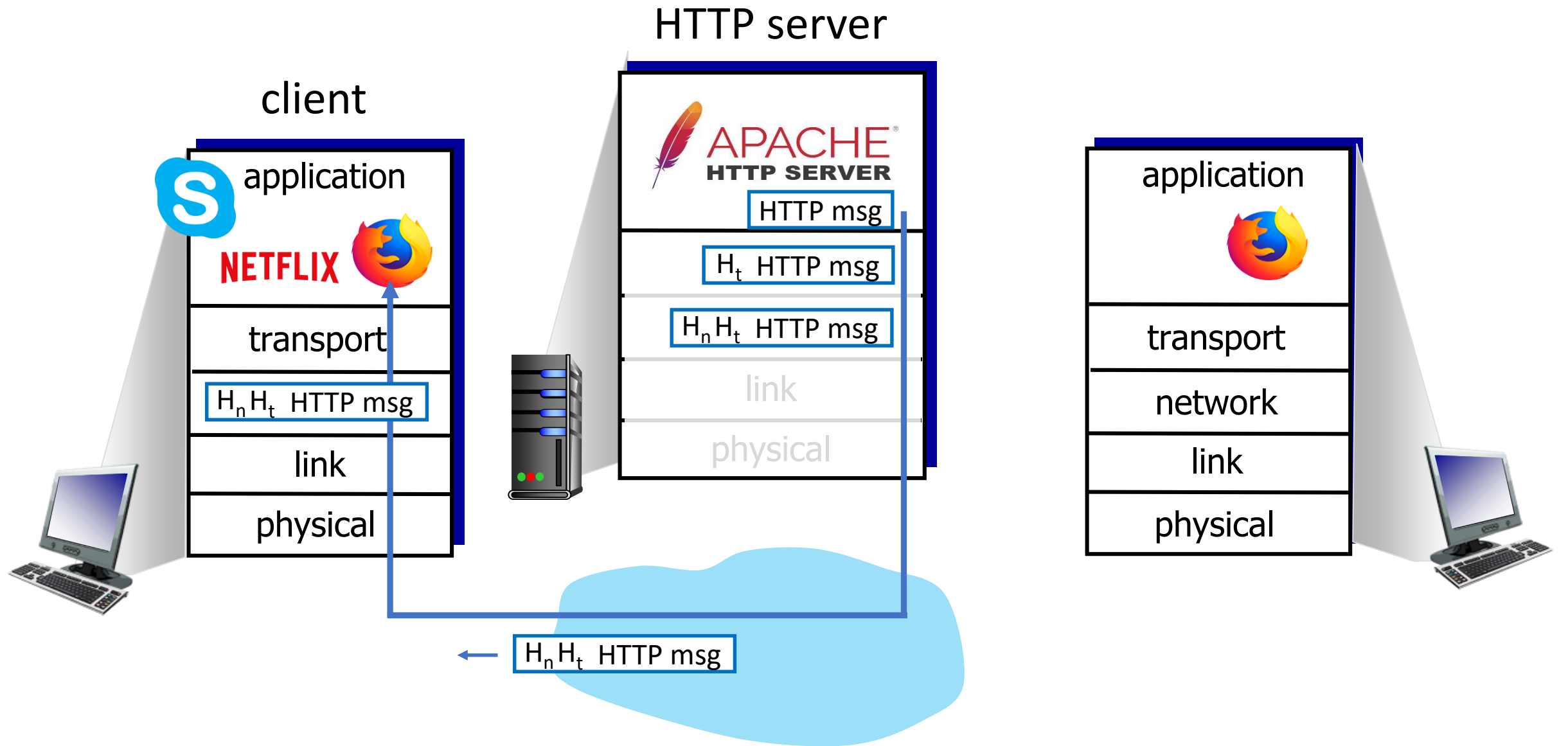
Q: *how did transport layer know to deliver message to Firefox browser process rather than Netflix process or Skype process?*
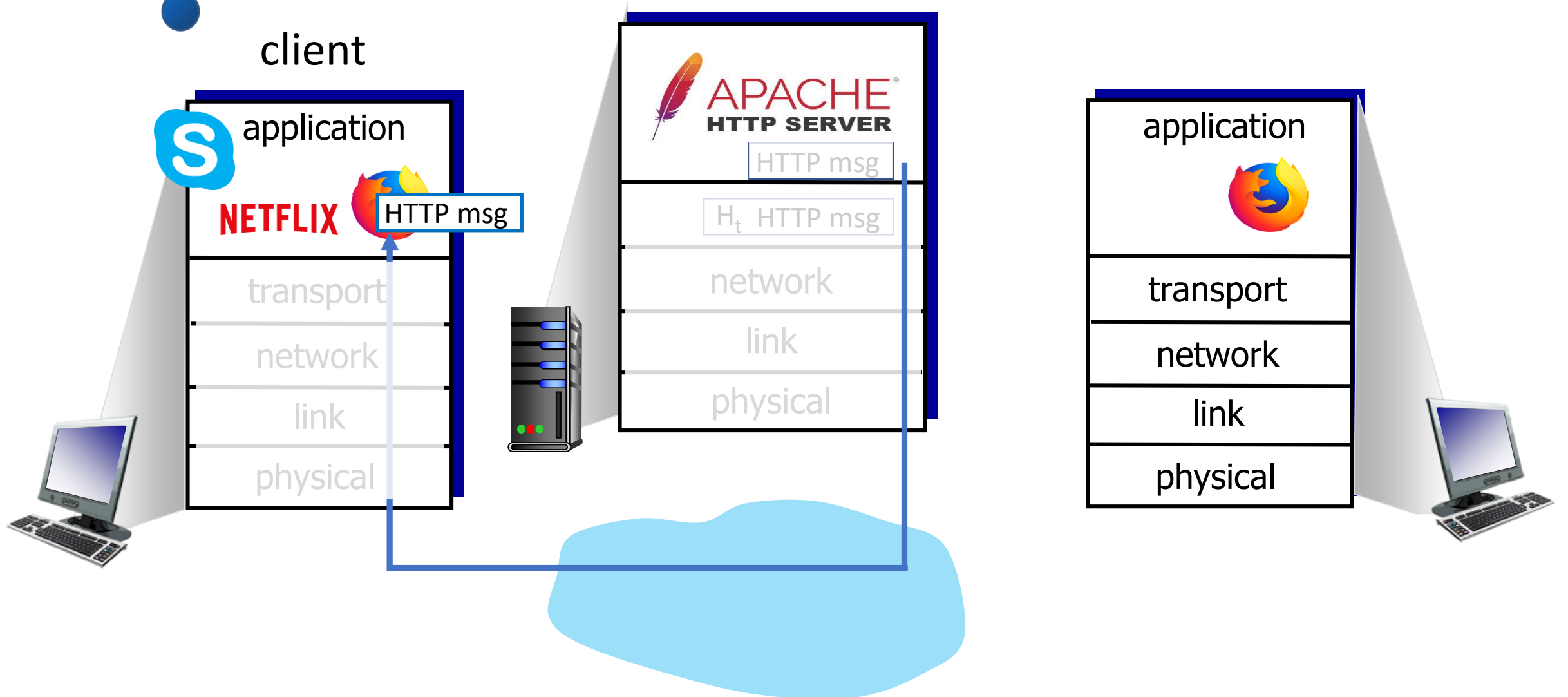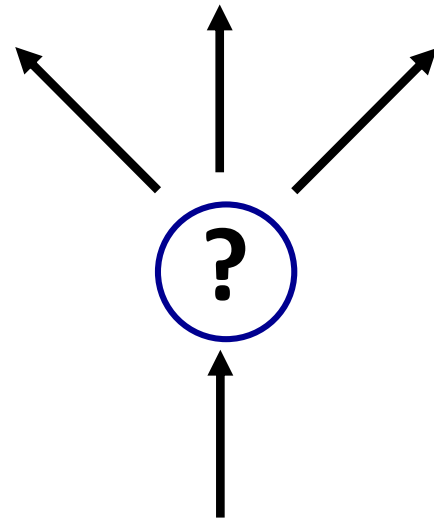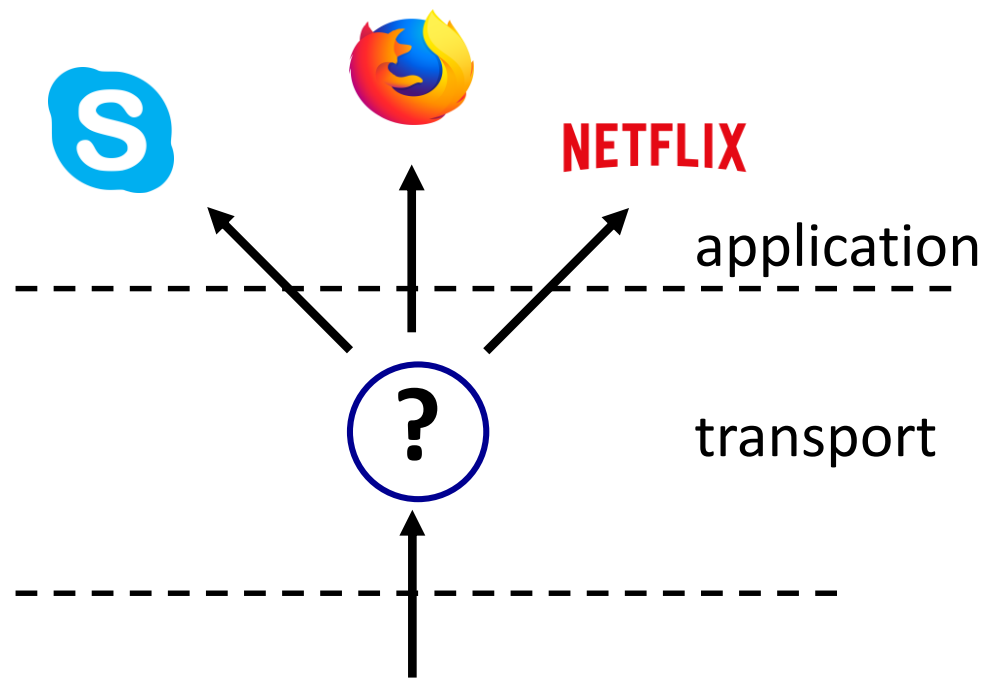
client

application

NETFLIX

HTTP msg

transport

network

link

physical

APACHE® HTTP SERVER

HTTP msg

H_t HTTP msg

network

link

physical

application

transport

network

link

physical

de-multiplexing

application

transport

de-multiplexing

Demultiplexing

multiplexing

application
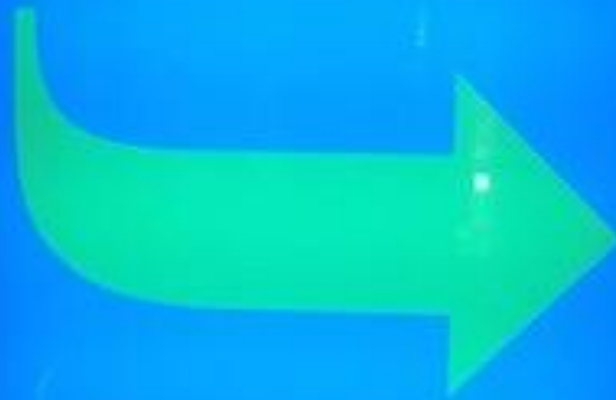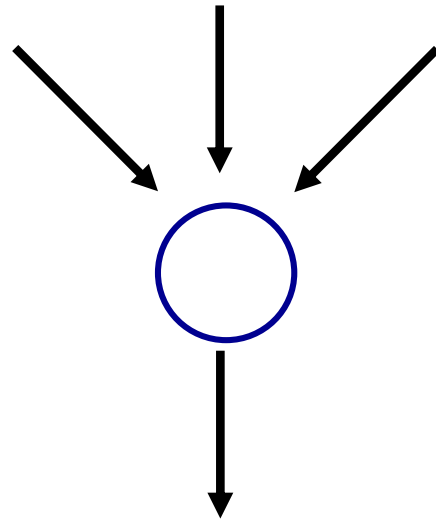
transport
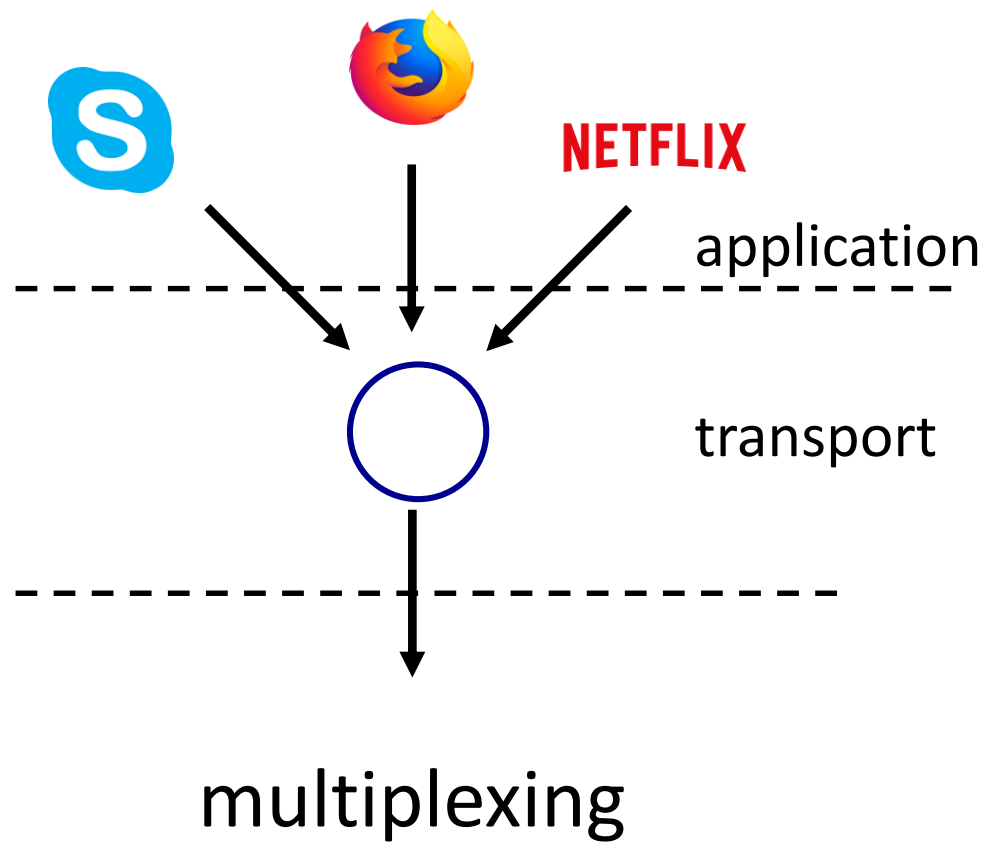
multiplexing

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

32 bits

source port #	dest port #
other header fields	
application data (payload)	

TCP/UDP segment format
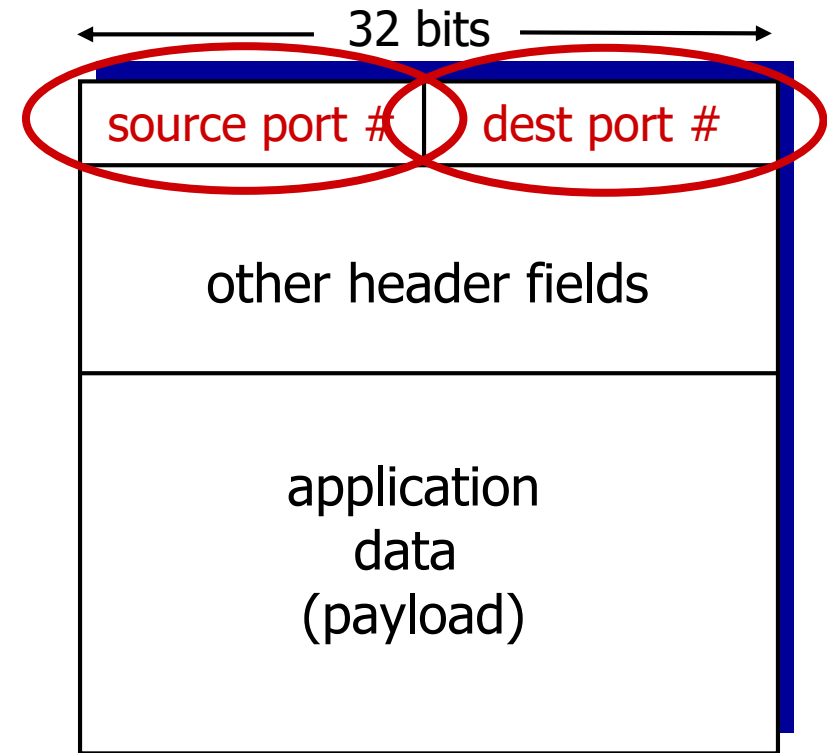
Connectionless demultiplexing

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

```
DatagramSocket serverSocket =
    new DatagramSocket(6428);
```

```
DatagramSocket mySocket2 =
    new DatagramSocket(9157);
```

```
DatagramSocket mySocket1 =
    new DatagramSocket(5775);
```

application

P3

transport

network

link

physical

application

P1

transport

network

link

physical

application

P4

transport

network

link

physical

B
source port: 6428
dest port: 9157

D
source port: ?
dest port: ?

A
source port: 9157
dest port: 6428

C
source port: ?
dest port: ?
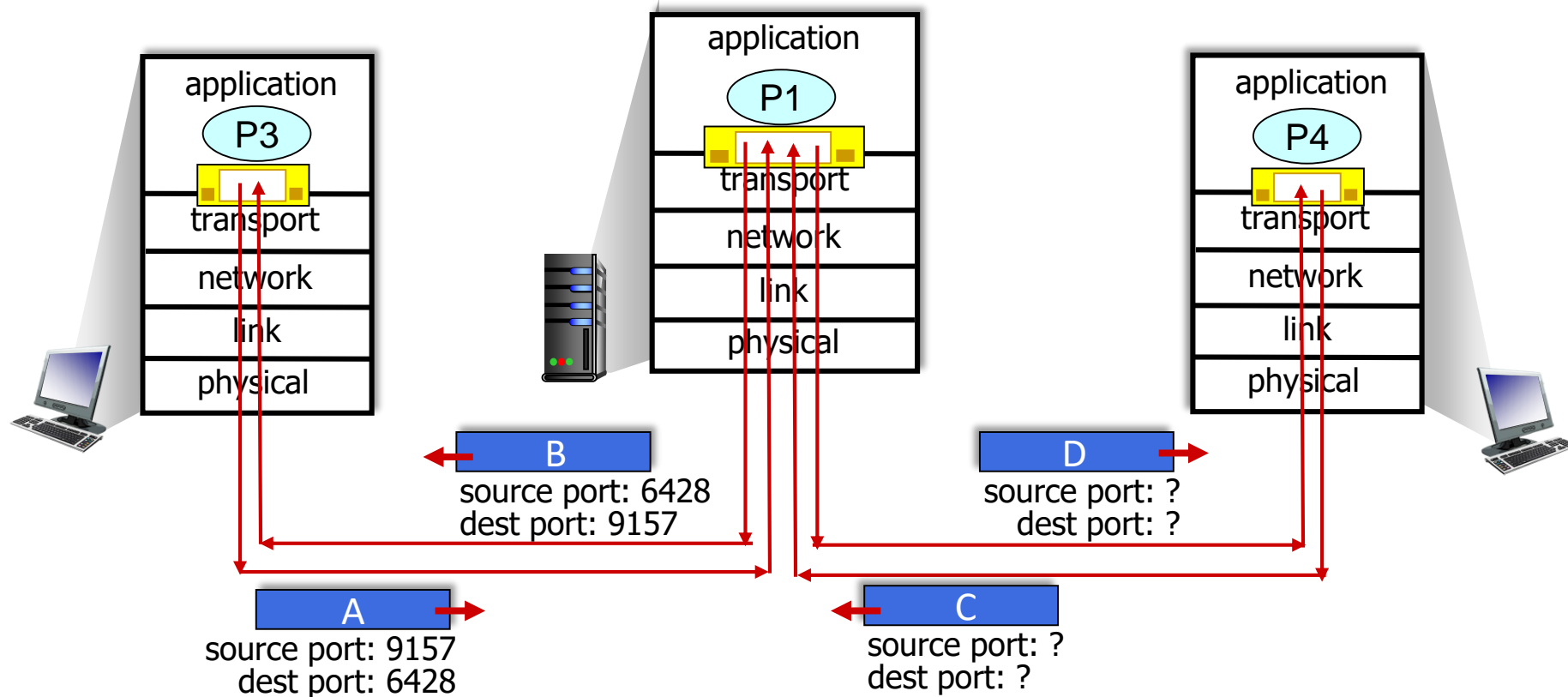
Connection-oriented demultiplexing

- 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

host: IP
address A

server: IP
address B

host: IP
address C

source IP,port: B,80
dest IP,port: A,9157

source IP,port: A,9157
dest IP,port: B,80

source IP,port: C,5775
dest IP,port: B,80

source IP,port: C,9157
dest IP,port: B,80
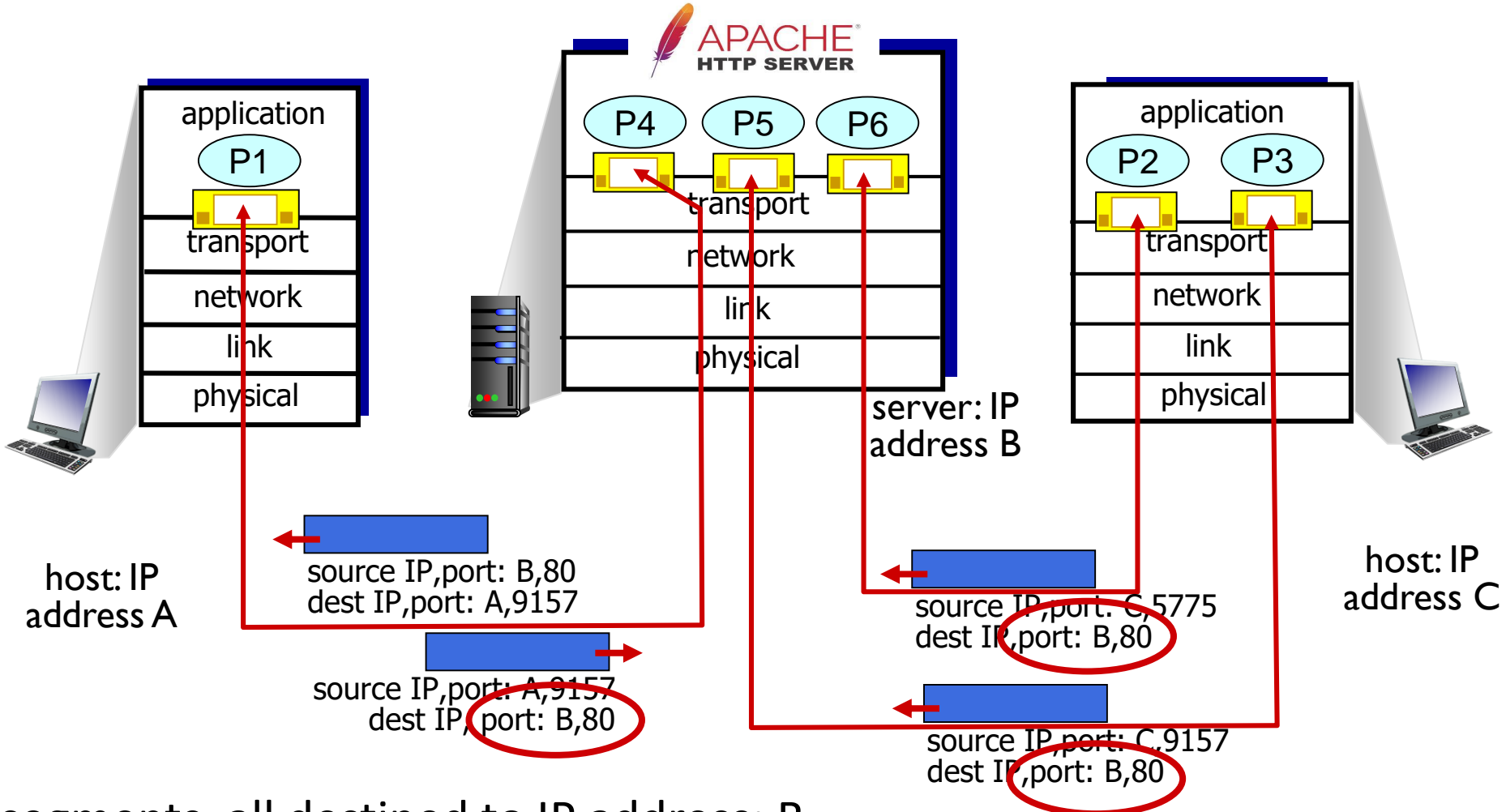
Three segments, all destined to IP address: B,
dest port: 80 are demultiplexed to *different* sockets

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

Chapter 3: roadmap

- Transport-layer services

- Multiplexing and demultiplexing

- **Connectionless transport: UDP**

- Principles of reliable data transfer

- Connection-oriented transport: TCP

- Principles of congestion control

- TCP congestion control

- Evolution of transport-layer functionality

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: User Datagram Protocol [RFC 768]

INTERNET STANDARD

RFC 768   J. Postel
                                                                      ISI
                                                          28 August 1980

```
                       User Datagram Protocol
                       ----------------------

Introduction
------------

This User Datagram  Protocol  (UDP)  is  defined  to  make  available  a
datagram   mode  of  packet-switched   computer   communication  in  the
environment  of  an  interconnected  set  of  computer  networks.   This
protocol   assumes   that  the  Internet  Protocol  (IP)  [1] is used as the
underlying protocol.

This protocol  provides  a procedure  for application  programs  to send
messages  to other programs  with a minimum  of protocol mechanism.  The
protocol  is transaction oriented, and delivery and duplicate protection
are not guaranteed.  Applications requiring ordered reliable delivery of
streams of data should use the Transmission Control Protocol (TCP) [2].

Format
------

          0      7 8     15 16    23 24    31
         +--------+--------+--------+--------+
         |     Source      |   Destination   |
         |     Port        |      Port       |
         +--------+--------+--------+--------+
         |                 |                 |
         |     Length      |    Checksum     |
         +--------+--------+--------+--------+
         |
         |          data octets ...
         +---------------- ...
```

UDP: Transport Layer Actions

SNMP client

application
transport (UDP)
network (IP)
link
physical

SNMP server

application
transport (UDP)
network (IP)
link
physical

UDP: Transport Layer Actions

SNMP client

- application
- transport (UDP)
- network (IP)
- link
- physical

SNMP server

- application | SNMP msg
- transport (UDP) | UDP_h | SNMP msg
- network (IP)
- link
- physical

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

SNMP client

application

transport
(UDP)

SNMP msg

UDP_h | SNMP msg

link

physical

SNMP server

application

transport
(UDP)

network (IP)

link

physical

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

length, in bytes of UDP segment, including header

data to/from application layer

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

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

sender:

- treat contents of UDP segment (including UDP header fields and IP addresses) as sequence of 16-bit integers
- checksum: addition (one's complement sum) of segment content
- checksum value put into UDP checksum field

receiver:

- compute checksum of received segment
- check if computed checksum equals checksum field value:
  - not equal - error detected
  - equal - no error detected. *But maybe errors nonetheless?* More later ….

Internet checksum: an example

example: add two 16-bit integers

```
1 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0
1 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1
```
wraparound   (1) 1 0 1 1 1 0 1 1 1 0 1 1 1 0 1 1

sum   1 0 1 1 1 0 1 1 1 0 1 1 1 1 0 0

checksum   0 1 0 0 0 1 0 0 0 1 0 0 0 0 1 1

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

```
          1 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0          0 1
          1 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1          1 0
         ────────────────────────────────
wraparound  1 1 0 1 1 1 0 1 1 1 0 1 1 1 0 1 1
         ────────────────────────────────
sum         1 0 1 1 1 0 1 1 1 0 1 1 1 1 0 0
checksum    0 1 0 0 0 1 0 0 0 1 0 0 0 0 1 1
```

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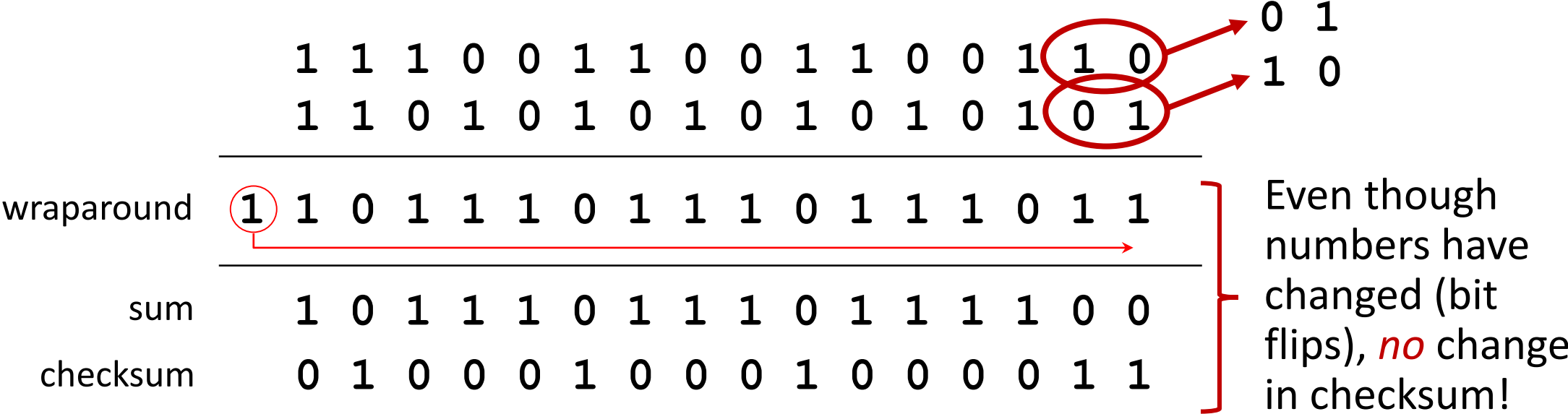
Summary: UDP

- "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)

Chapter 3: roadmap

- Transport-layer services
- Multiplexing and demultiplexing
- Connectionless transport: UDP
- **Principles of reliable data transfer**
- Connection-oriented transport: TCP
- Principles of congestion control
- TCP congestion control
- Evolution of transport-layer functionality

Principles of reliable data transfer

reliable service *abstraction*

Principles of reliable data transfer

reliable service *abstraction*

sending process

application
transport

data

reliable channel

receiving process

data

sending process

application
transport

data

sender-side of reliable data transfer protocol

transport
network

receiving process

data

receiver-side of reliable data transfer protocol

unreliable channel

reliable service *implementation*
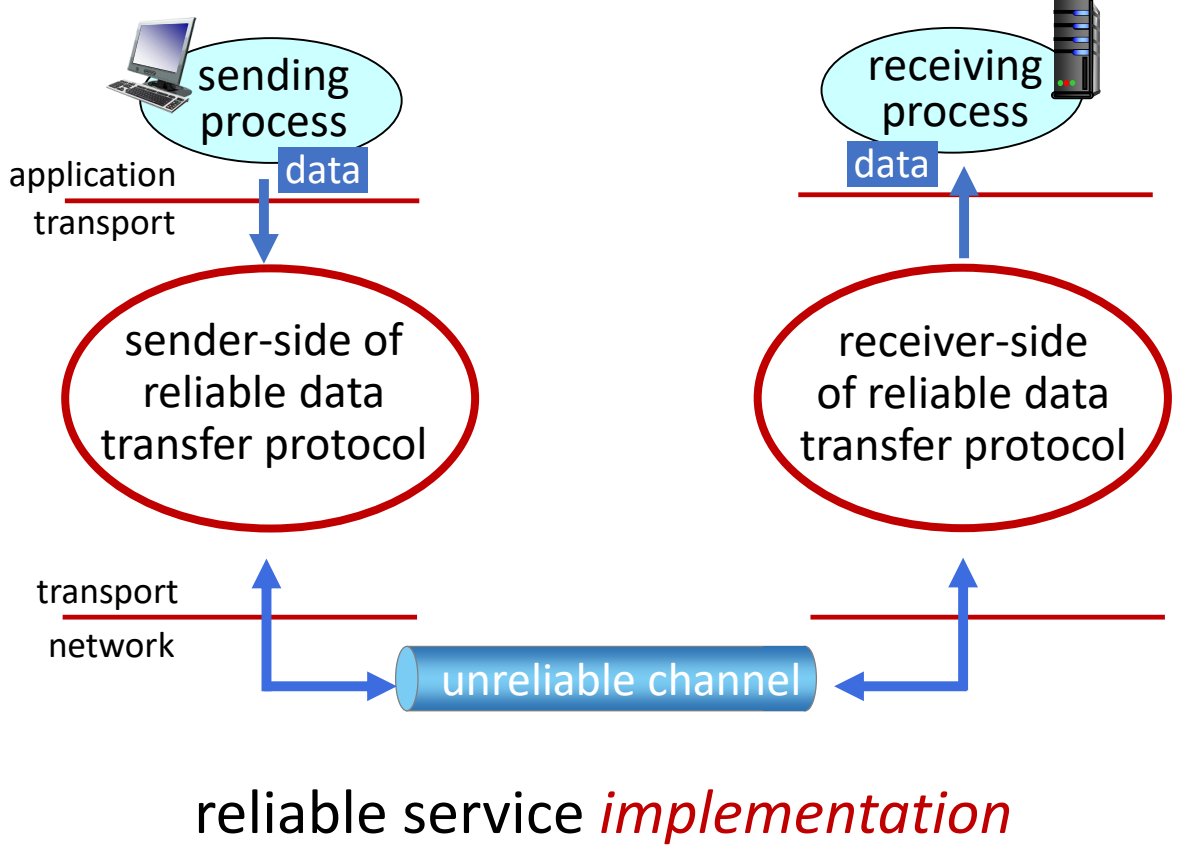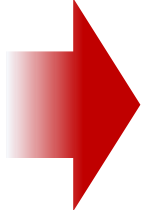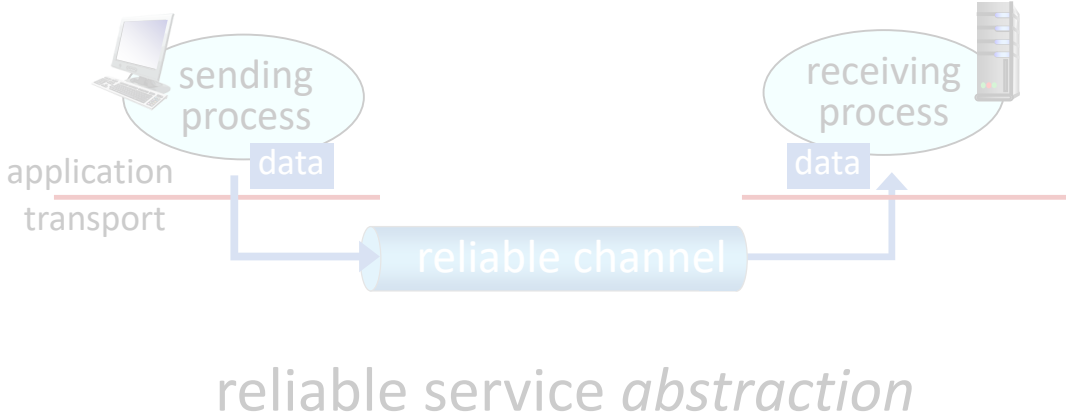
Principles of reliable data transfer

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

reliable service *implementation*
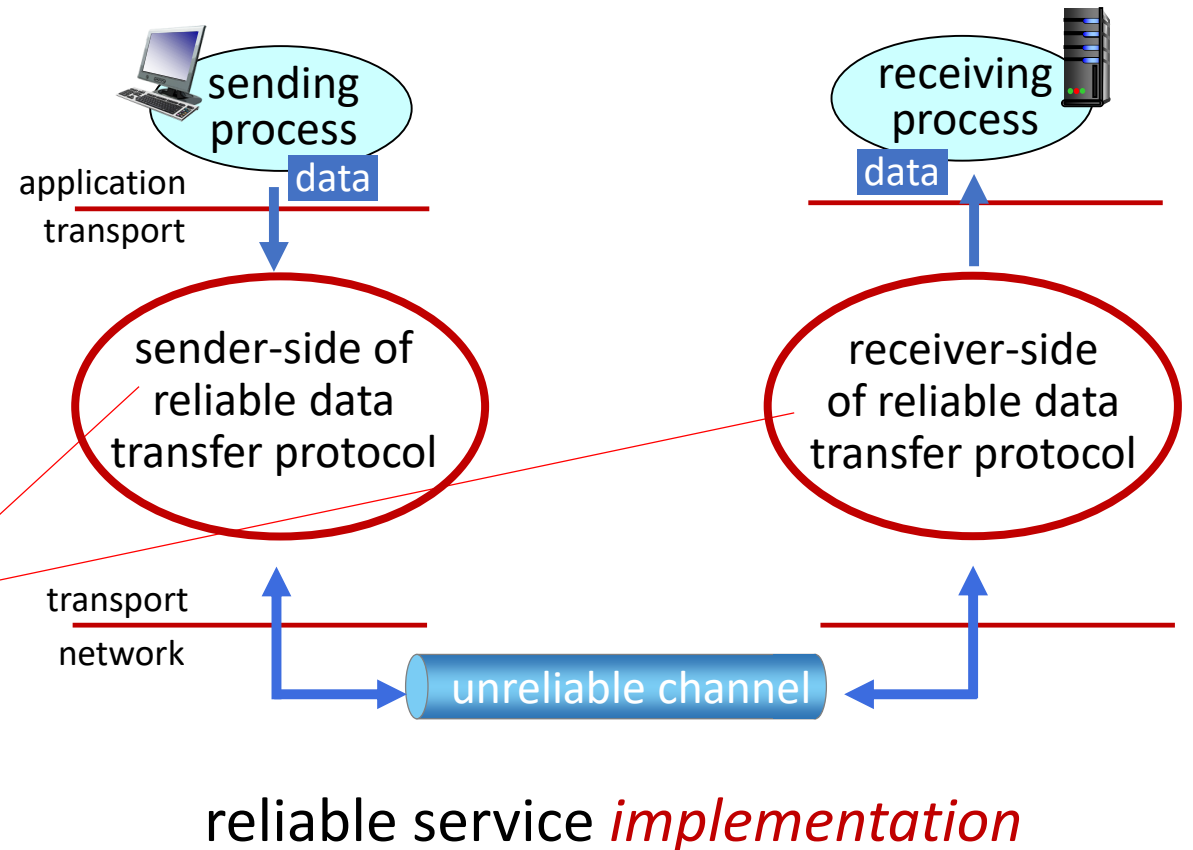
Principles of reliable data transfer

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

reliable service *implementation*
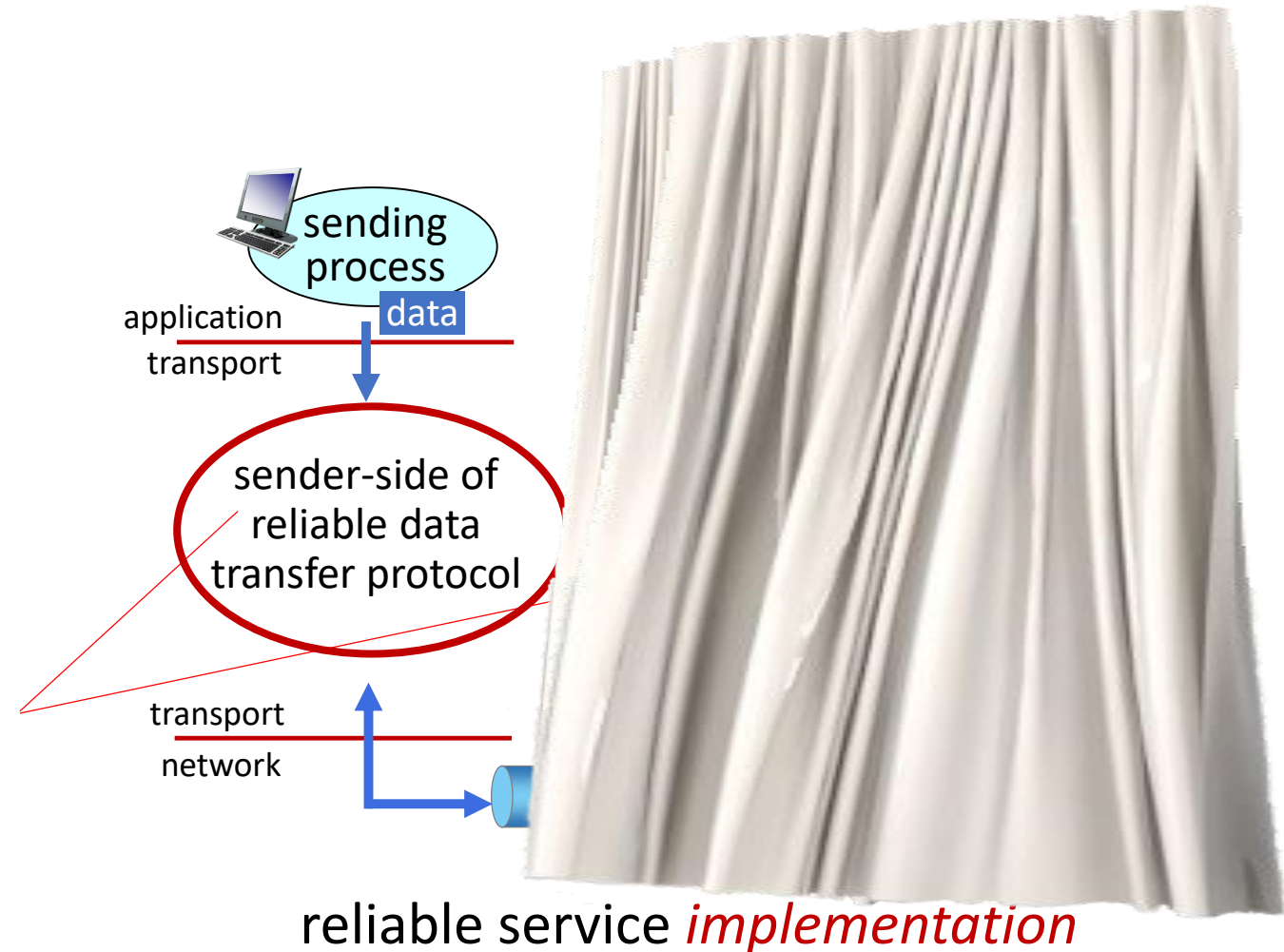
Reliable data transfer protocol (rdt): interfaces

rdt_send(): called from above, (e.g., by app.). Passed data to deliver to receiver upper layer

deliver_data(): called by `rdt` to deliver data to upper layer

sending process

receiving process

`rdt_send()`

data

data

`deliver_data()`

data

sender-side *implementation* of `rdt` reliable data transfer protocol

receiver-side *implementation* of `rdt` reliable data transfer protocol

packet

`udt_send()`

Header data

Header data

`rdt_rcv()`

unreliable channel

udt_send(): called by `rdt` to transfer packet over unreliable channel to receiver

Bi-directional communication over unreliable channel

rdt_rcv(): called when packet arrives on receiver side of channel

Reliable data transfer: getting started

We will:

- incrementally develop sender, receiver sides of reliable data transfer protocol (`rdt`)

- consider only unidirectional data transfer
  - but control info will flow in both directions!

- use finite state machines (FSM) to specify sender, receiver

state: when in this "state" next state uniquely determined by next event

$$\frac{\text{event causing state transition}}{\text{actions taken on state transition}}$$

state 1

$$\frac{\text{event}}{\text{actions}}$$

state 2
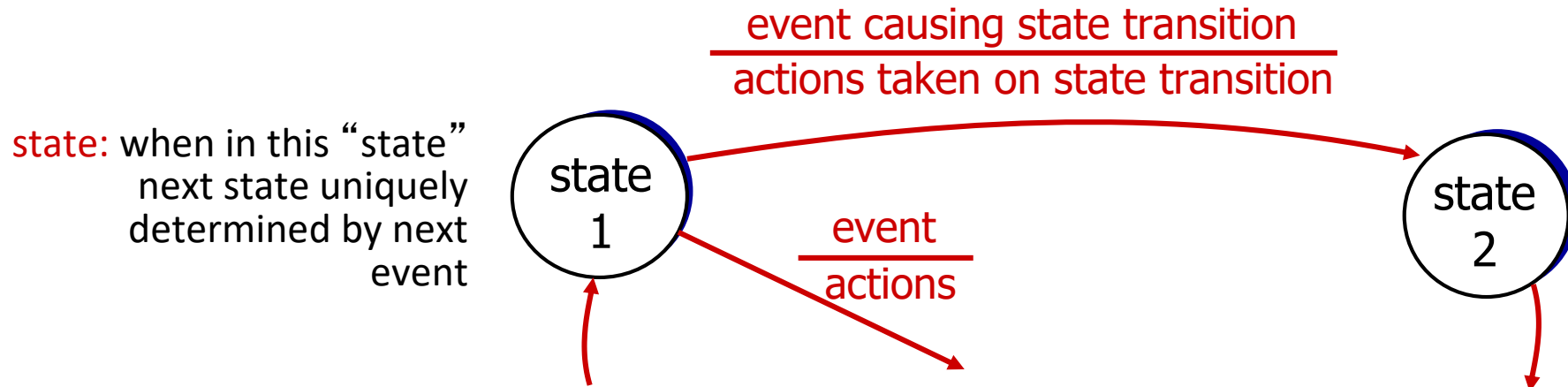
rdt1.0: reliable transfer over a reliable channel

- underlying channel perfectly reliable
  - no bit errors
  - no loss of packets (no packets are duplicated or reordered)

- *separate* FSMs for sender, receiver:
  - sender sends data into underlying channel
  - receiver reads data from underlying channel

sender

```
Wait for          rdt_send(data)
call from    ──────────────────────
above        packet = make_pkt(data)
             udt_send(packet)
```

receiver

```
Wait for          rdt_rcv(packet)
call from    ──────────────────────
below        extract (packet,data)
             deliver_data(data)
```
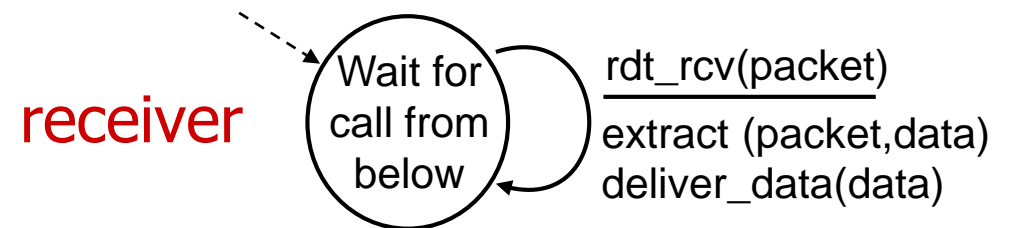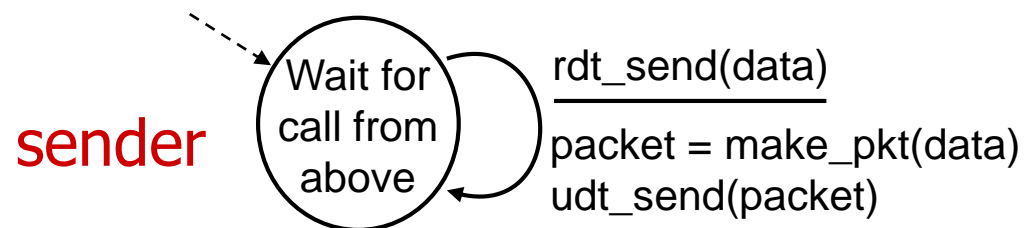
rdt2.0: channel with bit errors

- 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?

rdt2.0: 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

rdt2.0: FSM specifications

sender

rdt_send(data)
─────────────
sndpkt = make_pkt(data, checksum)
udt_send(sndpkt)

Wait for call from above

Wait for ACK or NAK

rdt_rcv(rcvpkt) && isNAK(rcvpkt)
─────────────
udt_send(sndpkt)

rdt_rcv(rcvpkt) && isACK(rcvpkt)
─────────────
Λ

rdt2.0: FSM specification

sender

rdt_send(data)
—————————————
sndpkt = make_pkt(data, checksum)
udt_send(sndpkt)

rdt_rcv(rcvpkt) &&
isNAK(rcvpkt)
—————————————
udt_send(sndpkt)

(Wait for call from above)

(Wait for ACK or NAK)

rdt_rcv(rcvpkt) && isACK(rcvpkt)
—————————————
Λ
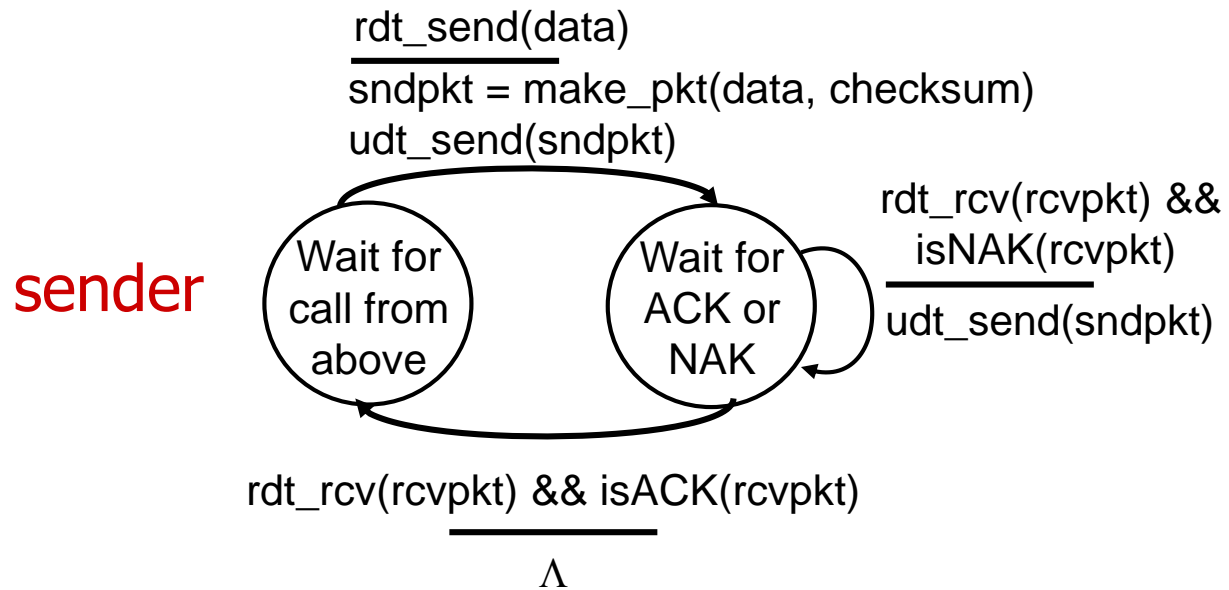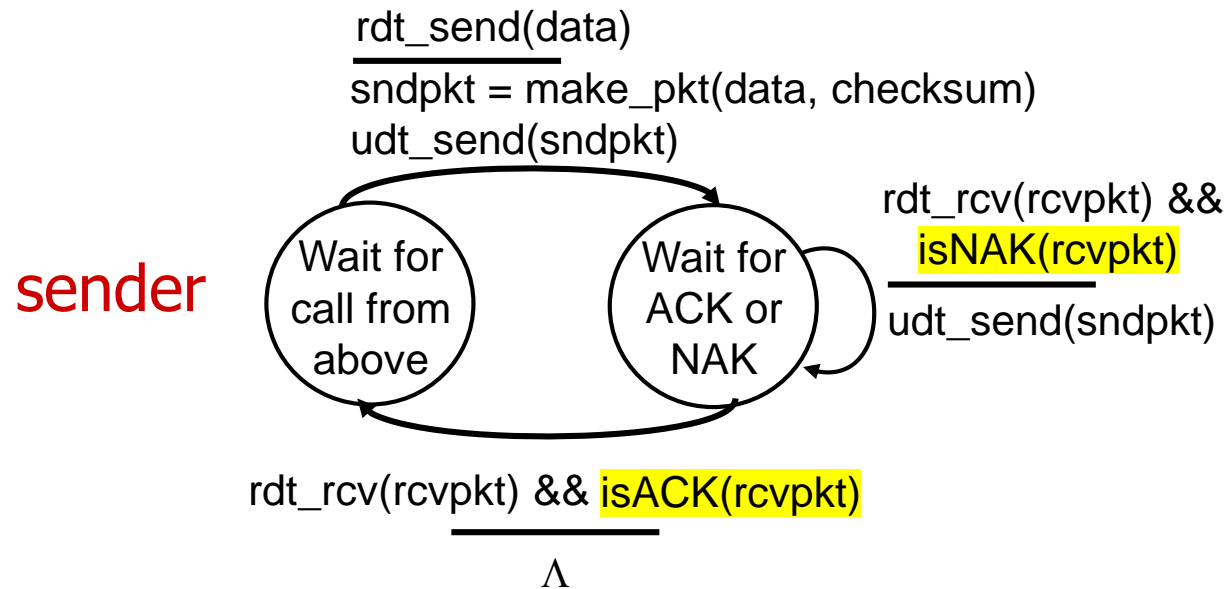
Note: "state" of receiver (did the receiver get my message correctly?) isn't known to sender unless somehow communicated from receiver to sender

- that's why we need a protocol!

rdt2.0: operation with no errors

rdt_send(data)
$\overline{}$
sndpkt = make_pkt(data, checksum)
udt_send(sndpkt)

sender

Wait for
call from
above

Wait for
ACK or
NAK

rdt_rcv(rcvpkt) &&
isNAK(rcvpkt)
$\overline{}$
udt_send(sndpkt)

rdt_rcv(rcvpkt) && isACK(rcvpkt)
$\overline{}$
Λ

rdt_rcv(rcvpkt) && corrupt(rcvpkt)
$\overline{}$
udt_send(NAK)

Wait for
call from
below

receiver

rdt_rcv(rcvpkt) && notcorrupt(rcvpkt)
$\overline{}$
extract(rcvpkt,data)
deliver_data(data)
udt_send(ACK)
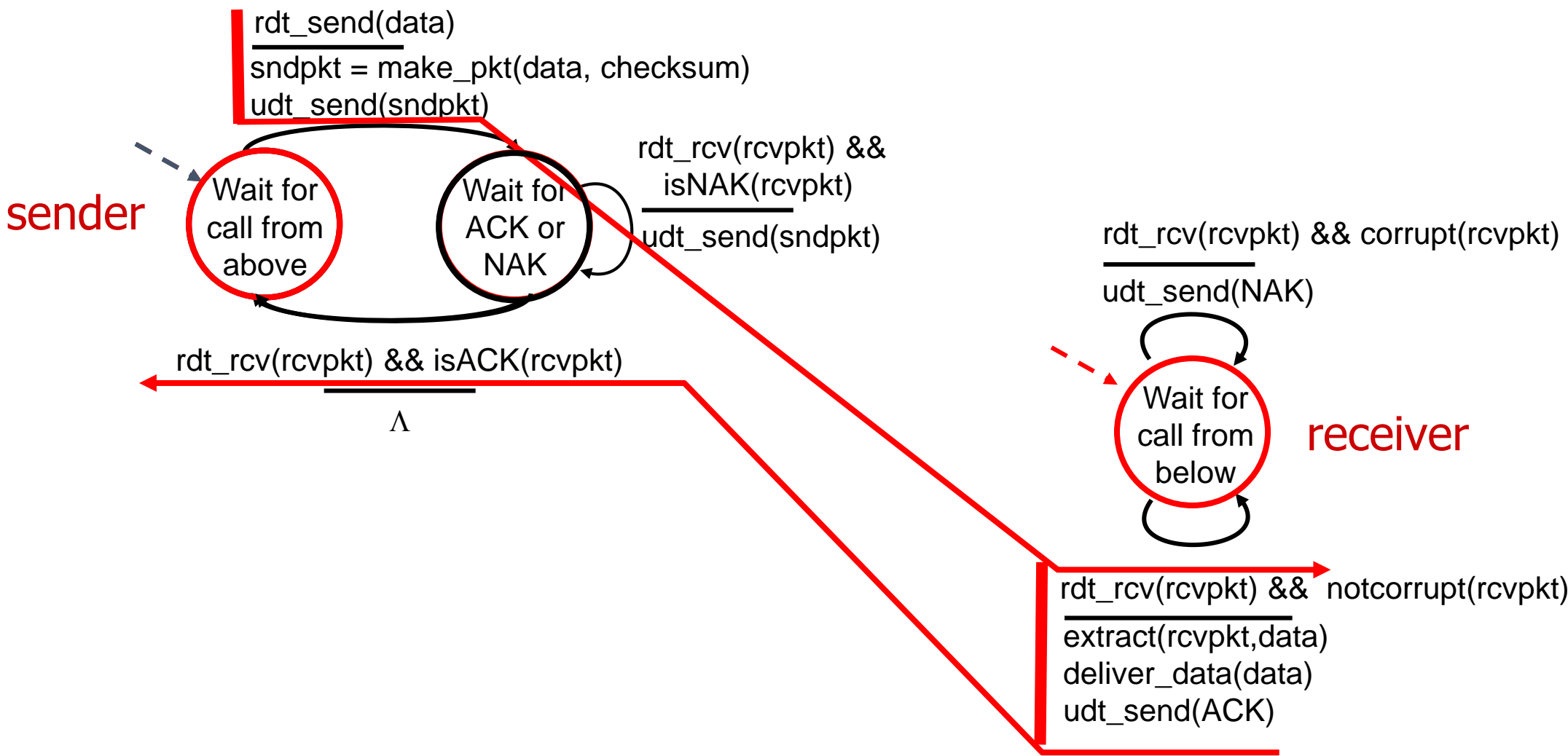
rdt2.0: corrupted packet scenario

sender

rdt_send(data)
_____
sndpkt = make_pkt(data, checksum)
udt_send(sndpkt)

Wait for call from above

Wait for ACK or NAK

rdt_rcv(rcvpkt) &&
isNAK(rcvpkt)
_____
udt_send(sndpkt)

rdt_rcv(rcvpkt) && isACK(rcvpkt)
_____
Λ

receiver

rdt_rcv(rcvpkt) && corrupt(rcvpkt)
_____
udt_send(NAK)

Wait for call from below

rdt_rcv(rcvpkt) && notcorrupt(rcvpkt)
_____
extract(rcvpkt,data)
deliver_data(data)
udt_send(ACK)
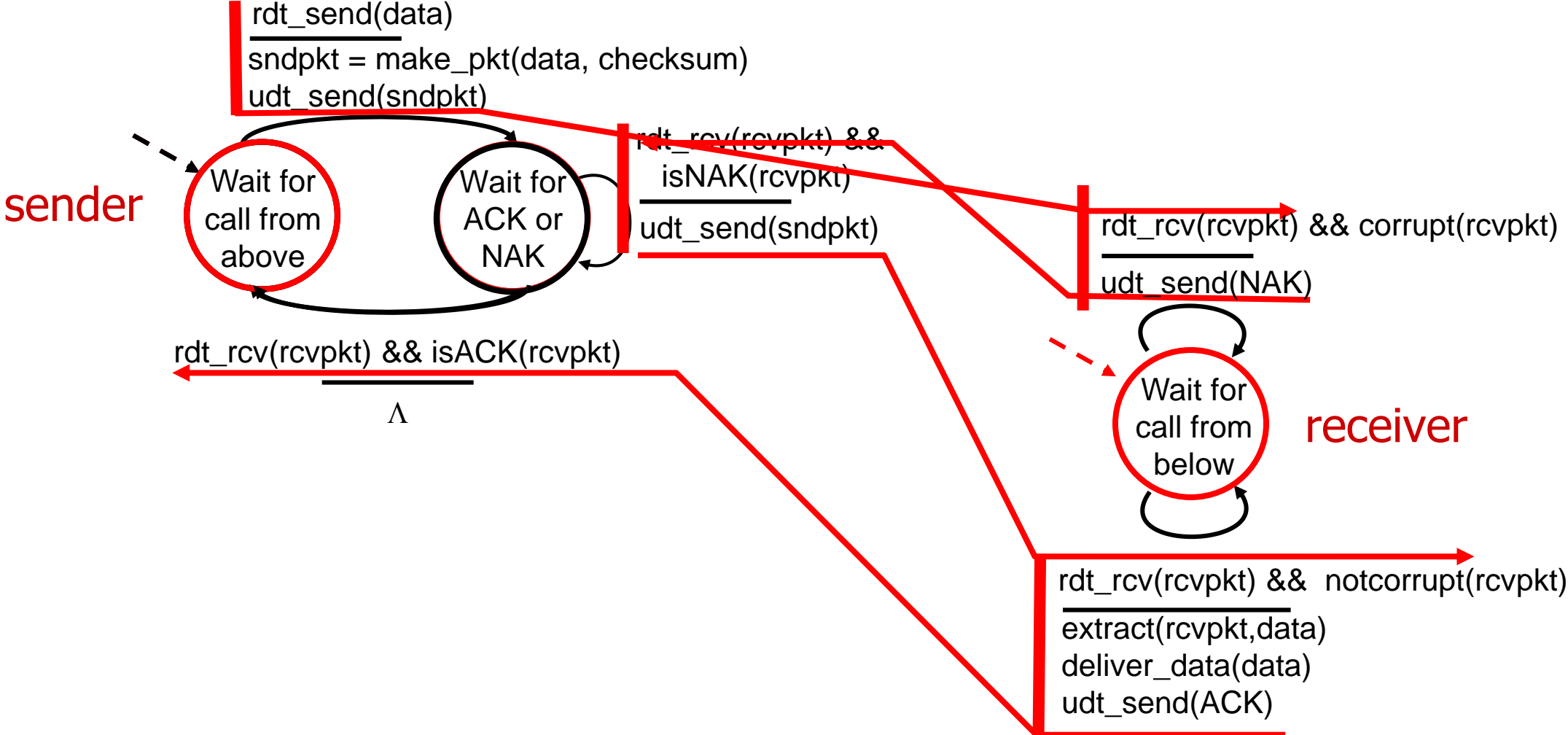
rdt2.0 has a fatal flaw!

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

rdt2.1: sender, handling garbled ACK/NAKs

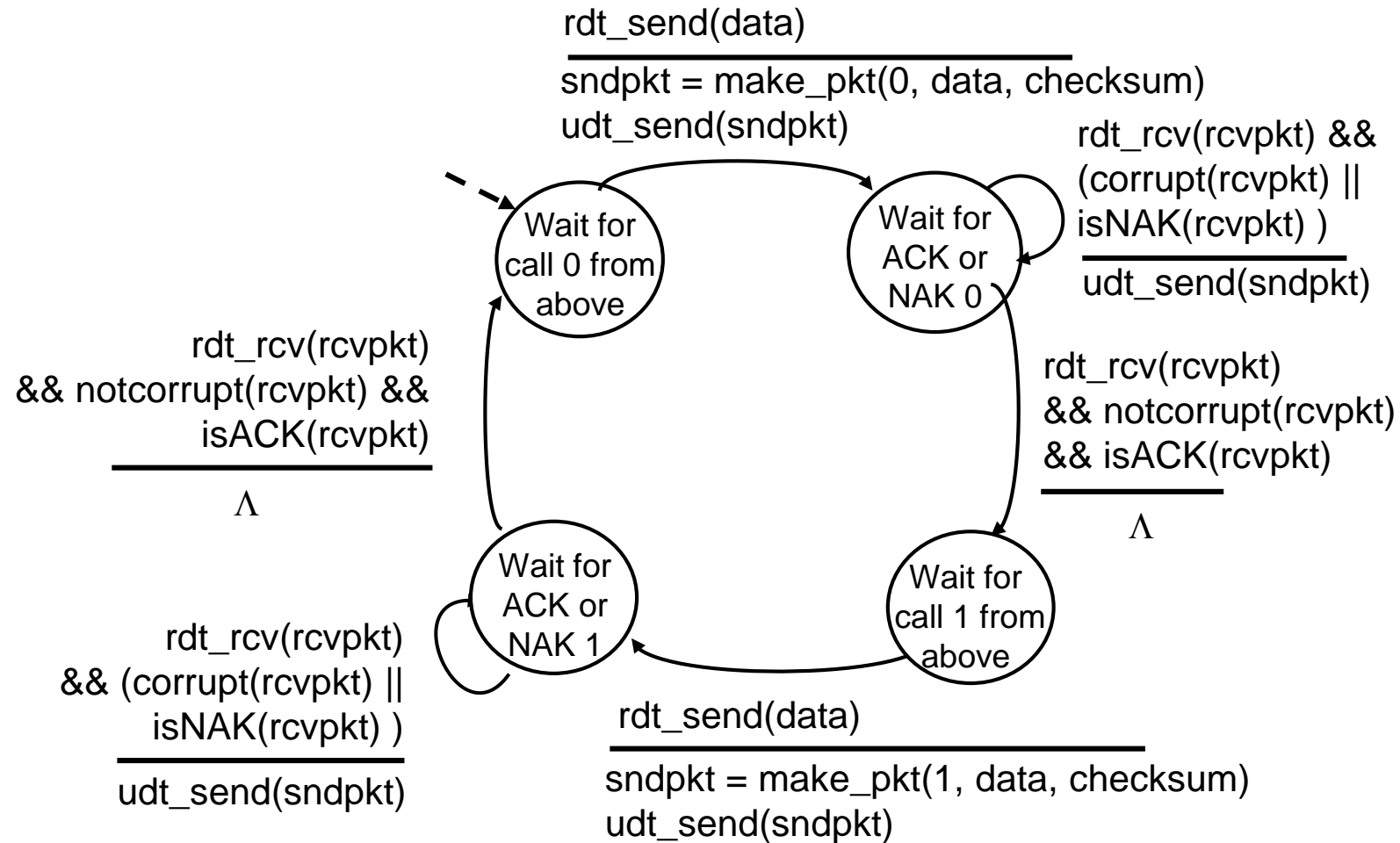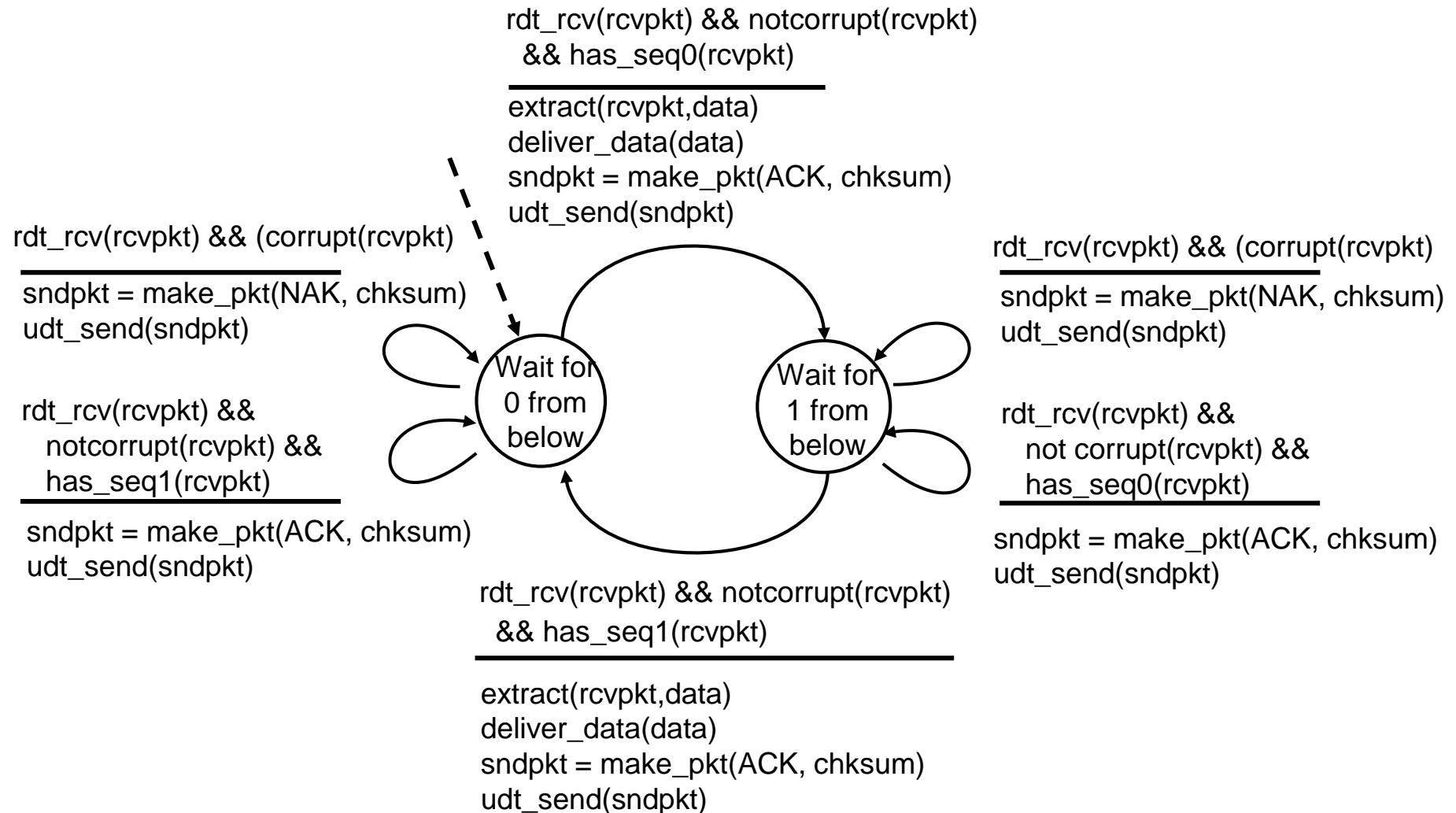# rdt2.1: receiver, handling garbled ACK/NAKs

rdt_rcv(rcvpkt) && notcorrupt(rcvpkt)
 && has_seq0(rcvpkt)
_____
extract(rcvpkt,data)
deliver_data(data)
sndpkt = make_pkt(ACK, chksum)
udt_send(sndpkt)

rdt_rcv(rcvpkt) && (corrupt(rcvpkt)
_____
sndpkt = make_pkt(NAK, chksum)
udt_send(sndpkt)

rdt_rcv(rcvpkt) &&
  notcorrupt(rcvpkt) &&
  has_seq1(rcvpkt)
_____
sndpkt = make_pkt(ACK, chksum)
udt_send(sndpkt)

rdt_rcv(rcvpkt) && (corrupt(rcvpkt)
_____
sndpkt = make_pkt(NAK, chksum)
udt_send(sndpkt)

rdt_rcv(rcvpkt) &&
  not corrupt(rcvpkt) &&
  has_seq0(rcvpkt)
_____
sndpkt = make_pkt(ACK, chksum)
udt_send(sndpkt)

Wait for
0 from
below

Wait for
1 from
below

rdt_rcv(rcvpkt) && notcorrupt(rcvpkt)
 && has_seq1(rcvpkt)
_____
extract(rcvpkt,data)
deliver_data(data)
sndpkt = make_pkt(ACK, chksum)
udt_send(sndpkt)

rdt2.1: discussion

sender:

- seq # added to pkt
- two seq. #s (0,1) will suffice. Why?
- must check if received ACK/NAK corrupted
- twice as many states
  - state must "remember" whether "expected" pkt should have seq # of 0 or 1
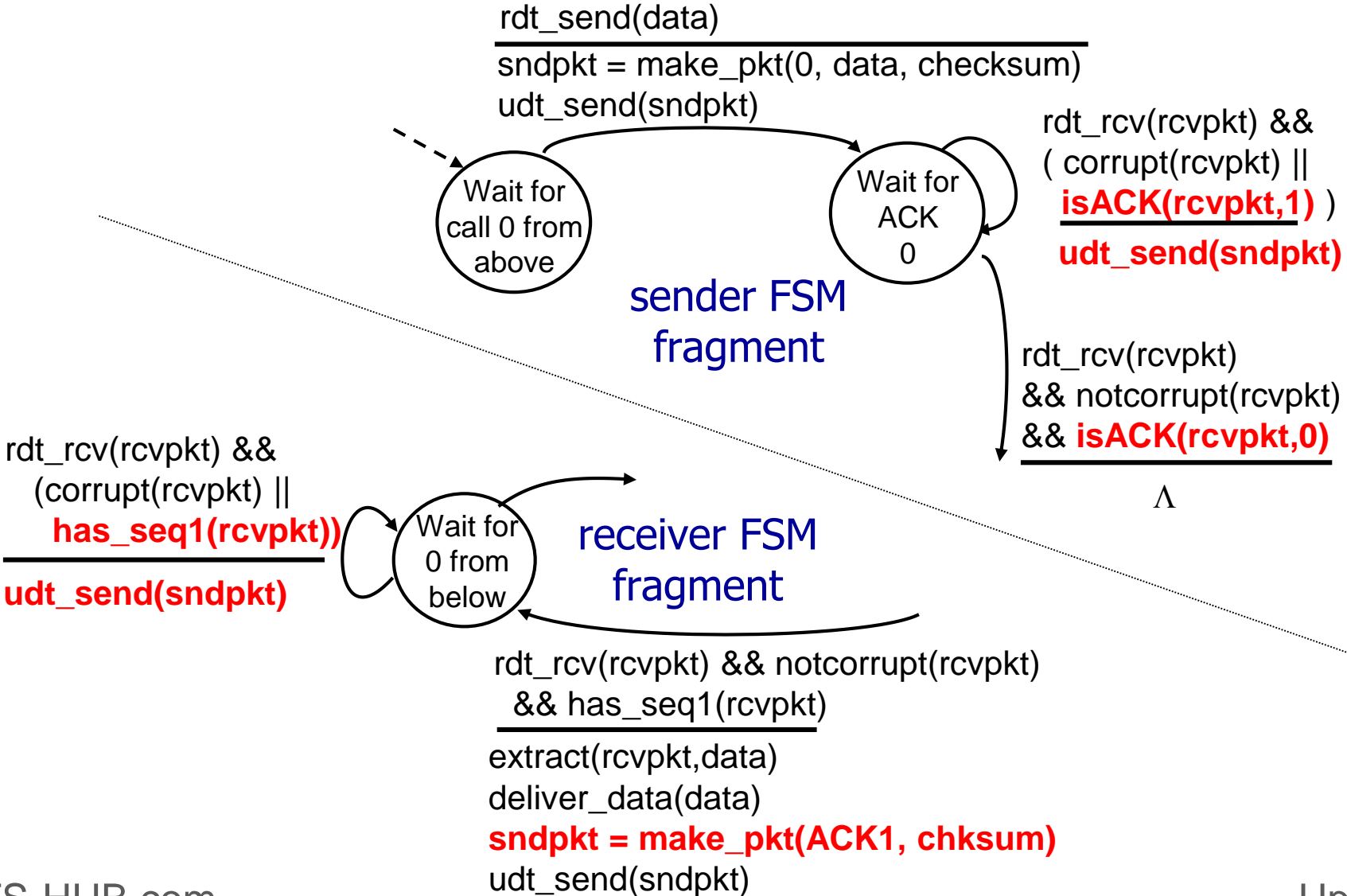
receiver:

- must check if received packet is duplicate
  - state indicates whether 0 or 1 is expected pkt seq #
- note: receiver can *not* know if its last ACK/NAK received OK at sender

rdt2.2: a NAK-free protocol

- same functionality as rdt2.1, using ACKs only

- instead of NAK, receiver sends ACK for last pkt received OK
  - receiver must *explicitly* include seq # of pkt being ACKed

- duplicate ACK at sender results in same action as NAK: *retransmit current pkt*

As we will see, TCP uses this approach to be NAK-free

rdt2.2: sender, receiver fragments

rdt_send(data)
___
sndpkt = make_pkt(0, data, checksum)
udt_send(sndpkt)

Wait for call 0 from above

Wait for ACK 0

rdt_rcv(rcvpkt) &&
(corrupt(rcvpkt) ||
isACK(rcvpkt,1))
___
udt_send(sndpkt)

rdt_rcv(rcvpkt)
&& notcorrupt(rcvpkt)
&& **isACK(rcvpkt,0)**
___
Λ

sender FSM fragment

rdt_rcv(rcvpkt) &&
(corrupt(rcvpkt) ||
has_seq1(rcvpkt))
___
udt_send(sndpkt)

Wait for 0 from below

receiver FSM fragment

rdt_rcv(rcvpkt) && notcorrupt(rcvpkt)
&& has_seq1(rcvpkt)
___
extract(rcvpkt,data)
deliver_data(data)
sndpkt = make_pkt(ACK1, chksum)
udt_send(sndpkt)

rdt3.0: channels with errors *and* loss

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?
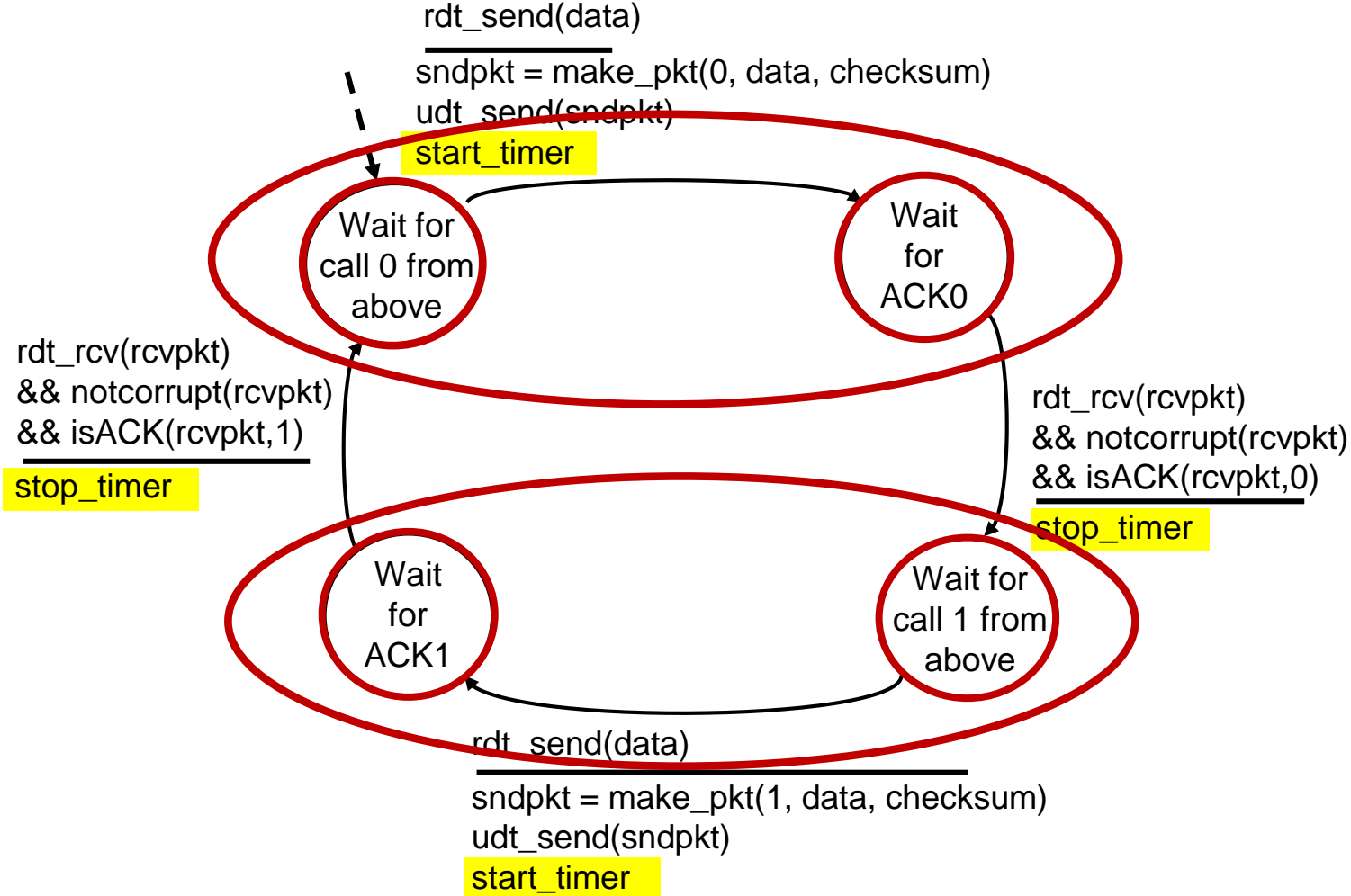
rdt3.0: channels with errors *and* loss

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
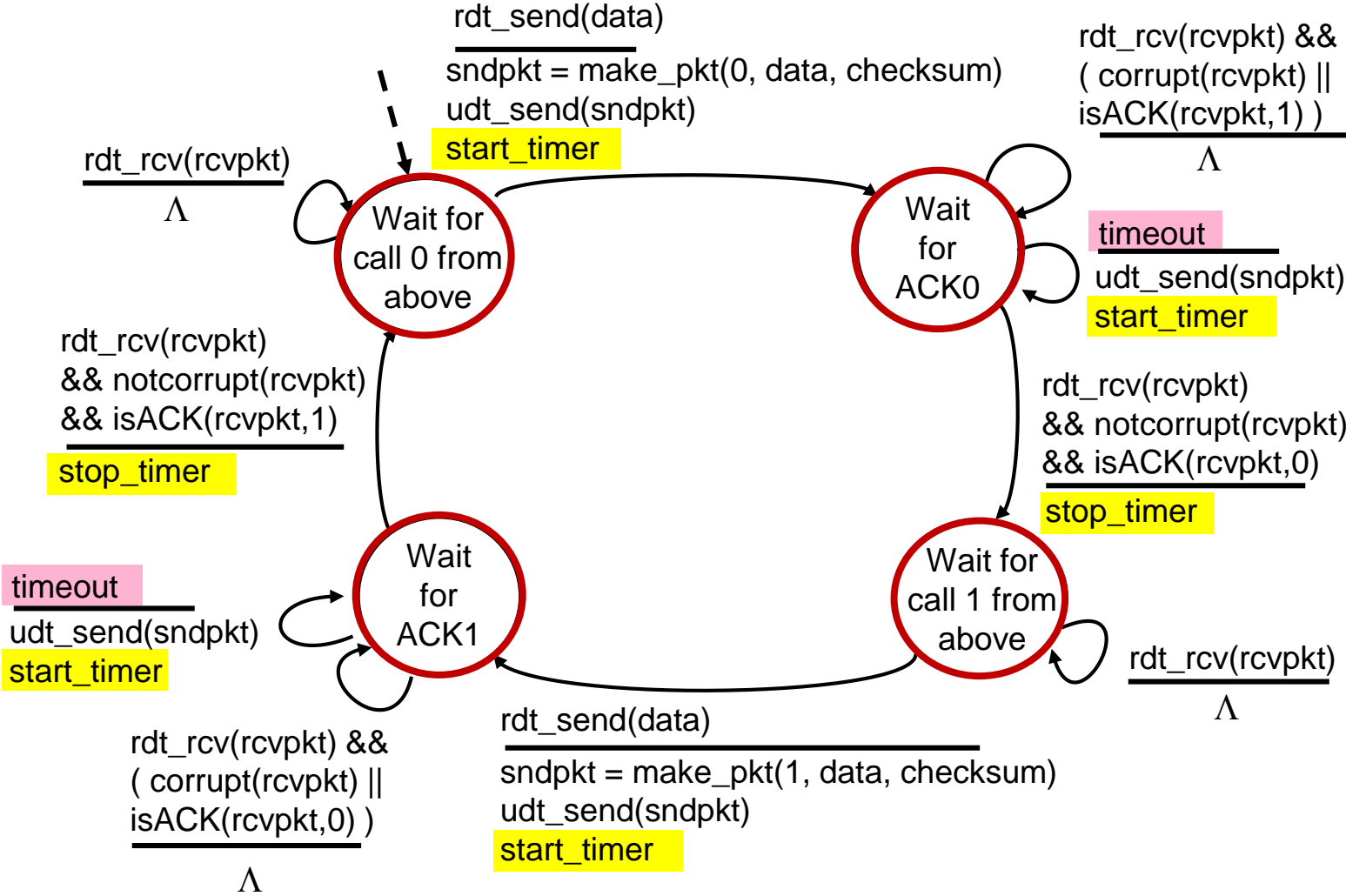- use countdown timer to interrupt after "reasonable" amount of time
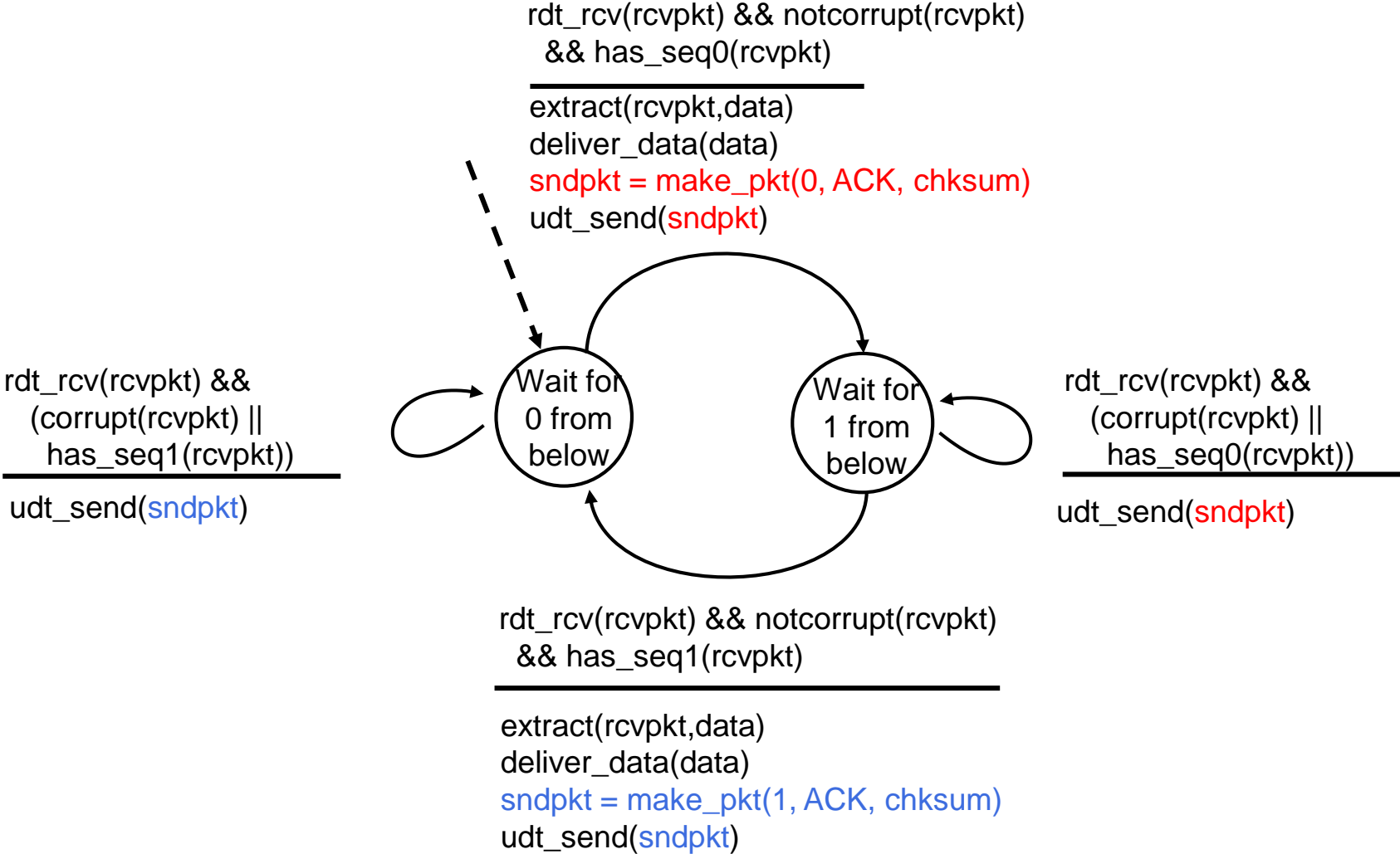
timeout

rdt3.0 sender

rdt_send(data)
_____

sndpkt = make_pkt(0, data, checksum)
udt_send(sndpkt)
start_timer

Wait for call 0 from above → Wait for ACK0

rdt_rcv(rcvpkt)
&& notcorrupt(rcvpkt)
&& isACK(rcvpkt,1)
_____
stop_timer

rdt_rcv(rcvpkt)
&& notcorrupt(rcvpkt)
&& isACK(rcvpkt,0)
_____
stop_timer

Wait for ACK1

Wait for call 1 from above

rdt_send(data)
_____
sndpkt = make_pkt(1, data, checksum)
udt_send(sndpkt)
start_timer

rdt3.0 sender

rdt3.0 receiver

rdt_rcv(rcvpkt) && notcorrupt(rcvpkt)
&& has_seq0(rcvpkt)
_____
extract(rcvpkt,data)
deliver_data(data)
sndpkt = make_pkt(0, ACK, chksum)
udt_send(sndpkt)

rdt_rcv(rcvpkt) &&
(corrupt(rcvpkt) ||
has_seq1(rcvpkt))
_____
udt_send(sndpkt)

Wait for 0 from below

Wait for 1 from below

rdt_rcv(rcvpkt) &&
(corrupt(rcvpkt) ||
has_seq0(rcvpkt))
_____
udt_send(sndpkt)

rdt_rcv(rcvpkt) && notcorrupt(rcvpkt)
&& has_seq1(rcvpkt)
_____
extract(rcvpkt,data)
deliver_data(data)
sndpkt = make_pkt(1, ACK, chksum)
udt_send(sndpkt)

rdt3.0 in action

sender   **receiver**

send pkt0 → pkt0 → rcv pkt0
send ack0

rcv ack0 ← ack0
send pkt1 → pkt1 → rcv pkt1
send ack1

rcv ack1 ← ack1
send pkt0 → pkt0 → rcv pkt0
send ack0

← ack0

(a) no loss

sender   **receiver**

send pkt0 → pkt0 → rcv pkt0
send ack0

rcv ack0 ← ack0
send pkt1 → pkt1 → **X**
loss

timeout
resend pkt1 → pkt1 → rcv pkt1
send ack1

rcv ack1 ← ack1
send pkt0 → pkt0 → rcv pkt0
send ack0

← ack0

(b) packet loss

rdt3.0 in action

(c) ACK loss

sender

send pkt0 → pkt0 → rcv pkt0 / send ack0

rcv ack0 / send pkt1 ← ack0

send pkt1 → pkt1 → rcv pkt1 / send ack1

ack1 ✗ loss

timeout → resend pkt1 → pkt1 → rcv pkt1 (detect duplicate) / send ack1

rcv ack1 / send pkt0 ← ack1

send pkt0 → pkt0 → rcv pkt0 / send ack0

← ack0

receiver

(c) ACK loss

(d) premature timeout/ delayed ACK

sender

send pkt0 → pkt0 → rcv pkt0 / send ack0

rcv ack0 / send pkt1 ← ack0

send pkt1 → pkt1 → rcv pkt1 / send ack1

timeout → resend pkt1 ← ack1

resend pkt1 → pkt1 → rcv pkt1 (detect duplicate) / send ack1

rcv ack1 / send pkt0 ← ack1

send pkt0 → pkt0 → rcv pkt0 / send ack0

rcv ack1 (ignore) ← ack1

← ack0

pkt1 →

receiver

(d) premature timeout/ delayed ACK

Performance of rdt3.0 (stop-and-wait)

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

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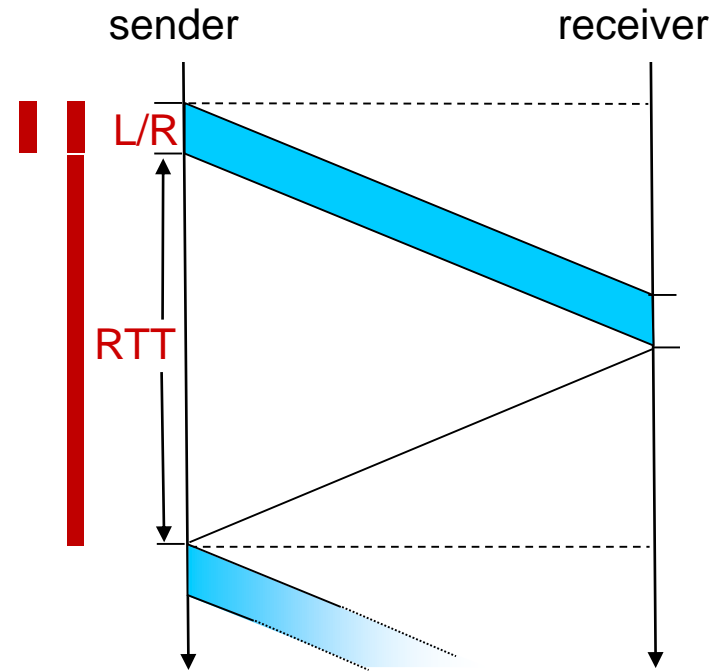
  - time to transmit packet into channel:

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

rdt3.0: stop-and-wait operation

sender

receiver

first packet bit transmitted, t = 0

first packet bit arrives

last packet bit arrives, send ACK

RTT

ACK arrives, send next packet, t = RTT + L / R
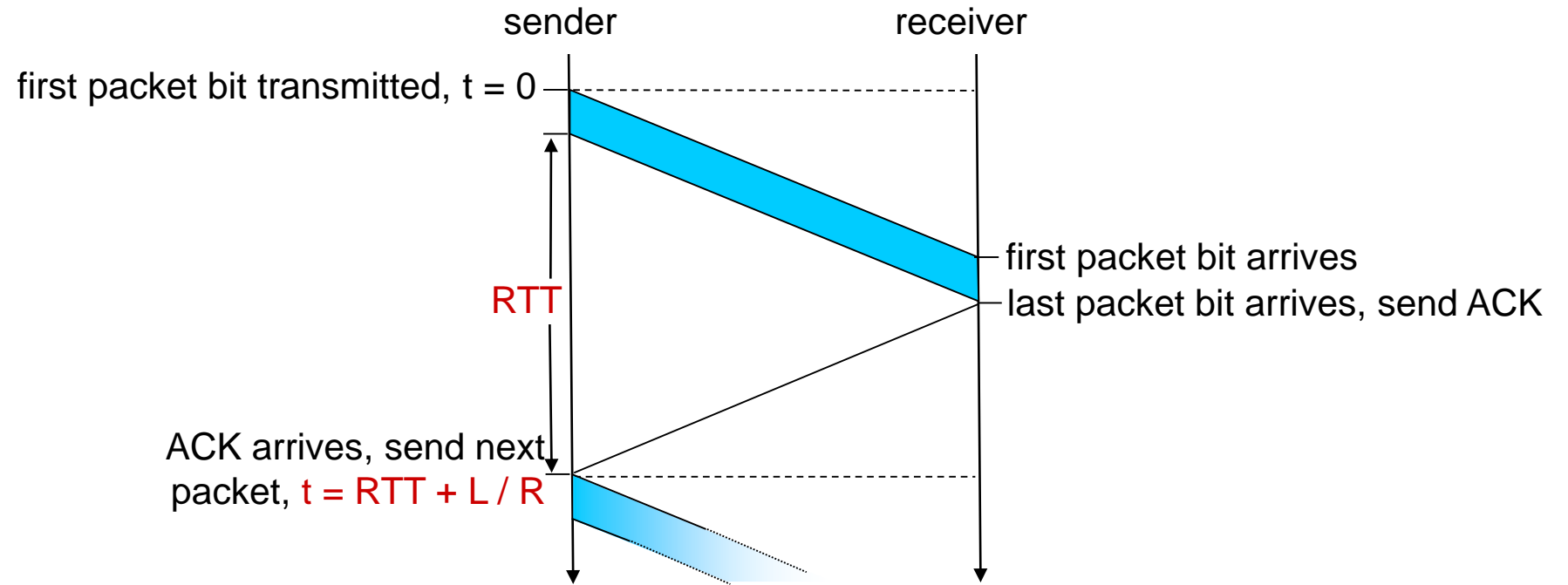
rdt3.0: stop-and-wait operation

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

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

$$= 0.00027$$

sender   receiver

L/R

RTT

- rdt 3.0 protocol performance stinks!
- Protocol limits performance of underlying infrastructure (channel)
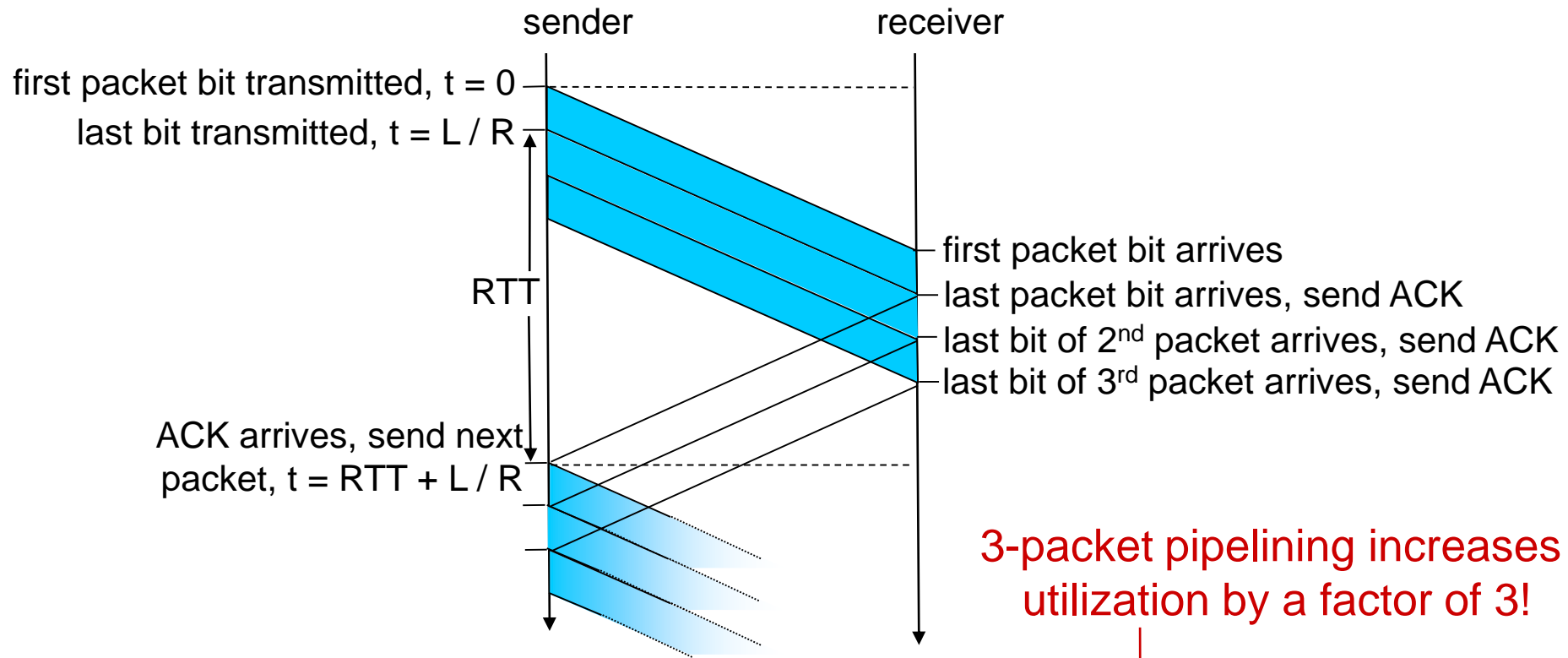
rdt3.0: pipelined protocols operation

pipelining: sender allows multiple, "in-flight", yet-to-be-acknowledged packets

- range of sequence numbers must be increased
- buffering at sender and/or receiver
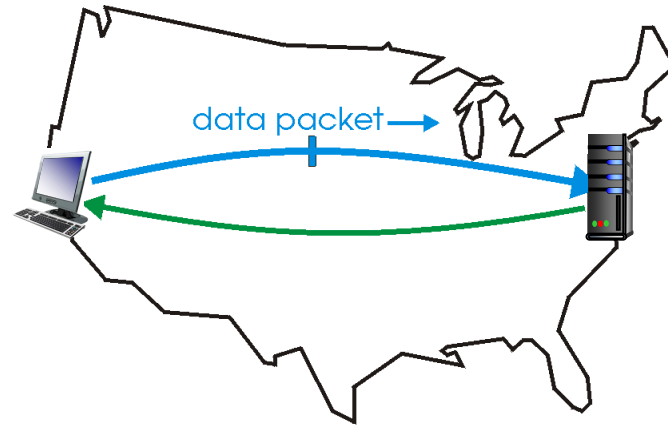
(a) a stop-and-wait protocol in operation

Pipelining: increased utilization

sender   receiver

first packet bit transmitted, t = 0

last bit transmitted, t = L / R

RTT

first packet bit arrives

last packet bit arrives, send ACK

last bit of 2nd packet arrives, send ACK

last bit of 3rd packet arrives, send ACK

ACK arrives, send next
packet, t = RTT + L / R

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

$$U_{sender} = \frac{3L / R}{RTT + L / R} = \frac{0.024}{30.008} = 0.0008$$

Go-Back-N: sender

- sender: "window" of up to N, consecutive transmitted but unACKed pkts
 - k-bit seq # in pkt header
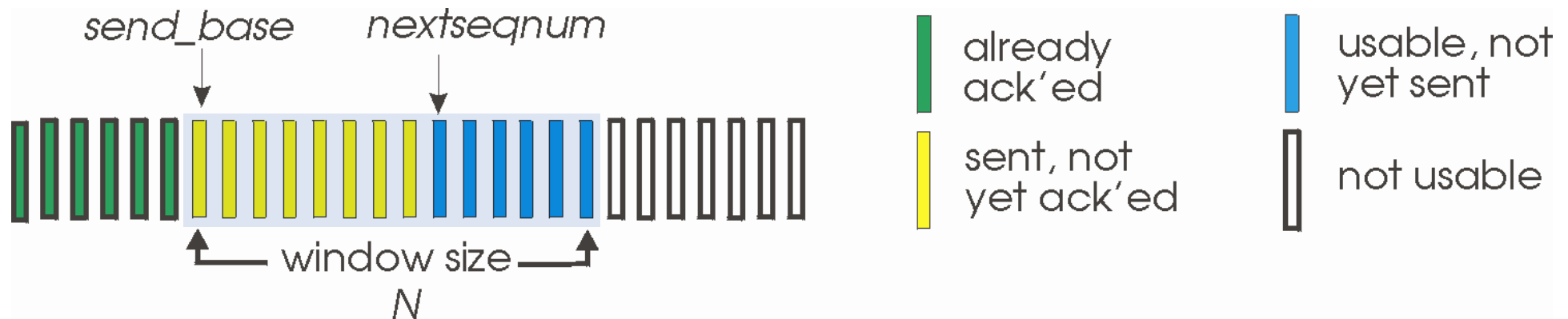
- *cumulative ACK:* ACK(n): ACKs all packets up to, including seq # n
  - on receiving ACK(n): move window forward to begin at $n+1$
- timer for oldest in-flight packet
- *timeout(n):* retransmit packet n and all higher seq # packets in window

Go-Back-N: receiver

- ■ ACK-only: always send ACK for correctly-received packet so far, with highest *in-order* seq #
  - 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:

... ▮▮▮▮▮▯▮▮▮▯▯▯▯▯ ...

↑
`rcv_base`

▮ received and ACKed

▮ Out-of-order: received but not ACKed

▯ Not received
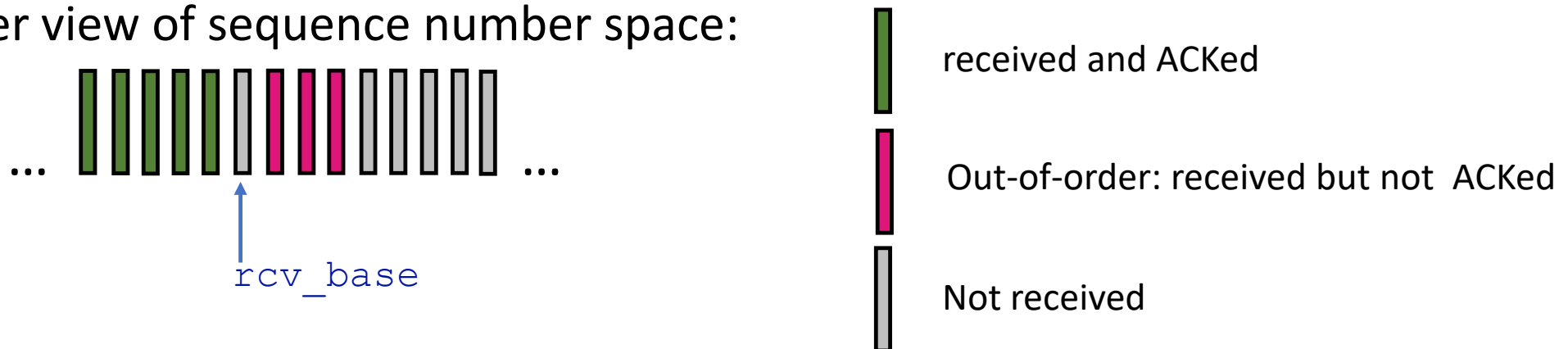
Go-Back-N in action

sender window (N=4)   sender   receiver   `rcv_base = 0`

`0 1 2 3` 4 5 6 7 8   send pkt0

`0 1 2 3` 4 5 6 7 8   send pkt1   receive pkt0, send ack0 (`rcv_base = 1`)

`0 1 2 3` 4 5 6 7 8   send pkt2   receive pkt1, send ack1 (`rcv_base = 2`)
                                              X *loss*

`0 1 2 3` 4 5 6 7 8   send pkt3

                             (wait)   receive pkt3, discard,
                                                                  (re)send ack1 (`rcv_base = 2`)

0 `1 2 3 4` 5 6 7 8   rcv ack0, send pkt4

0 1 `2 3 4 5` 6 7 8   rcv ack1, send pkt5   receive pkt4, discard,
                                                                  (re)send ack1 (`rcv_base = 2`)

                             ignore duplicate ACK   receive pkt5, discard,
                                                                  (re)send ack1 (`rcv_base = 2`)

                             pkt 2 timeout

0 1 `2 3 4 5` 6 7 8   send pkt2
0 1 `2 3 4 5` 6 7 8   send pkt3
0 1 `2 3 4 5` 6 7 8   send pkt4   rcv pkt2, deliver, send ack2 (`rcv_base = 3`)
0 1 `2 3 4 5` 6 7 8   send pkt5   rcv pkt3, deliver, send ack3 (`rcv_base = 4`)
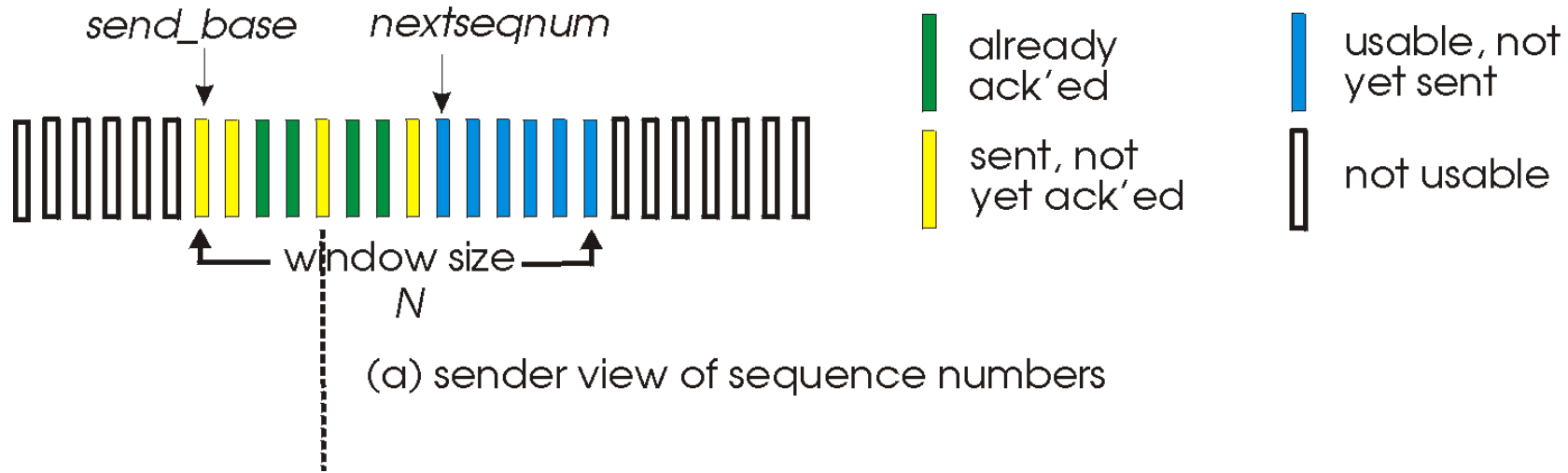   rcv pkt4, deliver, send ack4 (`rcv_base = 5`)
                                                           rcv pkt5, deliver, send ack5 (`rcv_base = 6`)

Selective repeat: the approach

- *pipelining*: *multiple* packets in flight

- *receiver individually ACKs* all correctly received packets
  - buffers packets, as needed, for in-order delivery to upper 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
    - limits pipelined, "in flight" packets to be within this window

Selective repeat: sender, receiver windows

(a) sender view of sequence numbers

send_base *nextseqnum*

■ already ack'ed

■ sent, not yet ack'ed

■ usable, not yet sent

□ not usable

window size N

Selective repeat: sender and receiver

sender

data from above:
- if next available seq # in window, send packet

timeout(*n*):
- resend packet *n*, restart timer

ACK(*n*) in [send_base,send_base+N-1]:
- mark packet *n* as received
- if n smallest unACKed packet, advance window base to next unACKed seq #

receiver

packet *n* in [rcv_base, rcv_base+N-1]
- send ACK(*n*)
- out-of-order: buffer
- in-order: deliver (also deliver buffered, in-order packets), advance window to next not-yet-received packet

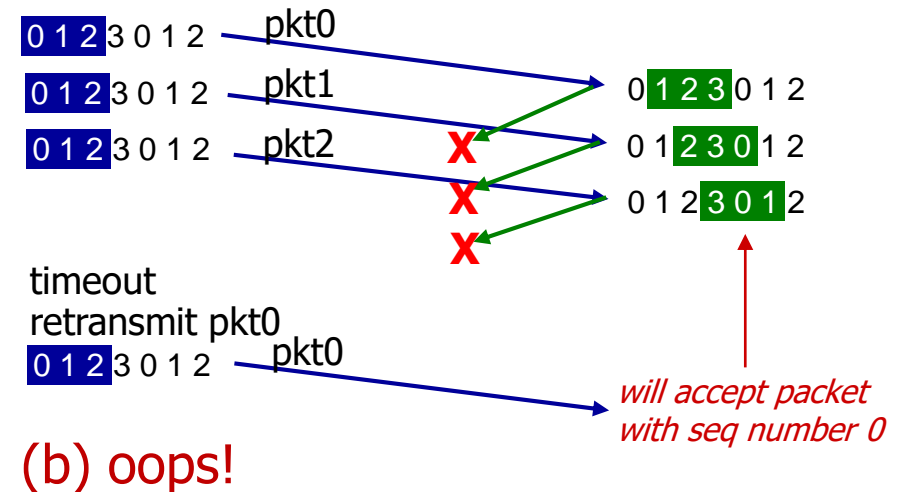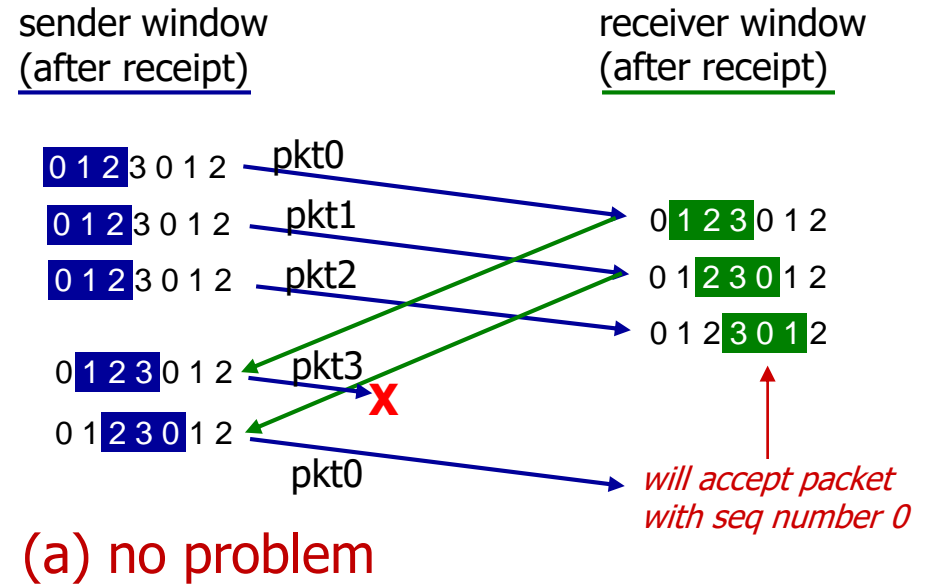### packet *n* in [rcv_base-N,rcv_base-1]
- ACK(*n*)

otherwise:
- ignore

Selective Repeat in action

sender window (N=4)

sender

receiver

`0 1 2 3 4 5 6 7 8`   send pkt0

`0 1 2 3 4 5 6 7 8`   send pkt1

`0 1 2 3 4 5 6 7 8`   send pkt2

`0 1 2 3 4 5 6 7 8`   send pkt3

(wait)

`0 1 2 3 4 5 6 7 8`   rcv ack0, send pkt4

`0 1 2 3 4 5 6 7 8`   rcv ack1, send pkt5

`0 1 2 3 4 5 6 7 8`   record ack3 arrived

pkt 2 timeout

`0 1 2 3 4 5 6 7 8`   send pkt2

`0 1 2 3 4 5 6 7 8`   (but not 3,4,5)

`0 1 2 3 4 5 6 7 8`

X *loss*

receive pkt0, send ack0

receive pkt1, send ack1

receive pkt3, buffer,
send ack3

receive pkt4, buffer,
send ack4

receive pkt5, buffer,
send ack5

rcv pkt2; deliver pkt2,
pkt3, pkt4, pkt5; send ack2

Q: what happens when ack2 arrives?

Selective repeat: a dilemma!

example:

- seq #s: 0, 1, 2, 3 (base 4 counting)
- window size=3

0 1 2 3 0 1 2   pkt0

0 1 2 3 0 1 2   pkt1

0 1 2 3 0 1 2   pkt2

0 1 2 3 0 1 2   pkt3

X

0 1 2 3 0 1 2

pkt0

0 1 2 3 0 1 2

0 1 2 3 0 1 2

0 1 2 3 0 1 2

*will accept packet
with seq number 0*

(a) no problem

0 1 2 3 0 1 2   pkt0

0 1 2 3 0 1 2   pkt1

0 1 2 3 0 1 2   pkt2   X

X

X

timeout
retransmit pkt0

0 1 2 3 0 1 2   pkt0

0 1 2 3 0 1 2

0 1 2 3 0 1 2

0 1 2 3 0 1 2

*will accept packet
with seq number 0*

(b) oops!

Selective repeat: a dilemma!

example:

- seq #s: 0, 1, 2, 3 (base 4 counting)
- window size = 3

Q: what relationship is needed between sequence # size and window size to avoid problem in scenario (b)?

sequence # size >= 2 * window size

0 1 2 3 0 1 2   pkt0
0 1 2 3 0 1 2   pkt1
0 1 2 3 0 1 2   pkt2

0 1 2 3 0 1 2   pkt3

0 1 **2 3** 0 1 2

0 1 2 **3 0** 1 2

0 1 2 **3 0 1** 2

- *receiver can't see sender side*
- *receiver behavior identical in both cases!*
- *something's (very) wrong!*

will accept packet with seq number 0

0 **1 2 3** 0 1 2

0 1 **2 3 0** 1 2

0 1 2 **3 0 1** 2

0 1 2 3 0 1 2   pkt2

timeout
retransmit pkt0
0 1 2 3 0 1 2   pkt0

(b) oops!

will accept packet with seq number 0
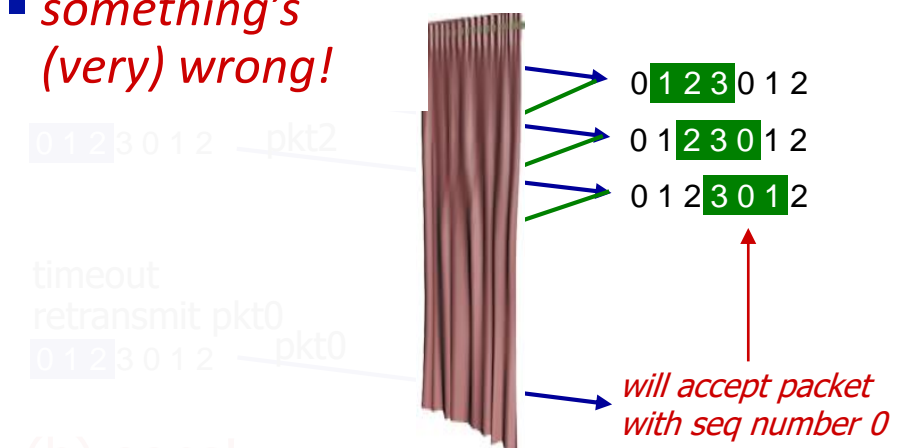
Selective repeat:
a dilemma!

sequence # size >= 2 * window size

example:

- window size = 3
- seq #s: 0, 1, 2, 3, 4, 5

Example:

➢ if window size is 16 ➡ SQN = 32 (0,…, 31). So, we need 5 bits for the SQN

➢ if window size is 2^8 ➡ we need 9 bits for the SQN = 512 (0,…, 511).

<u>sender window</u>
(after receipt)

<u>receiver window</u>
(after receipt)
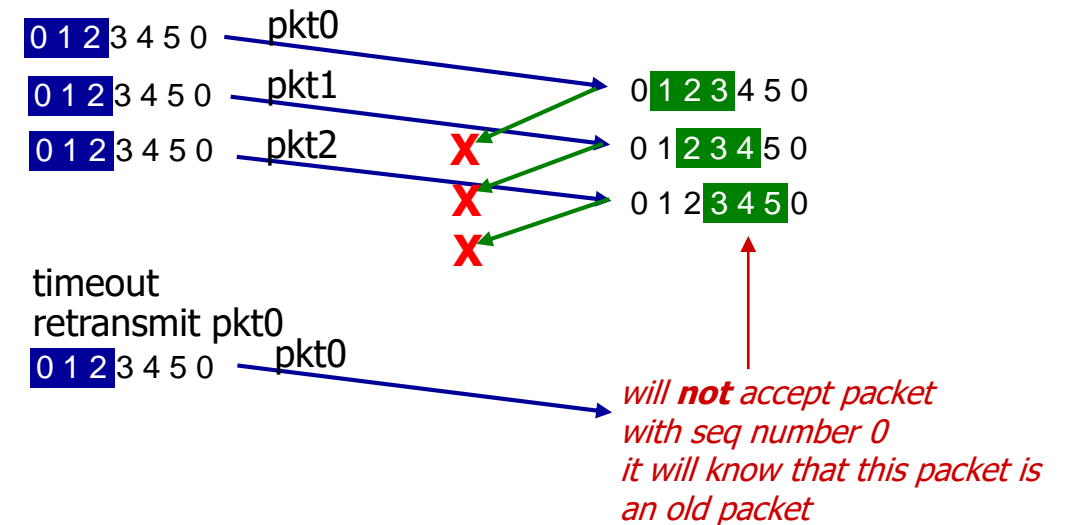
0 1 2 3 4 5 0 ——— pkt0

0 1 2 3 4 5 0 ——— pkt1   0 1 2 3 4 5 0

0 1 2 3 4 5 0 ——— pkt2   **X**   0 1 2 3 4 5 0

                             X   0 1 2 3 4 5 0

                             X

timeout
retransmit pkt0

0 1 2 3 4 5 0 ——— pkt0

*will **not** accept packet with seq number 0 it will know that this packet is an old packet*

Chapter 3: roadmap

- Transport-layer services

- Multiplexing and demultiplexing

- Connectionless transport: UDP

- Principles of reliable data transfer

- **Connection-oriented transport: TCP**

  - segment structure
  - reliable data transfer
  - flow control
  - connection management

- Principles of congestion control

- TCP congestion control

TCP: overview   RFCs: 793,1122, 2018, 5681, 7323

- **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:**
  - handshaking (exchange of control messages) initializes sender, receiver state before data exchange

- **flow controlled:**
  - sender will not overwhelm receiver

TCP segment structure

32 bits

ACK: seq # of next expected byte; A bit: this is an ACK

4-bit header length (in 32-bit Words)

Internet checksum

C, E: congestion notification

TCP options (used when a sender and receiver negotiate the MSS)

RST, SYN, FIN: connection management (used for connection setup and teardown)

PSH: indicates that the receiver should pass the data to the upper layer immediately

source port #	dest port #
sequence number	
acknowledgement number	

head len	not used	C	E	U	A	P	R	S	F	receive window

checksum	Urg data pointer

options (variable length)

application data (variable length)

segment seq #: counting bytes of data into bytestream (not segments!)

flow control: # bytes receiver willing to accept

URG: indicate that there is data in this segment that the sending-side upper- layer entity has marked as "urgent." The location of the last byte of this urgent data is indicated by the 16-bit **urgent data pointer** field.

data sent by application into TCP socket

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 (discards out-of-order segments or keeps the out-of-order bytes and waits for the missing bytes to fill in the gaps)

outgoing segment from sender

source port #	dest port #
sequence number	
acknowledgement number	
	rwnd
checksum	urg pointer

window size
N

sender sequence number space

| sent ACKed | sent, not-yet ACKed ("in-flight") | usable but not yet sent | not usable |
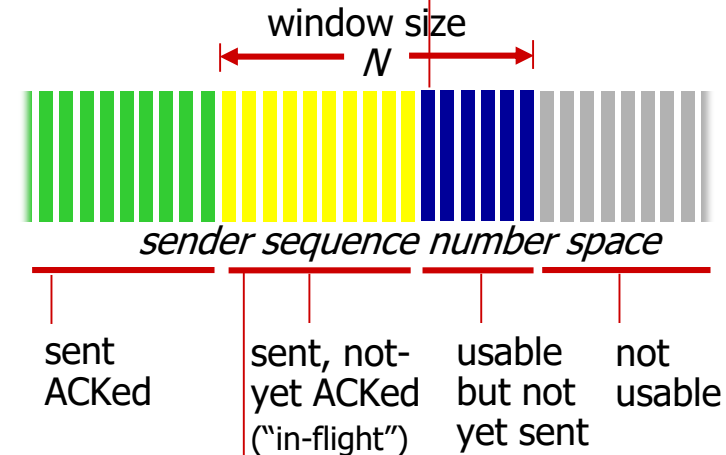
outgoing segment from receiver

source port #	dest port #
sequence number	
acknowledgement number	
	A	rwnd
checksum	urg pointer

TCP sequence numbers, ACKs

Host A
(assume **nextSeqNum** is **42**)

Host B
(assume **nextSeqNum** is **79**)

User types 'C'

Seq=42, ACK=79, data = 'C'

host ACKs receipt
of 'C', echoes back 'C'

Seq=79, ACK=43, data = 'C'

host ACKs receipt
of echoed 'C'

Seq=43, ACK=80
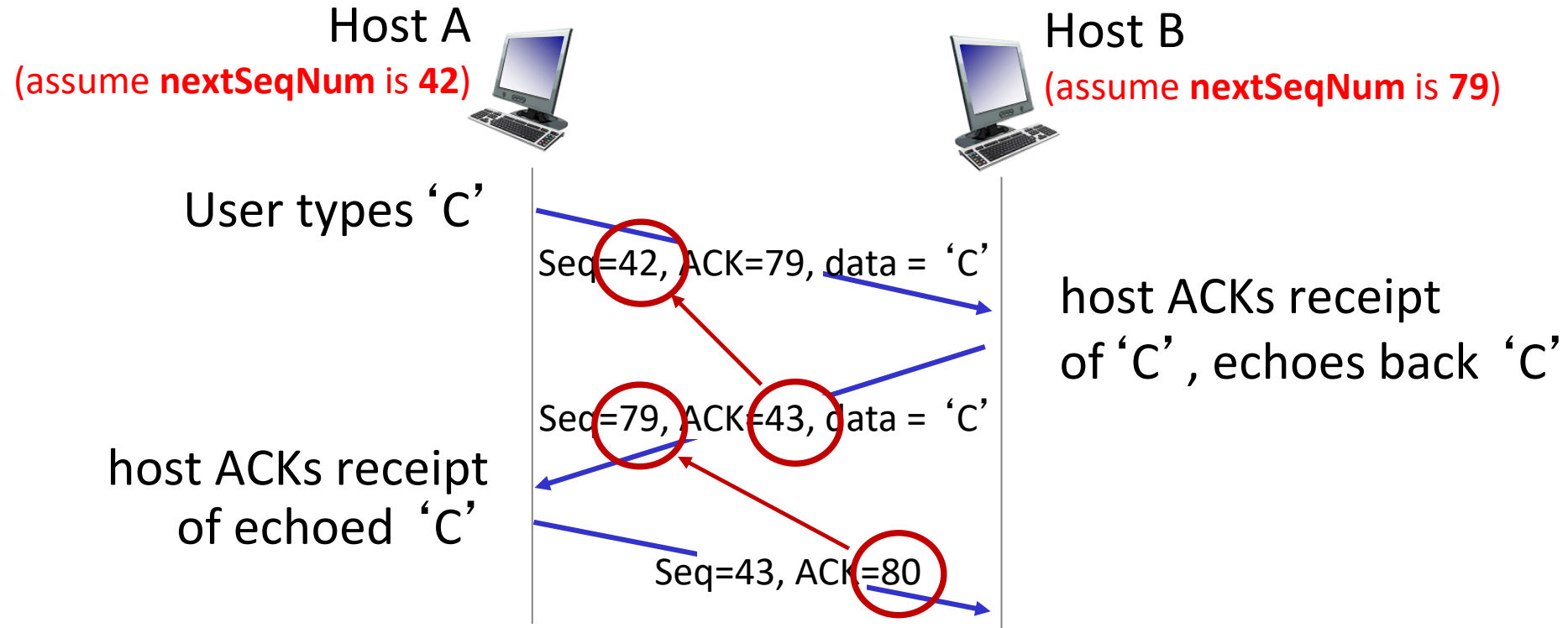
simple telnet scenario

TCP round trip time, timeout

Q: how to set TCP timeout value?

- longer than RTT, but RTT varies!
- *too short:* premature timeout, unnecessary retransmissions
- *too long:* slow reaction to segment loss

Q: how to estimate RTT?

- `SampleRTT:` measured time from segment transmission until ACK receipt
  - ignore retransmissions
- `SampleRTT` will vary, want estimated RTT "smoother"
  - average several *recent* measurements, not just current `SampleRTT`

TCP round trip time, timeout

$$\text{EstimatedRTT} = (1-\alpha)*\text{EstimatedRTT} + \alpha*\text{SampleRTT}$$

- exponential weighted moving average (EWMA)
- influence of past sample decreases exponentially fast
- typical value: $\alpha = 0.125$

Assume old **EstimatedRTT** = 25 ms, what is the new **EstimatedRTT** if the **SampleRTT** = 100 ms?
0.875*25 + 0.125*100 = 34.375 ms

Assume old **EstimatedRTT** = 25 ms, what is the new **EstimatedRTT** if the **SampleRTT** = 26 ms?
0.875*25 + 0.125*26 = 25.1255 ms

RTT: gaia.cs.umass.edu to fantasia.eurecom.fr

RTT (milliseconds)

time (seconds)

◆ sampleRTT
■ EstimatedRTT
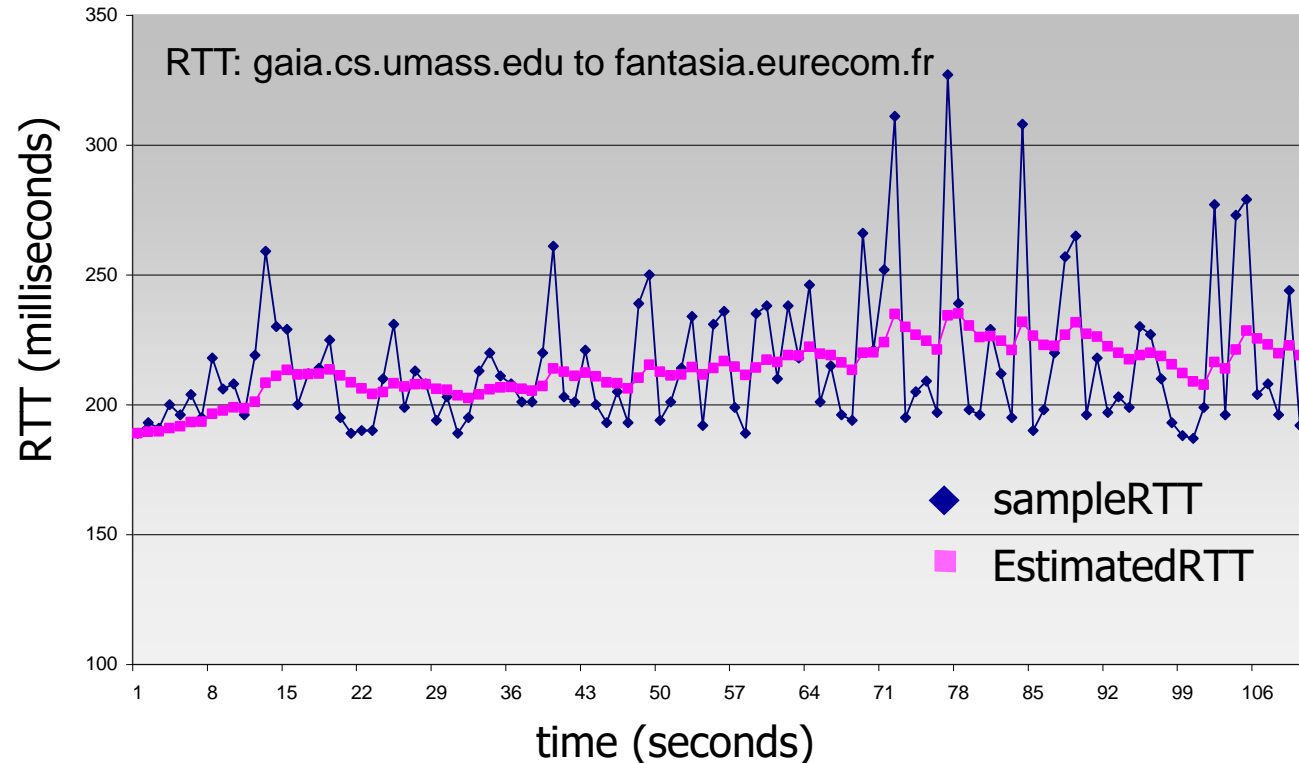
TCP round trip time, timeout

- timeout interval: `EstimatedRTT` plus "safety margin"

  - large variation in `EstimatedRTT`: want a larger safety margin

`TimeoutInterval = EstimatedRTT + 4*DevRTT`

estimated RTT   "safety margin"

- `DevRTT`: EWMA of `SampleRTT` deviation from `EstimatedRTT`:

`DevRTT = (1-β)*DevRTT + β*|SampleRTT-EstimatedRTT|`

(typically, β = 0.25)

TCP Sender (simplified)

event: data received from application

- create segment with seq #

- seq # is byte-stream number of first data byte in segment

- start timer if not already running
  - think of timer as for oldest unACKed segment
  - expiration interval: **`TimeOutInterval`**

event: timeout

- retransmit segment that caused timeout
- restart timer

event: ACK received

- if ACK acknowledges previously unACKed segments
  - update what is known to be ACKed
  - start timer if there are still unACKed segments

TCP sender (simplified)

data received from application above
_____
create segment, seq. #: NextSeqNum
pass segment to IP (i.e., "send")
NextSeqNum = NextSeqNum + length(data)
if (timer currently not running)
    start timer

Λ
_____
NextSeqNum = InitialSeqNum
SendBase = InitialSeqNum

wait for event

timeout
_____
retransmit not-yet-acked segment
                with smallest seq. #
start timer

ACK received, with ACK field value y
_____
if (y > SendBase) {
    SendBase = y
    /* SendBase–1: last cumulatively ACKed byte */
    if (there are currently not-yet-acked segments)
        start timer
      else stop timer
    }

TCP Receiver: ACK generation [RFC 5681]

Event at receiver	TCP receiver action

TCP: retransmission scenarios

Host A   Host B

Seq=92, 8 bytes of data

timeout

ACK=100

X

Seq=92, 8 bytes of data

ACK=100

lost ACK scenario

Host A   Host B

SendBase=92

Seq=92, 8 bytes of data

timeout

Seq=100, 20 bytes of data

ACK=100

ACK=120

SendBase=100

Seq=92, 8 bytes of data

SendBase=120

send cumulative ACK for 120

ACK=120

SendBase=120
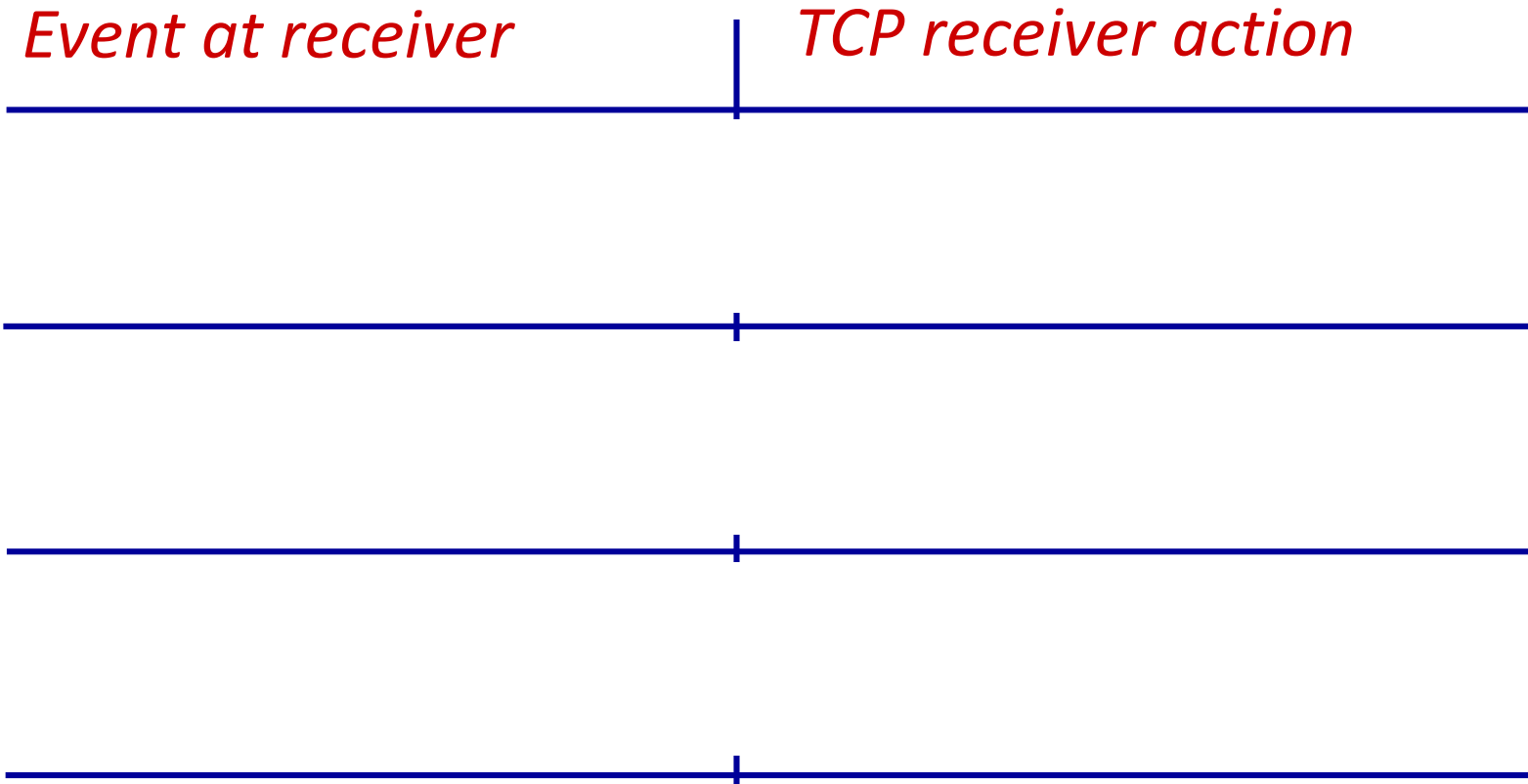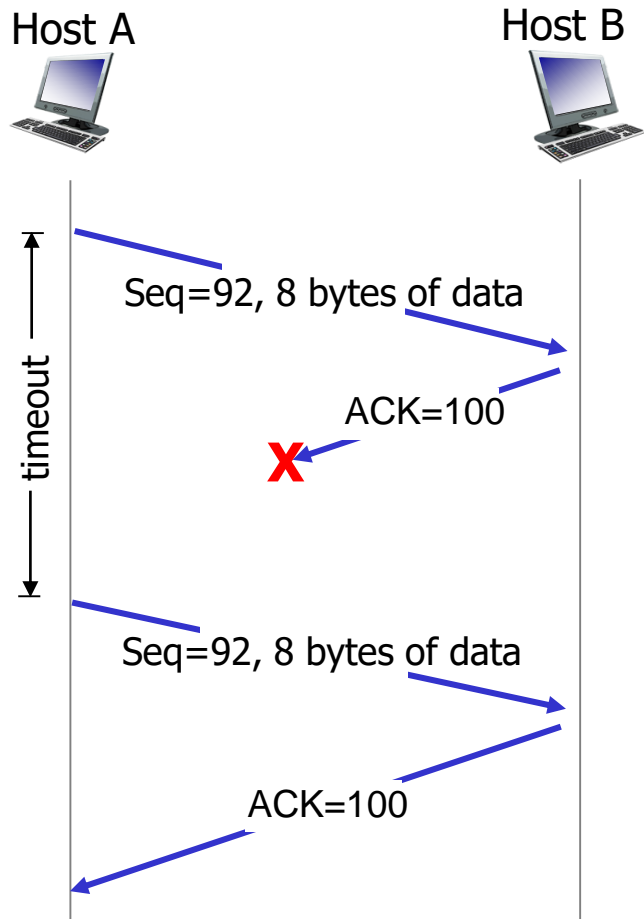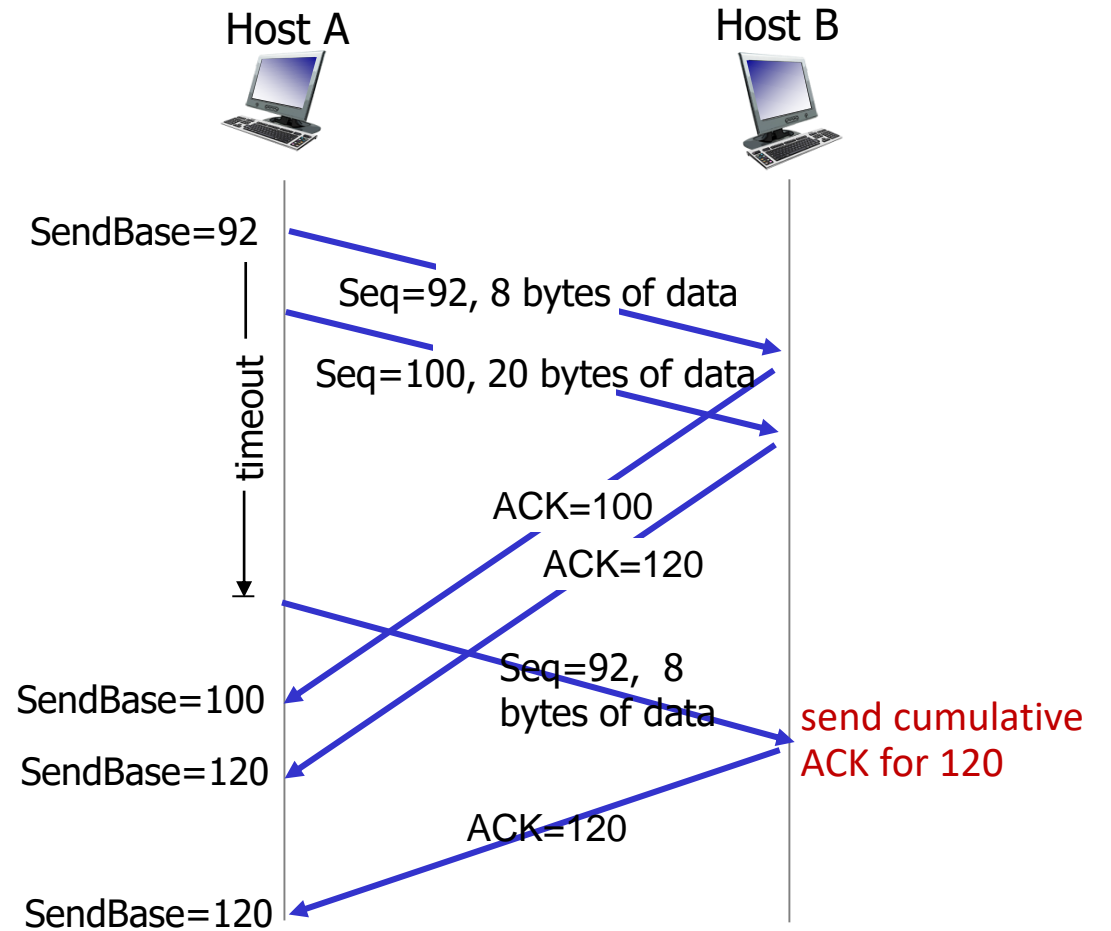
premature timeout

TCP: retransmission scenarios

Host A   Host B

Seq=92, 8 bytes of data

Seq=100, 20 bytes of data

ACK=100

X

ACK=120

Seq=120, 15 bytes of data

cumulative ACK covers
for earlier lost ACK
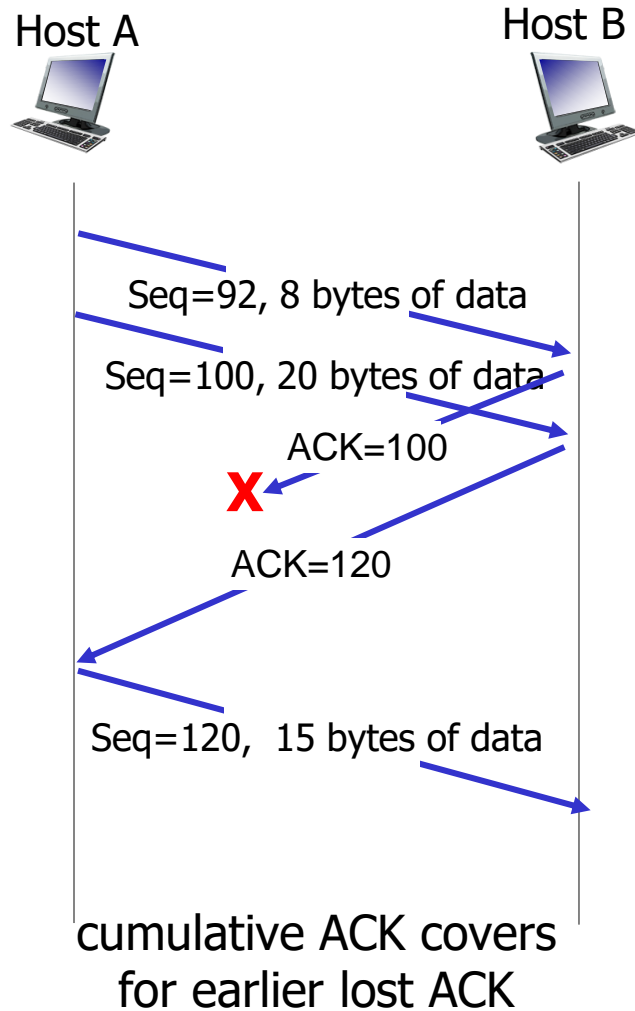
TCP fast retransmit

if sender receives 3 additional ACKs for same data ("triple duplicate ACKs"), resend unACKed segment with smallest seq #

- likely that unACKed segment lost, so don't wait for timeout

💡 Receipt of three duplicate ACKs indicates 3 segments received after a missing segment – lost segment is likely. So retransmit!
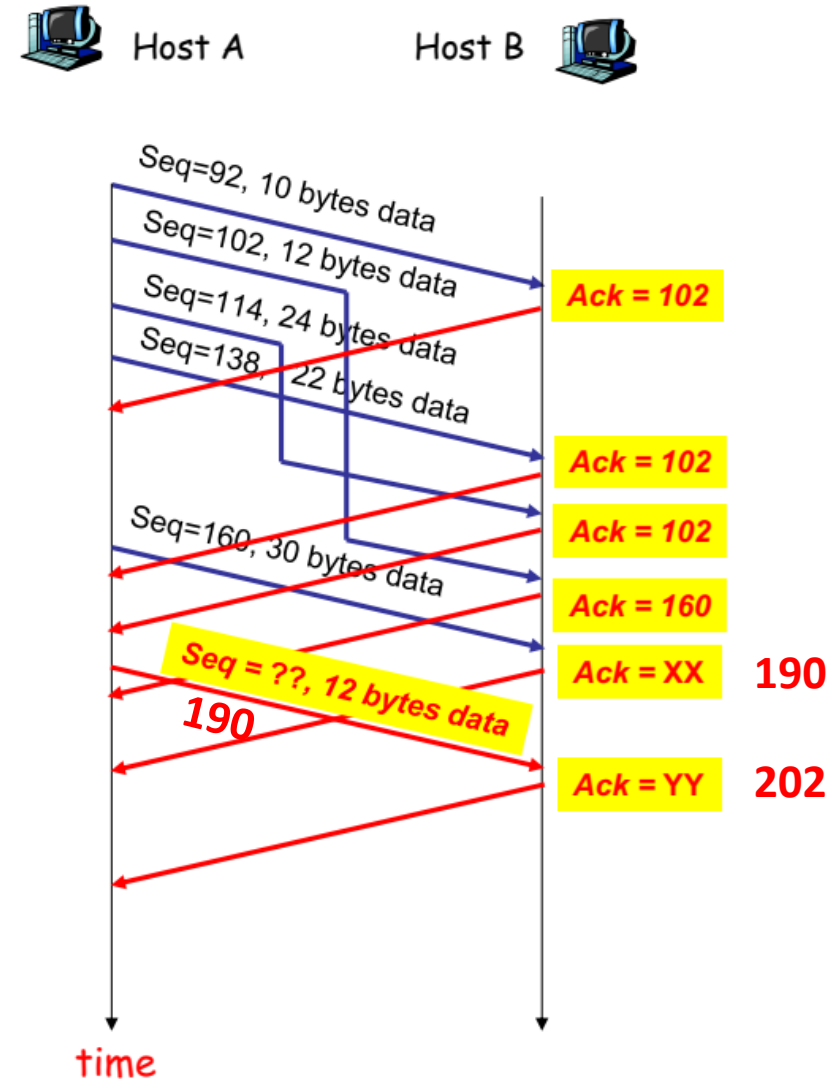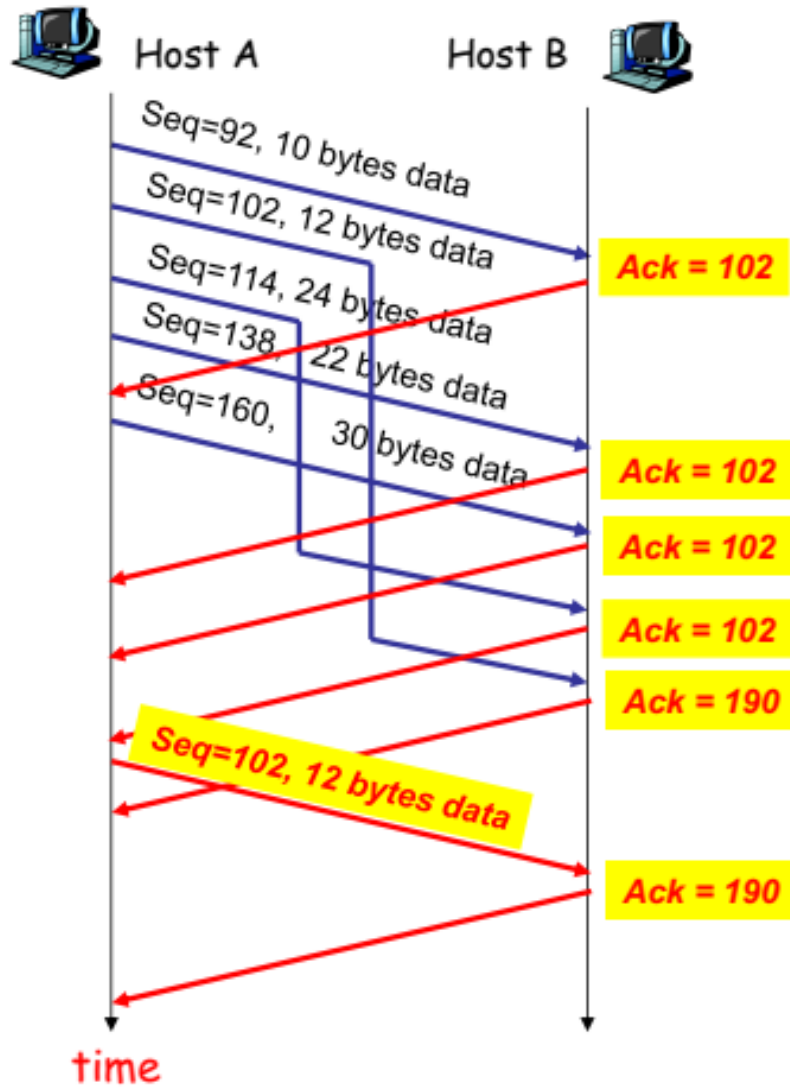
Host A   Host B

Seq=92, 8 bytes of data
Seq=100, 20 bytes of data   X

ACK=100
ACK=100
ACK=100
ACK=100

Seq=100, 20 bytes of data

timeout

Example

Chapter 3: roadmap

- Transport-layer services

- Multiplexing and demultiplexing

- Connectionless transport: UDP

- Principles of reliable data transfer

- **Connection-oriented transport: TCP**

  - segment structure
  - reliable data transfer
  - flow control
  - connection management

- Principles of congestion control

- TCP congestion control

TCP flow control

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

Application removing data from TCP socket buffers

application process

TCP socket receiver buffers

TCP code

Network layer delivering IP datagram payload into TCP socket buffers

IP code

from sender

receiver protocol stack

TCP flow control

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

Application removing data from TCP socket buffers ——

Network layer delivering IP datagram payload into TCP socket buffers ——

application process

TCP socket receiver buffers

TCP code

IP code

from sender

receiver protocol stack

TCP flow control

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

receive window ——— flow control: # bytes receiver willing to accept

Application removing data from TCP socket buffers ———

application process

TCP socket receiver buffers

TCP code

IP code

from sender

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

Application removing data from TCP socket buffers

application process

TCP socket receiver buffers

TCP code

IP code

from sender

receiver protocol stack

TCP flow control

- TCP receiver "advertises" free buffer space in the receive window (**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

to application process

RcvBuffer

buffered data

rwnd

free buffer space

TCP segment payloads

TCP receiver-side buffering

TCP flow control

- TCP receiver "advertises" free buffer space in the receive window (**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**
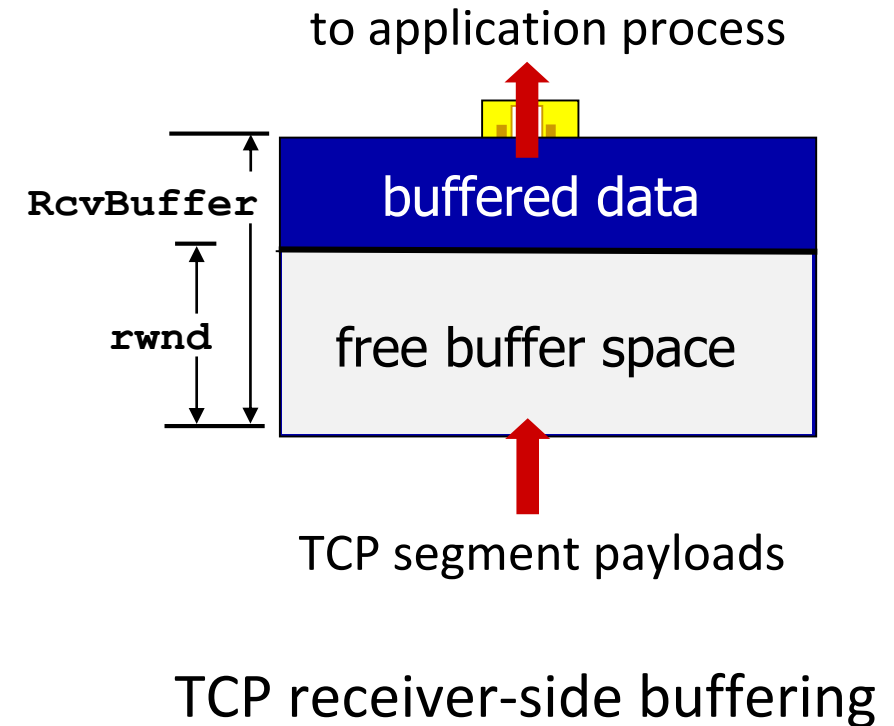
- guarantees receive buffer will not overflow

flow control: # bytes receiver willing to accept

receive window
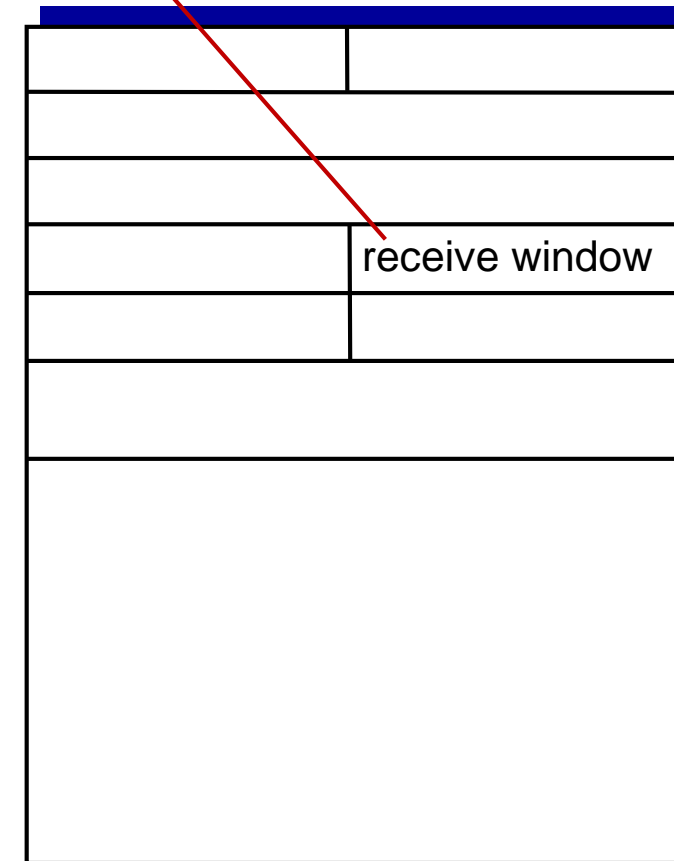
TCP segment format

TCP flow control

- **At Receiver:**
  - `LastByteRead` (by the application)
  - `LastByteRcvd` (arrived from the network)
  - `LastByteRcvd - LastByteRead <= RcvBuffer`
  - `rwnd = RcvBuffer - (LastByteRcvd - LastByteRead)`

- **At Sender:**
  - `LastByteSent`
  - `LastByteAcked`
  - `LastByteSent - LastByteAcked <= rwnd`

to application process

RcvBuffer

buffered data

rwnd

free buffer space

TCP segment payloads

TCP receiver-side buffering
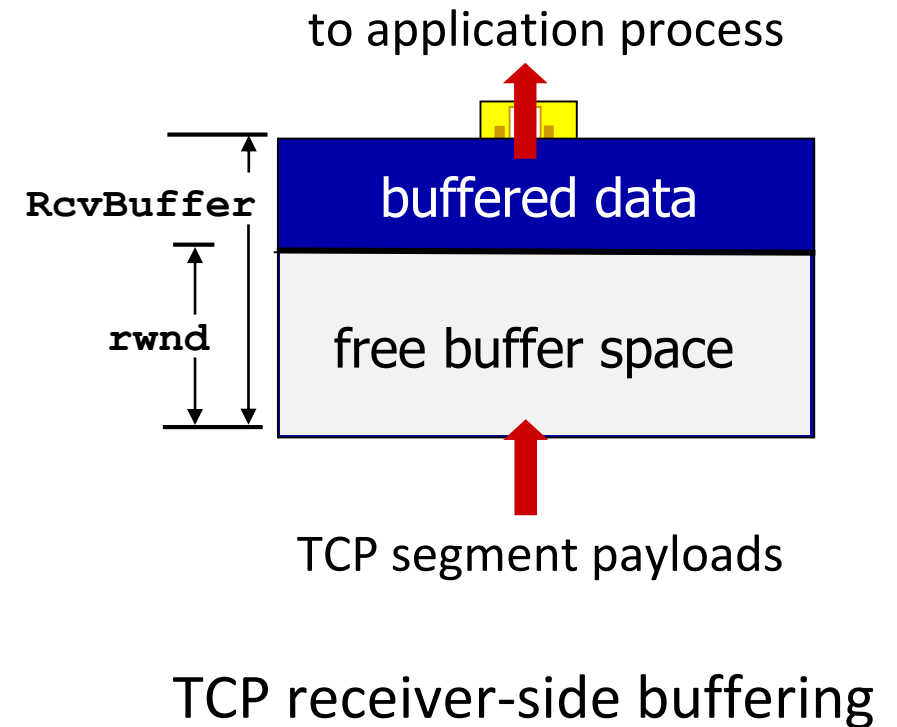
TCP connection management

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)

application

connection state: ESTAB
connection variables:
   seq # client-to-server
        server-to-client
   rcvBuffer size
    at server,client

network

application

connection state: ESTAB
connection Variables:
   seq # client-to-server
        server-to-client
   rcvBuffer size
    at server,client
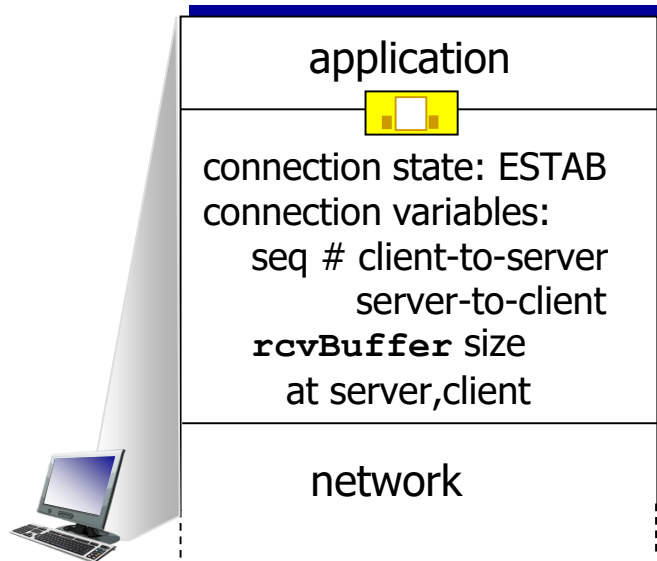
network

```
Socket clientSocket =
   newSocket("hostname","port number");
```
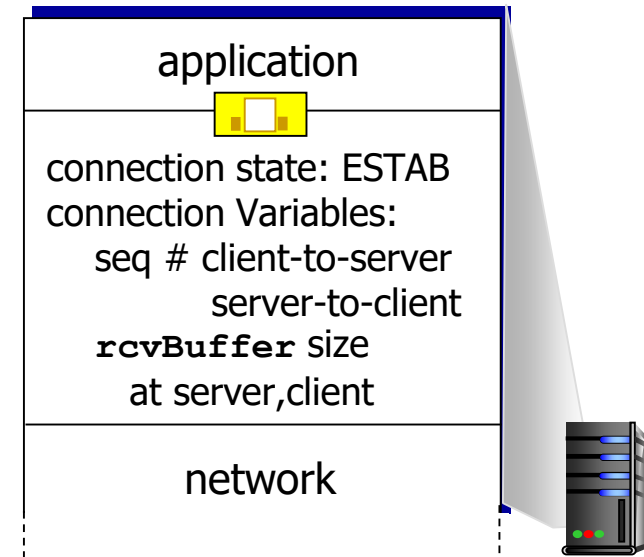
```
Socket connectionSocket =
   welcomeSocket.accept();
```

Agreeing to establish a connection

2-way handshake:

Let's talk

OK

ESTAB

ESTAB

choose x

req_conn(x)

acc_conn(x)

ESTAB

ESTAB
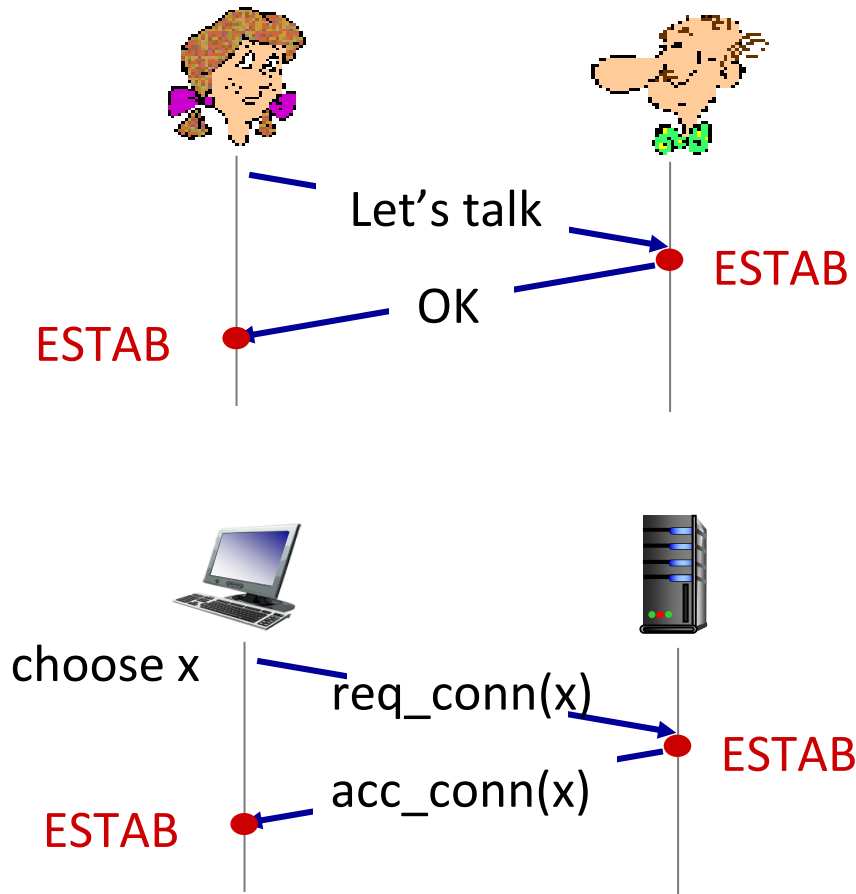
Q: will 2-way handshake always work in network?

- variable delays
- retransmitted messages (e.g. req_conn(x)) due to message loss
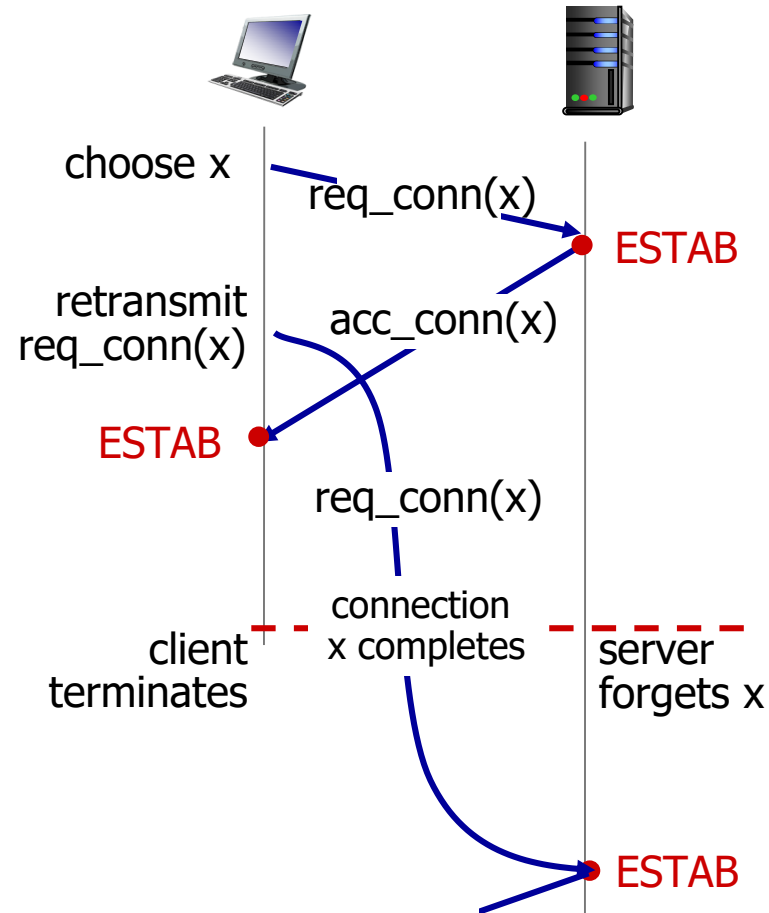- message reordering
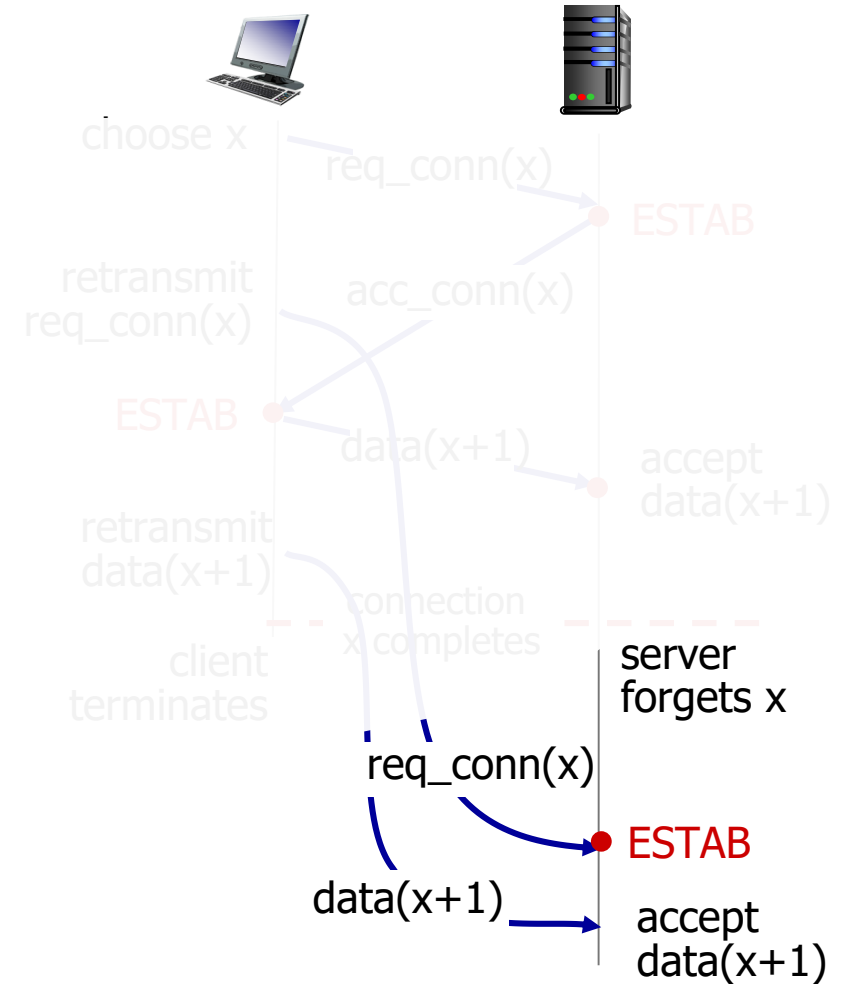- can't "see" other side

2-way handshake scenarios

choose x

req_conn(x)

ESTAB

acc_conn(x)

ESTAB

data(x+1)

accept
data(x+1)

ACK(x+1)

connection
x completes
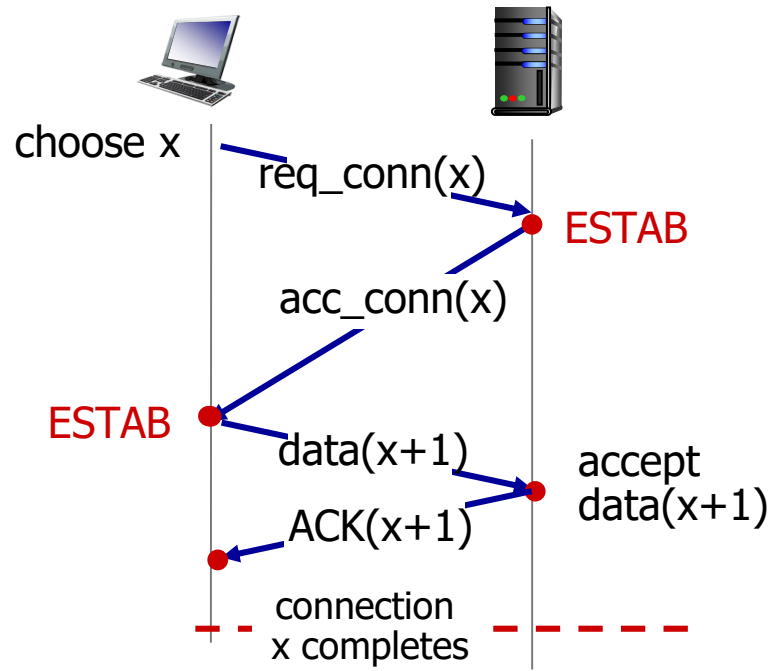
No problem!

2-way handshake scenarios

Problem: half open connection! (no client)

2-way handshake scenarios

choose x

req_conn(x)

ESTAB

retransmit
req_conn(x)

acc_conn(x)

ESTAB

data(x+1)

accept
data(x+1)

retransmit
data(x+1)

connection
x completes

client
terminates

server
forgets x

req_conn(x)

ESTAB

data(x+1)

accept
data(x+1)

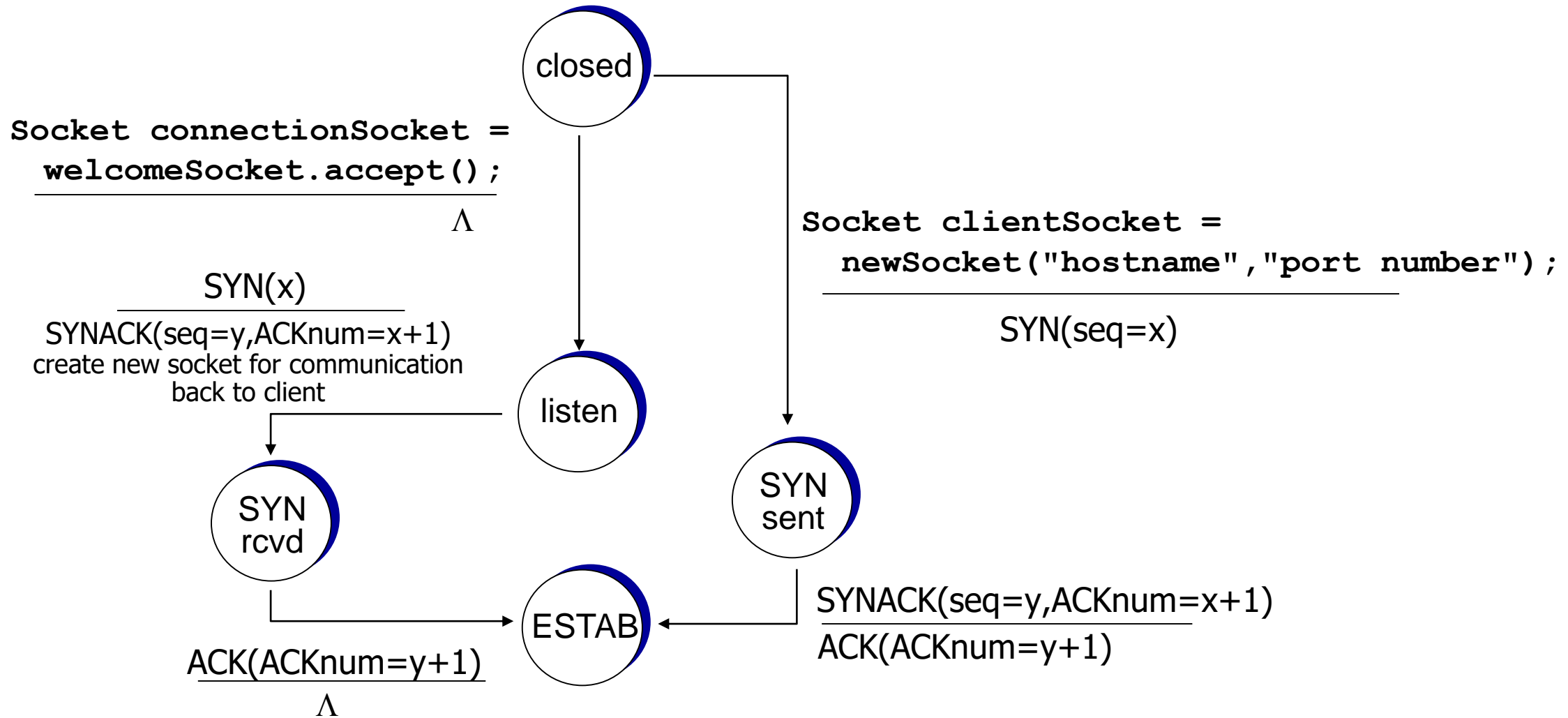Problem: dup data accepted!

TCP 3-way handshake

Server state

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

Client state

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

LISTEN

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

LISTEN

choose init seq num, x
send TCP SYN msg

SYNSENT

SYNbit=1, Seq=x

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 FSM

closed

$$\frac{\text{Socket connectionSocket =}}{\text{welcomeSocket.accept();}}$$
$$\Lambda$$

$$\frac{\text{Socket clientSocket =}}{\text{newSocket("hostname","port number");}}$$
$$\text{SYN(seq=x)}$$

$$\frac{\text{SYN(x)}}{\substack{\text{SYNACK(seq=y,ACKnum=x+1)}\\\text{create new socket for communication}\\\text{back to client}}}$$

listen

SYN
rcvd

SYN
sent

ESTAB

$$\frac{\text{ACK(ACKnum=y+1)}}{\Lambda}$$

$$\frac{\text{SYNACK(seq=y,ACKnum=x+1)}}{\text{ACK(ACKnum=y+1)}}$$

A human 3-way handshake protocol

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 (FINACK)
  - Waits around a bit (to respond to any retransmitted FIN messages before timing out) and then closes the connection

- simultaneous FIN exchanges can be handled

Closing a TCP connection
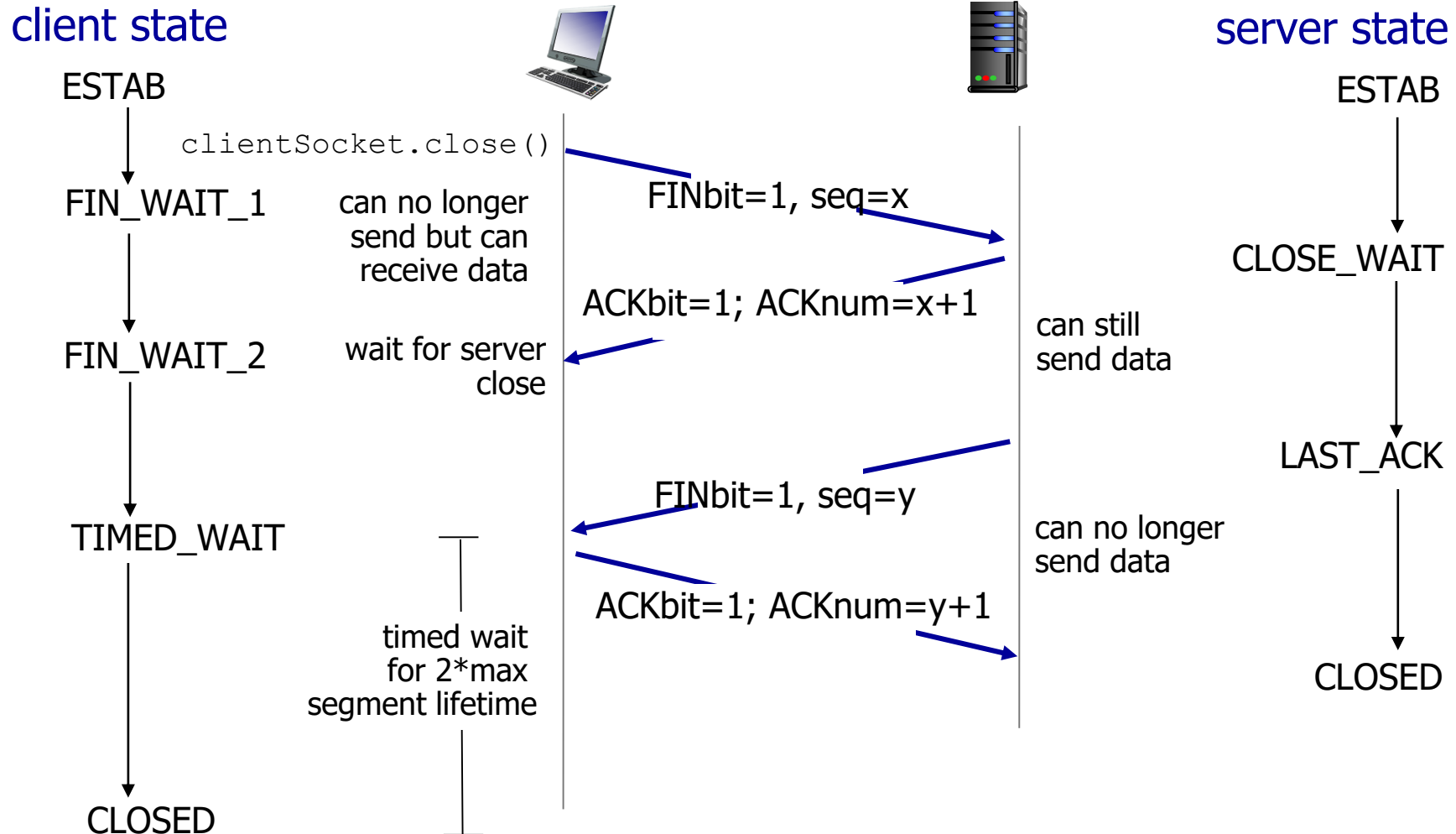
client state

ESTAB

FIN_WAIT_1

FIN_WAIT_2

TIMED_WAIT

CLOSED

clientSocket.close()

can no longer send but can receive data

wait for server close

timed wait for 2*max segment lifetime

FINbit=1, seq=x

ACKbit=1; ACKnum=x+1

FINbit=1, seq=y

ACKbit=1; ACKnum=y+1

server state

ESTAB

CLOSE_WAIT

LAST_ACK

CLOSED

can still send data

can no longer send data

Chapter 3: roadmap

- Transport-layer services
- Multiplexing and demultiplexing
- Connectionless transport: UDP
- Principles of reliable data transfer
- Connection-oriented transport: TCP
- **Principles of congestion control** (This topic will not be covered)
- TCP congestion control
- Evolution of transport-layer functionality

Chapter 3: roadmap

- Transport-layer services
- Multiplexing and demultiplexing
- Connectionless transport: UDP
- Principles of reliable data transfer
- Connection-oriented transport: TCP
- Principles of congestion control
- **TCP congestion control**
- Evolution of transport-layer functionality

TCP congestion control: AIMD

- *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 (MSS) every RTT until loss detected

Multiplicative Decrease

cut sending rate in half at each loss event
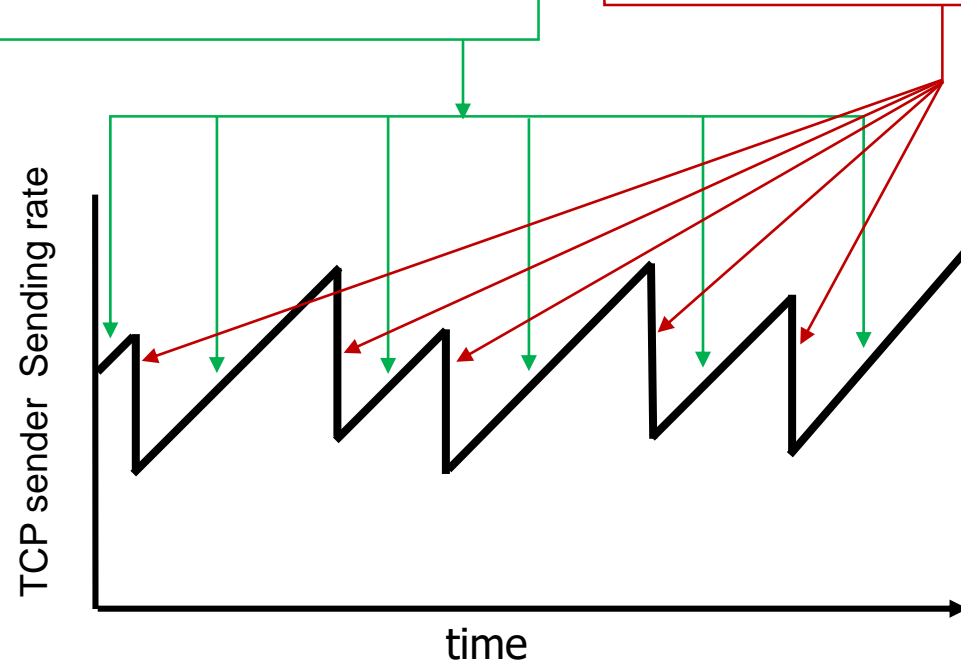
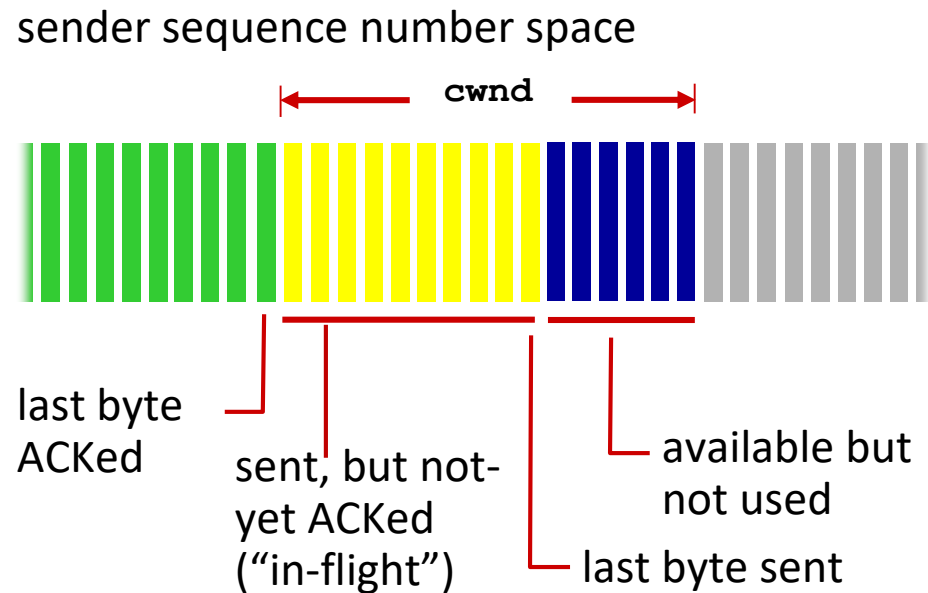**AIMD** sawtooth behavior: *probing* for bandwidth

TCP sender Sending rate

time

TCP AIMD: more

Multiplicative decrease detail: sending rate is

- Cut in half on loss detected by triple duplicate ACK
- Cut to 1 MSS (maximum segment size) when loss detected by timeout

Why AIMD?

- AIMD – a distributed, asynchronous-optimization algorithm – has been shown to:
  - optimize congested flow rates network wide!
  - have desirable stability properties

TCP congestion control: details

sender sequence number space

last byte
ACKed

sent, but not-
yet ACKed
("in-flight")

last byte sent

available but
not used

TCP sending behavior:

- *roughly:* send `cwnd` bytes, wait RTT for ACKS, then send more bytes

TCP rate $\approx \dfrac{\texttt{cwnd}}{\text{RTT}}$ bytes/sec

- TCP sender limits transmission: `LastByteSent-LastByteAcked ≤ min(cwnd, rwnd)`
- `cwnd` is dynamically adjusted in response to observed network congestion (implementing TCP congestion control)

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

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

Host A   Host B

RTT

one segment

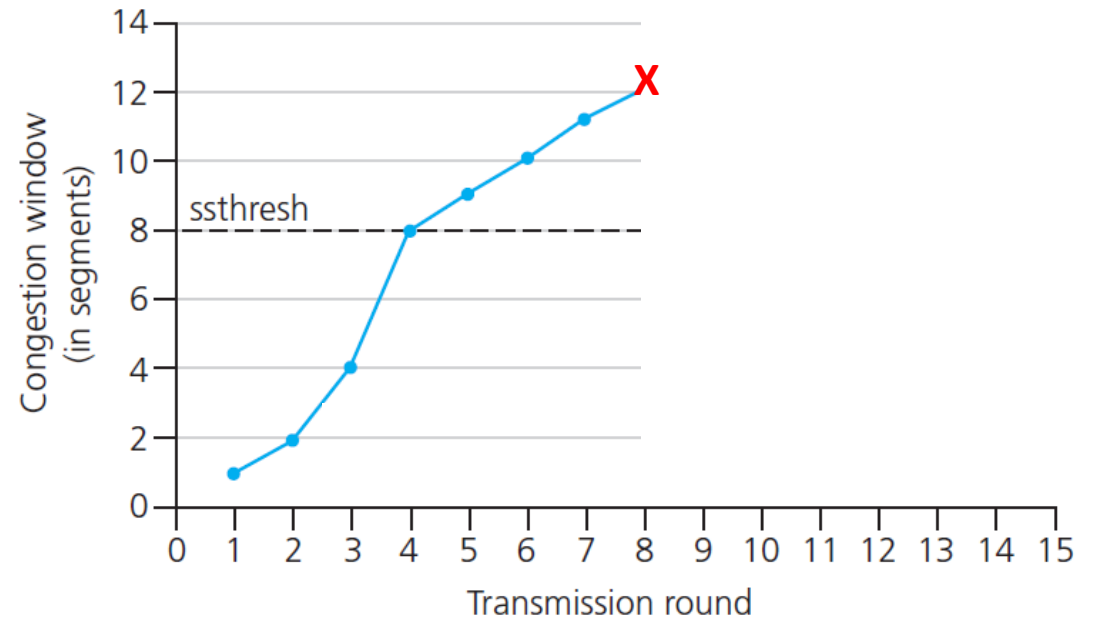two segments

four segments
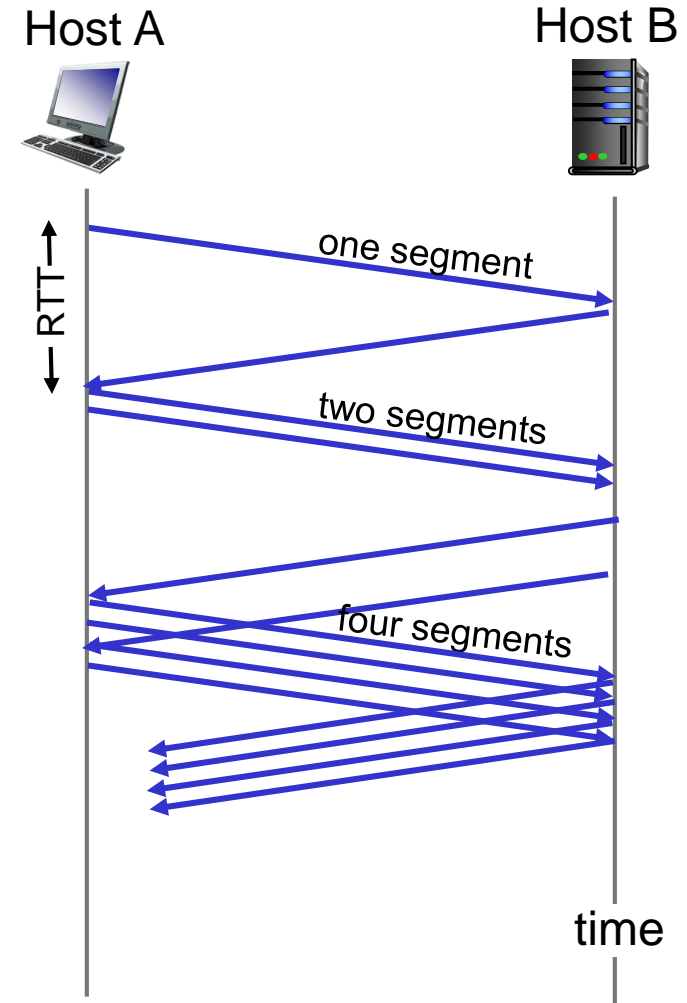
time

TCP: from slow start to congestion avoidance

Q: when should the exponential increase switch to linear?

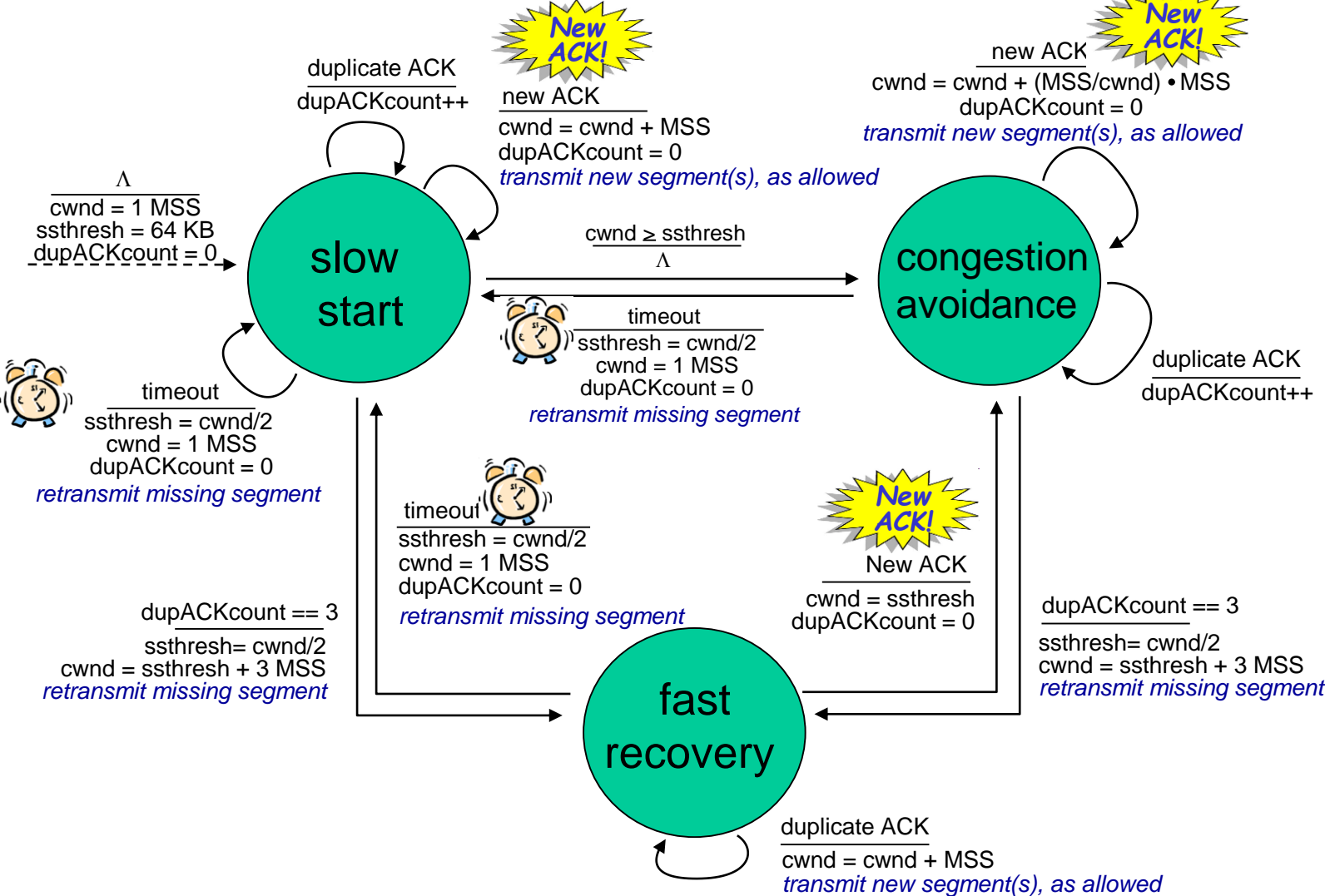*A:* when `cwnd` gets to 1/2 of its value before timeout.

Implementation:

- variable `ssthresh` (slow start threshold)
- on loss event, `ssthresh` is set to 1/2 of `cwnd` just before loss event

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

Summary: TCP congestion control

slow start

Λ
cwnd = 1 MSS
ssthresh = 64 KB
dupACKcount = 0

duplicate ACK
―――――――――
dupACKcount++

New ACK!
new ACK
―――――――――
cwnd = cwnd + MSS
dupACKcount = 0
transmit new segment(s), as allowed

timeout
―――――――――
ssthresh = cwnd/2
cwnd = 1 MSS
dupACKcount = 0
retransmit missing segment

cwnd ≥ ssthresh
―――――――――
Λ

timeout
―――――――――
ssthresh = cwnd/2
cwnd = 1 MSS
dupACKcount = 0
retransmit missing segment

congestion avoidance

New ACK!
new ACK
―――――――――
cwnd = cwnd + (MSS/cwnd) • MSS
dupACKcount = 0
transmit new segment(s), as allowed

i.e., increases the value of
`cwnd` by just a single
MSS every RTT

duplicate ACK
―――――――――
dupACKcount++

dupACKcount == 3
―――――――――
ssthresh = cwnd/2
cwnd = ssthresh + 3 MSS
retransmit missing segment

timeout
―――――――――
ssthresh = cwnd/2
cwnd = 1 MSS
dupACKcount = 0
retransmit missing segment

New ACK!
New ACK
―――――――――
cwnd = ssthresh
dupACKcount = 0

dupACKcount == 3
―――――――――
ssthresh= cwnd/2
cwnd = ssthresh + 3 MSS
retransmit missing segment

fast recovery

duplicate ACK
―――――――――
cwnd = cwnd + MSS
transmit new segment(s), as allowed

Evolution of TCP's congestion window

Fast recovery is a recommended, but not required, component of TCP

- **TCP Tahoe** (an early version of TCP) unconditionally cut its congestion window to 1 MSS and entered the slow-start phase after either a timeout-indicated or triple-duplicate-ACK-indicated loss event
- **TCP Reno** (newer version of TCP) incorporated fast recovery

Example

TCP Reno

- Identify the intervals of time when TCP slow start is operating?
  - [1,4] and [24,27]

- Identify the intervals of time when TCP congestion avoidance is operating?
  - [4,8], [9,17], [18,23], and [27,29]

- After the 17th transmission round, is segment loss detected by a triple duplicate ACK or by timeout?
  - Triple duplicate ACK

- After the 23rd transmission round, is segment loss detected by a triple duplicate ACK or by timeout?
  - Timeout

- What is the initial value of Threshold at the first transmission round?
  - 8

- What is the value of Threshold at the 18th transmission round?
  - 14/2=7

- What is the value of Threshold at the 26th transmission round?
  - 12/2=6

- Assuming a packet loss is detected after the 29th round by a timeout, what will be the values of the congestion-window size and of Threshold?
  - cwnd=1 and ssthreshold= 8/2=4

- During what transmission round is the 40th segment sent?
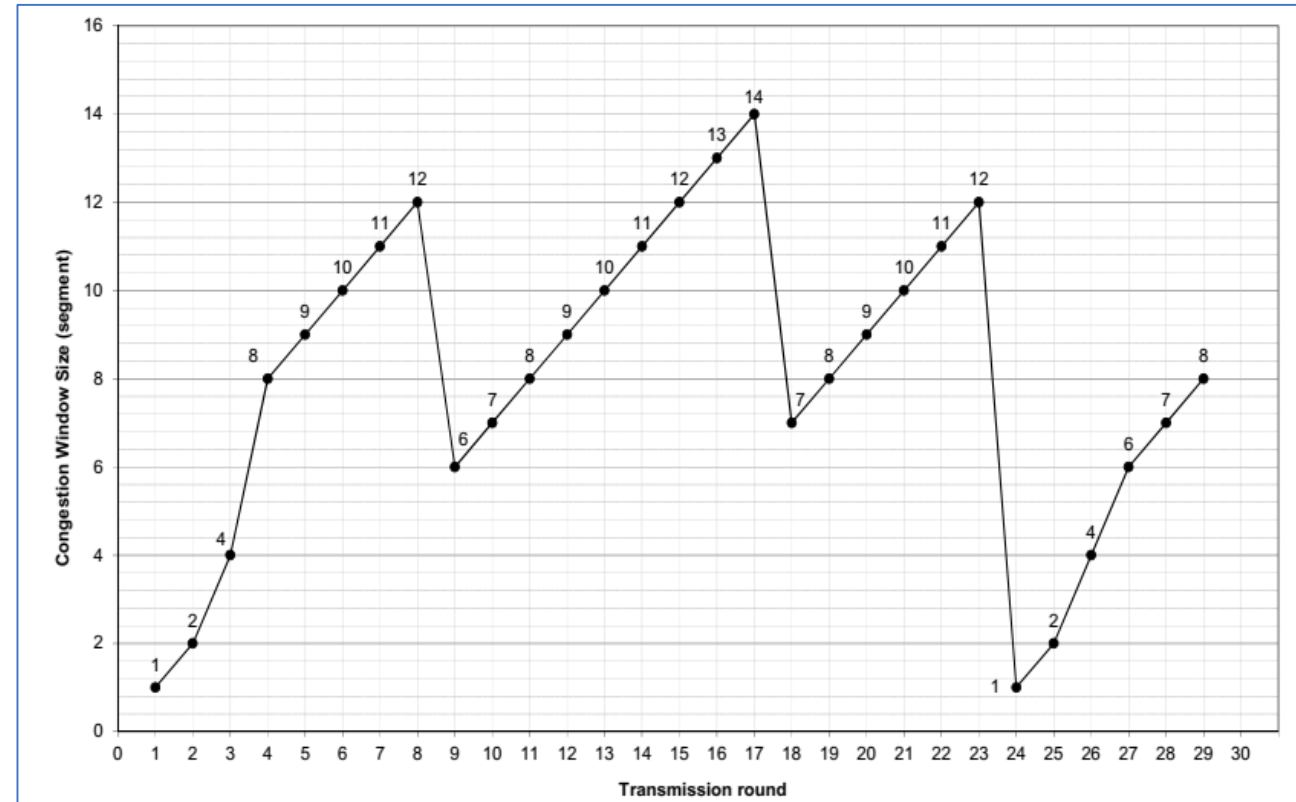

| Round	Segments	Total Segments
1	1	1
2	2	3
3	4	7
4	8	15
5	9	24
6	10	34
7	11	45

TCP CUBIC

- Is there a better way than AIMD to "probe" for usable bandwidth?
- Insight/intuition:
  - W_{max}: sending rate at which congestion loss was detected
  - congestion state of bottleneck link probably (?) hasn't changed much
  - after cutting rate/window in half on loss, initially ramp to to W_{max} *faster*, but then approach W_{max} more *slowly*


— classic TCP

- - - TCP CUBIC - higher throughput in this example
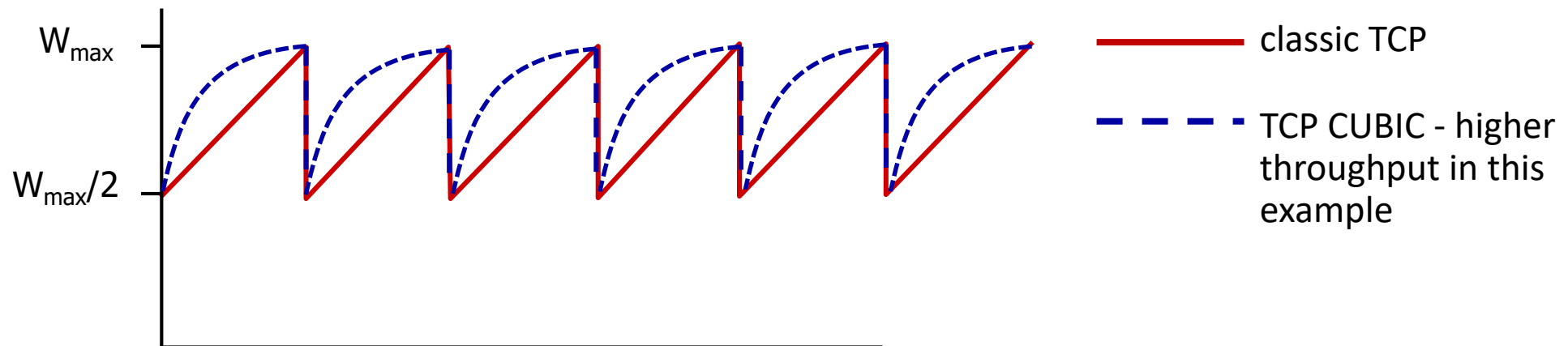
TCP throughput

- avg. TCP thruput as function of window size, RTT?
  - ignore slow start, assume there is always data to send

- W: window size (measured in bytes) where loss occurs
  - avg. window size (# in-flight bytes) is ¾ W
  - avg. thruput is 3/4W per RTT

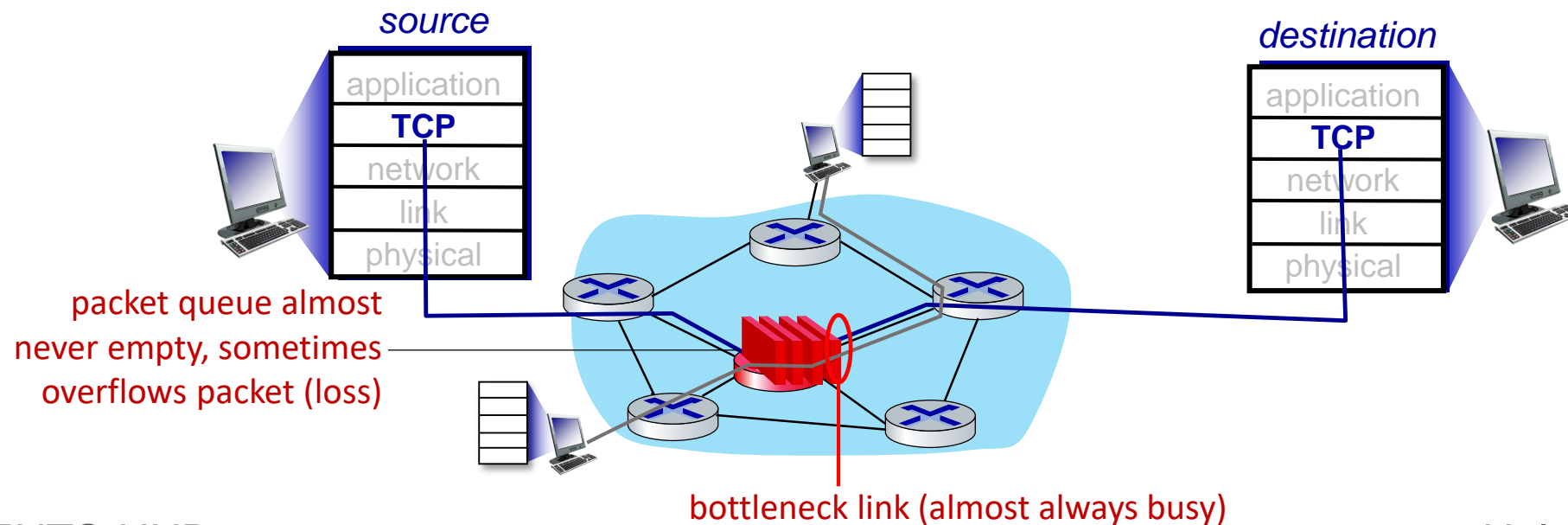avg TCP thruput = $\frac{3}{4} \frac{W}{RTT}$ bytes/sec
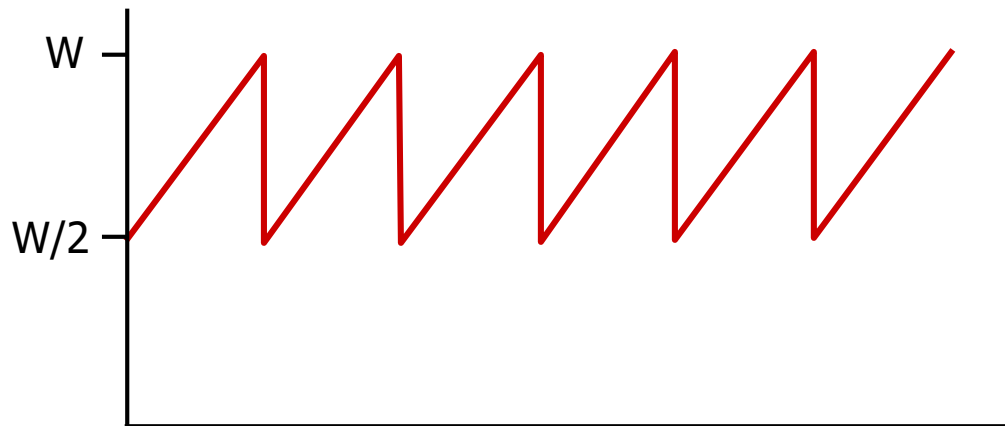
TCP CUBIC

- ▪ K: point in time when TCP window size will reach W_{max}
  - K itself is tunable
- ▪ increase W as a function of the *cube* of the distance between current time and K
  - larger increases when further away from K
  - smaller increases (cautious) when nearer K
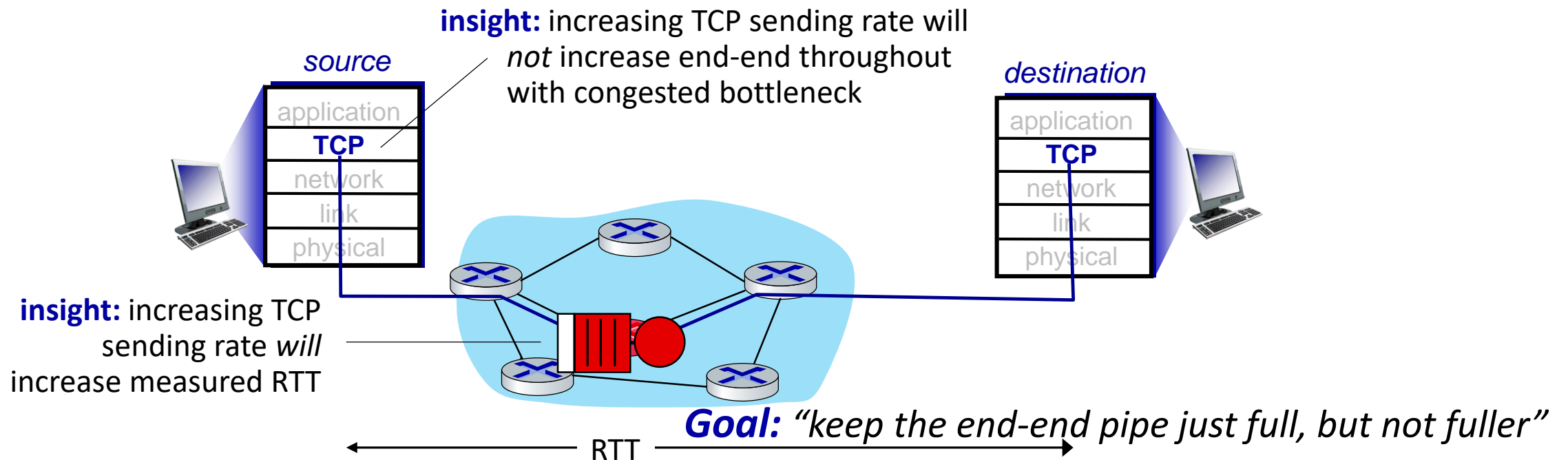- ▪ TCP CUBIC default in Linux, most popular TCP for popular Web servers
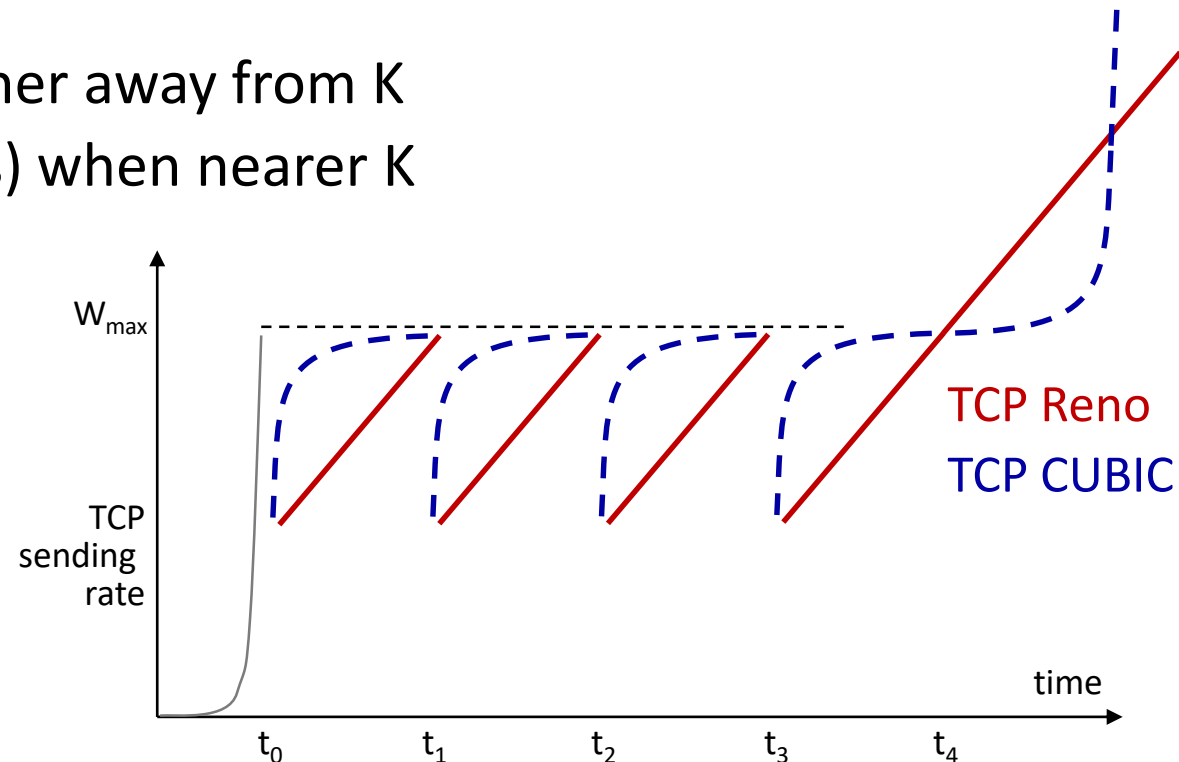
W_{max}

TCP sending rate

TCP Reno

TCP CUBIC

time

t_0 t_1 t_2 t_3 t_4

TCP and the congested "bottleneck link"

- TCP (classic, CUBIC) increase TCP's sending rate until packet loss occurs at some router's output: the *bottleneck link*

source

application
TCP
network
link
physical

destination

application
TCP
network
link
physical

packet queue almost never empty, sometimes overflows packet (loss)

bottleneck link (almost always busy)

TCP and the congested "bottleneck link"

- TCP (classic, CUBIC) increase TCP's sending rate until packet loss occurs at some router's output: the *bottleneck link*

- understanding congestion: useful to focus on congested bottleneck link

insight: increasing TCP sending rate will *not* increase end-end throughout with congested bottleneck

source

destination

application
TCP
network
link
physical

application
TCP
network
link
physical

insight: increasing TCP sending rate *will* increase measured RTT

Goal: "keep the end-end pipe just full, but not fuller"

RTT

Delay-based TCP congestion control

Keeping sender-to-receiver pipe "just full enough, but no fuller": keep bottleneck link busy transmitting, but avoid high delays/buffering
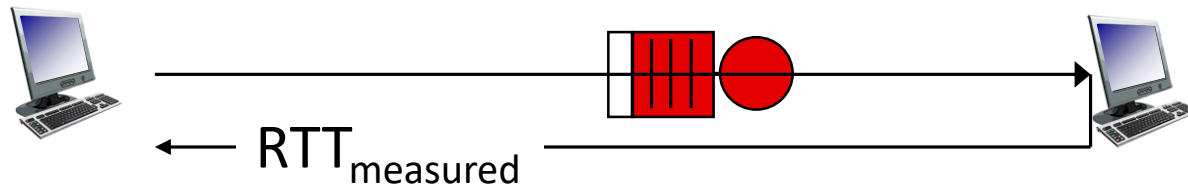
$$\text{measured throughput} = \frac{\text{\# bytes sent in last RTT interval}}{RTT_{measured}}$$

Delay-based approach:

- RTT_{min} - minimum observed RTT (uncongested path)
- uncongested throughput with congestion window `cwnd` is cwnd/RTT_{min}

```
if measured throughput "very close" to  uncongested throughput
     increase cwnd linearly          /* since path not congested */
else if measured throughput "far below" uncongested throughout
     decrease cwnd  linearly          /* since path is congested */
```
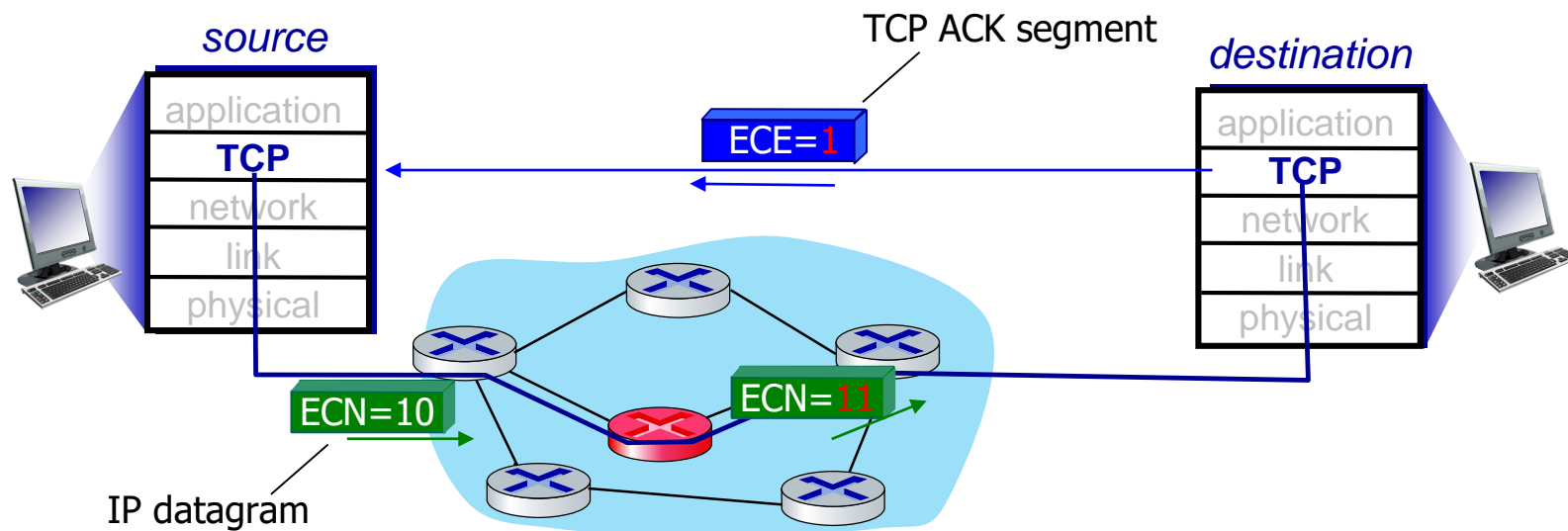
Delay-based TCP congestion control

- congestion control without inducing/forcing loss

- maximizing throughout ("keeping the just pipe full… ") while keeping delay low ("…but not fuller")

- a number of deployed TCPs take a delay-based approach
  - Bottleneck Bandwidth and Round-trip propagation time (BBR) congestion control algorithm deployed on Google's (internal) backbone network

Explicit congestion notification (ECN)

TCP deployments often implement *network-assisted* congestion control:

- two bits in IP header (ToS field) marked *by network router* to indicate congestion
  - *policy* to determine marking chosen by network operator
- congestion indication carried to destination
- destination sets ECE bit on ACK segment to notify sender of congestion
- involves both IP (IP header ECN bit marking) and TCP (TCP header E (a.k.a., ECE - ECN-Echo) and C (a.k.a., CWR - Congestion Window Reduced) bit marking)
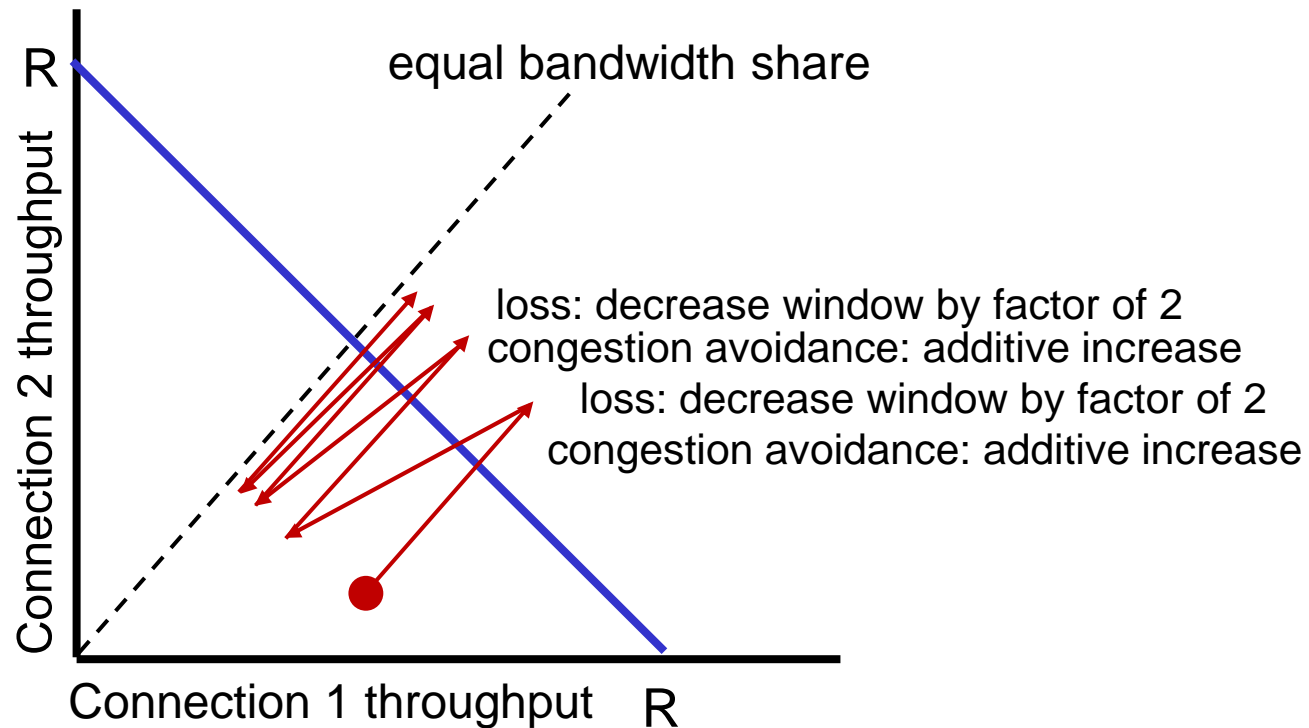
source

TCP ACK segment

destination

application
TCP
network
link
physical

ECE=1

application
TCP
network
link
physical

ECN=10

ECN=11

IP datagram

TCP fairness

Fairness goal: if *K* TCP sessions share same bottleneck link of bandwidth *R*, each should have average rate of *R/K*

TCP connection 1

TCP connection 2

bottleneck router capacity R
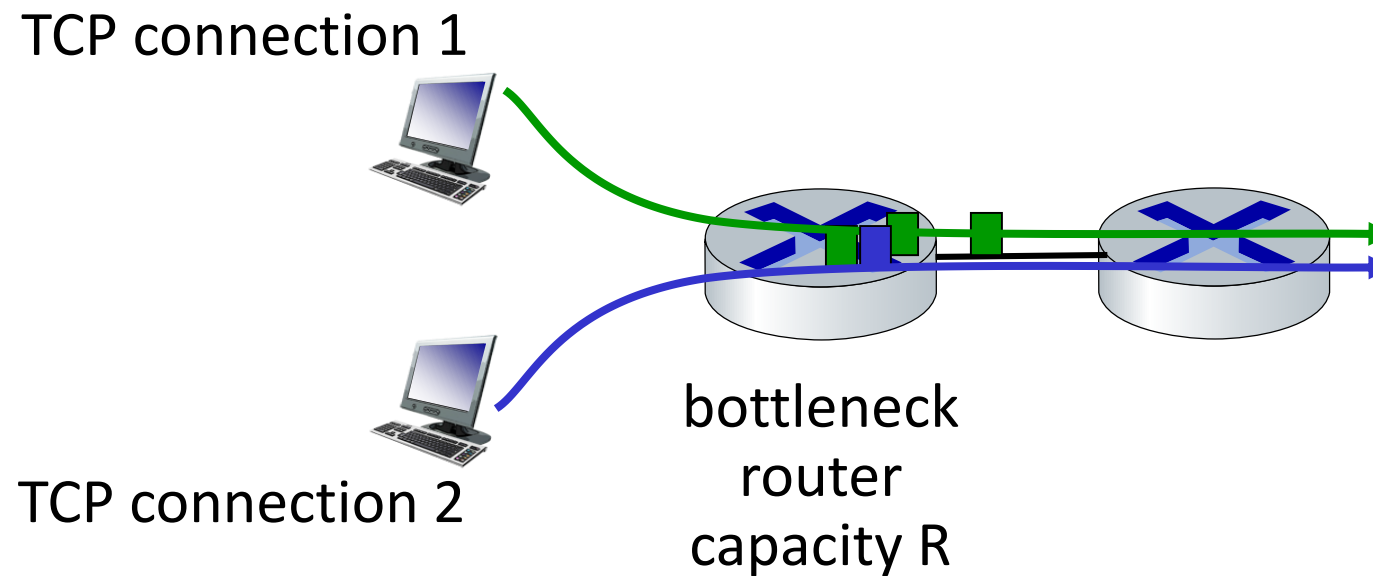
Q: is TCP Fair?

Example: two competing TCP sessions:

- additive increase gives slope of 1, as throughout increases
- multiplicative decrease decreases throughput proportionally

equal bandwidth share

Connection 2 throughput

R

loss: decrease window by factor of 2
congestion avoidance: additive increase
loss: decrease window by factor of 2
congestion avoidance: additive increase

Connection 1 throughput   R

Is TCP fair?
A: Yes, under idealized assumptions:
- same RTT
- fixed number of sessions only in congestion avoidance

Fairness: must all network apps be "fair"?

Fairness and UDP

- multimedia apps often do not use TCP
  - do not want rate throttled by congestion control
- instead use UDP:
  - send audio/video at constant rate, tolerate packet loss
- there is no "Internet police" policing use of congestion control

Fairness, parallel TCP connections

- application can open *multiple* parallel connections between two hosts
- web browsers do this , e.g., link of rate R with 9 existing connections:
  - new app asks for 1 TCP, gets rate R/10
  - new app asks for 11 TCPs, gets ≈ R/2

Transport layer: roadmap

- Transport-layer services
- Multiplexing and demultiplexing
- Connectionless transport: UDP
- Principles of reliable data transfer
- Connection-oriented transport: TCP
- Principles of congestion control
- TCP congestion control
- **Evolution of transport-layer functionality**

Evolving transport-layer functionality

- TCP, UDP: principal transport protocols for 40 years
- different "flavors" of TCP developed, for specific scenarios:

Scenario	Challenges
Long, fat pipes (large data transfers)	Many packets "in flight"; loss shuts down pipeline
Wireless networks	Loss due to noisy wireless links, mobility; TCP treat this as congestion loss
Long-delay links	Extremely long RTTs
Data center networks	Latency sensitive
Background traffic flows	Low priority, "background" TCP flows

- moving transport–layer functions to application layer, on top of UDP
  - HTTP/3: QUIC

TCP over "long, fat pipes"

- example: 1500 byte segments, 100ms RTT, want 10 Gbps throughput

- requires W = 83,333 in-flight segments

- throughput in terms of segment loss probability, L [Mathis 1997]:

$$\text{TCP throughput} = \frac{1.22 \cdot \text{MSS}}{\text{RTT} \sqrt{L}}$$

  ➡ to achieve 10 Gbps throughput, need a loss rate of L = $2.14 \cdot 10^{-10}$
    – *a very small loss rate!*

- versions of TCP for long, high-speed scenarios

QUIC: Quick UDP Internet Connections

- application-layer protocol, on top of UDP
  - increase performance of HTTP
  - deployed on many Google servers, apps (Chrome, mobile YouTube app)

|	
Application	HTTP/2
	TLS
Transport	TCP
Network	IP

HTTP/2 over TCP
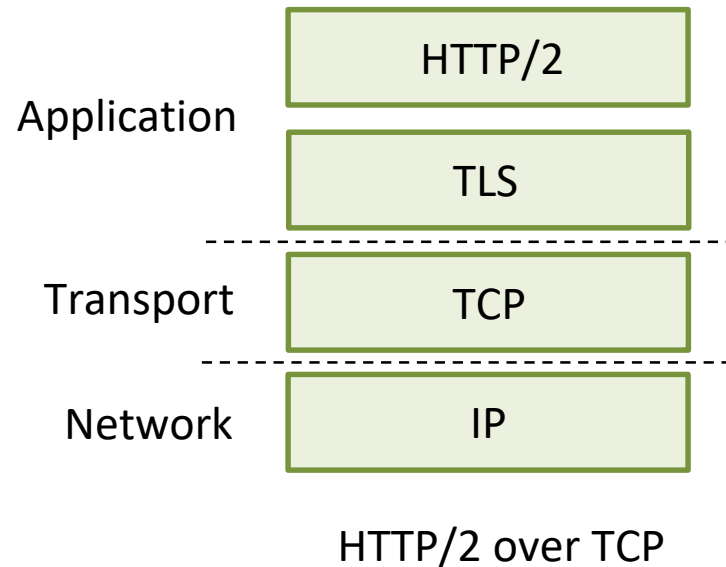
QUIC: Quick UDP Internet Connections

adopts approaches we've studied in this chapter for connection establishment, error control, congestion control

- **error and congestion control:** "Readers familiar with TCP's loss detection and congestion control will find algorithms here that parallel well-known TCP ones." [from QUIC specification]
- **connection establishment:** reliability, congestion control, authentication, encryption, state established in one RTT

■ multiple application-level "streams" multiplexed over single QUIC connection
  - separate reliable data transfer, security
  - common congestion control

QUIC: Connection establishment
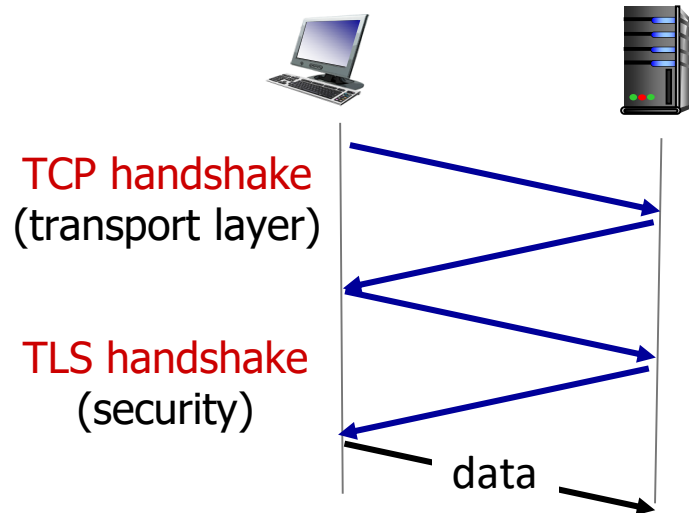
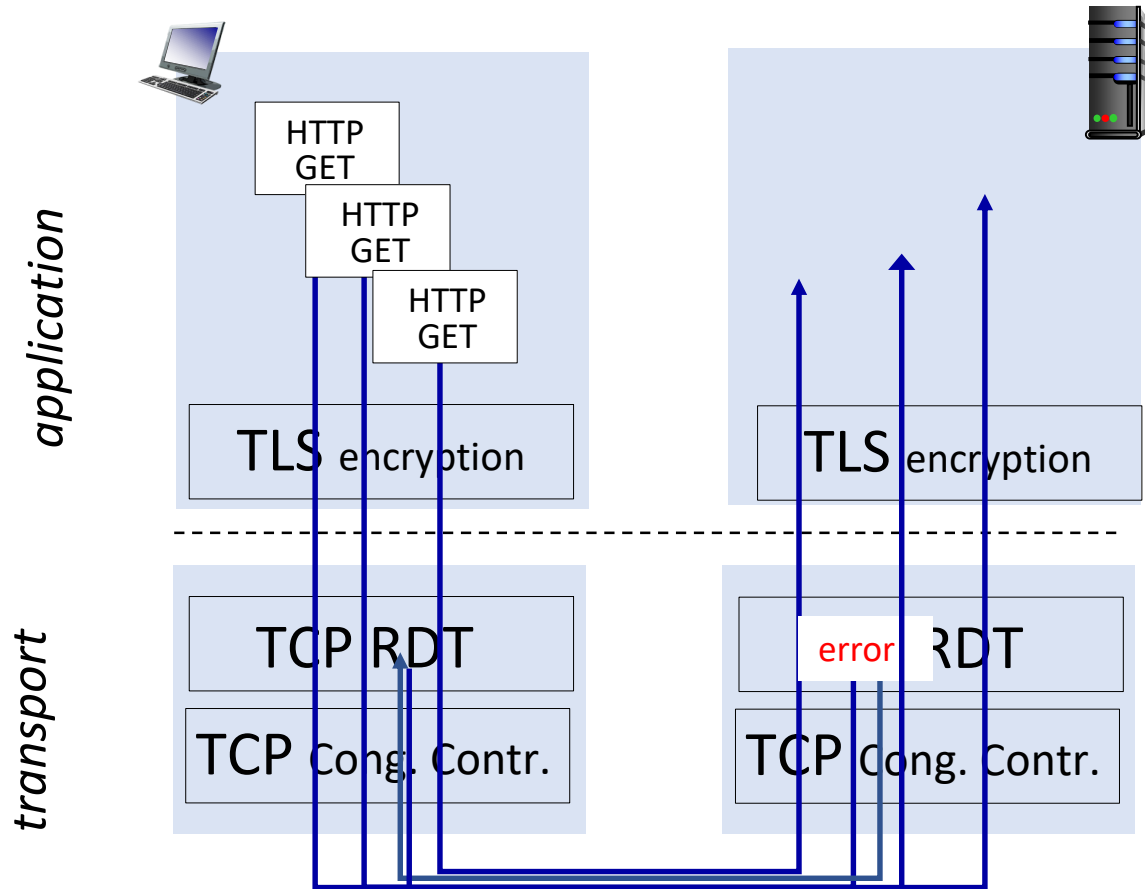TCP (reliability, congestion control state) + TLS (authentication, crypto state)

- 2 serial handshakes
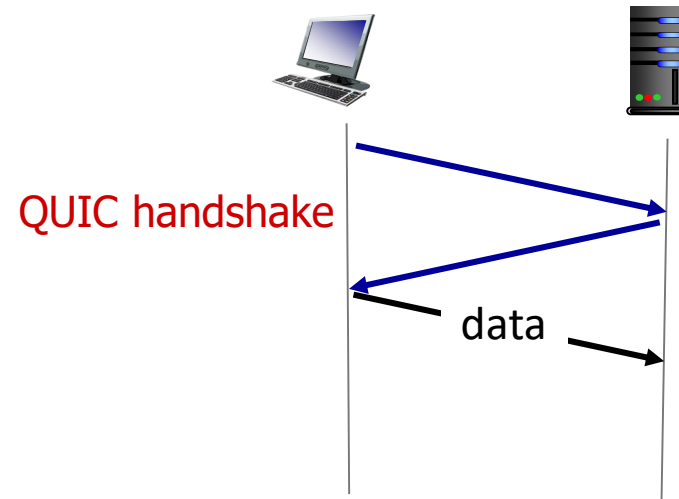
QUIC: reliability, congestion control, authentication, crypto state

- 1 handshake

QUIC: streams: parallelism, no HOL blocking

(a) HTTP 1.1

Chapter 3: summary

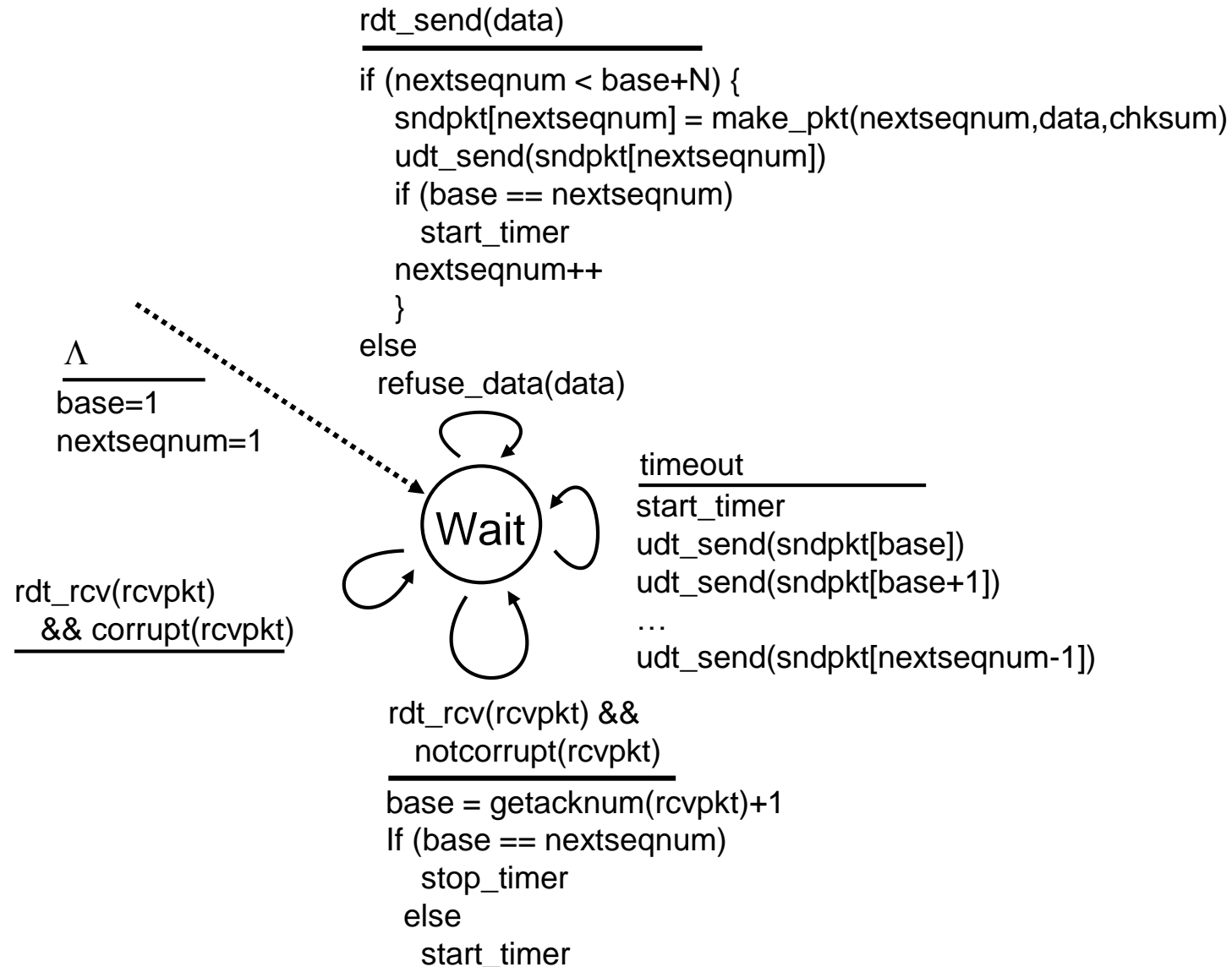- principles behind transport layer services:
  - multiplexing, demultiplexing
  - reliable data transfer
  - flow control
  - congestion control
- instantiation, implementation in the Internet
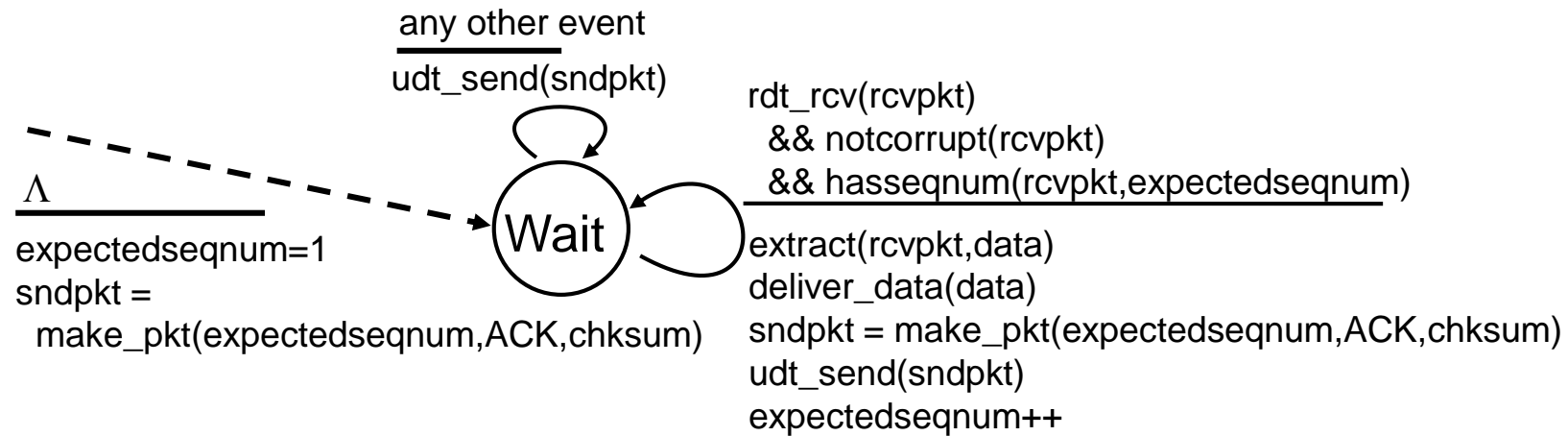 - UDP
  - TCP

Up next:

- leaving the network "edge" (application, transport layers)
- into the network "core"
- two network-layer chapters:
  - data plane
  - control plane

Additional Chapter 3 slides

Go-Back-N: sender extended FSM

rdt_send(data)
_____

if (nextseqnum < base+N) {
    sndpkt[nextseqnum] = make_pkt(nextseqnum,data,chksum)
    udt_send(sndpkt[nextseqnum])
    if (base == nextseqnum)
      start_timer
    nextseqnum++
    }
else
  refuse_data(data)

Λ
_____
base=1
nextseqnum=1

(Wait)

timeout
_____
start_timer
udt_send(sndpkt[base])
udt_send(sndpkt[base+1])
…
udt_send(sndpkt[nextseqnum-1])

rdt_rcv(rcvpkt)
  && corrupt(rcvpkt)
_____

rdt_rcv(rcvpkt) &&
  notcorrupt(rcvpkt)
_____
base = getacknum(rcvpkt)+1
If (base == nextseqnum)
  stop_timer
 else
  start_timer

Go-Back-N: receiver extended FSM

any other event
_____
udt_send(sndpkt)

rdt_rcv(rcvpkt)
&& notcorrupt(rcvpkt)
&& hasseqnum(rcvpkt,expectedseqnum)
_____

Λ
_____
expectedseqnum=1
sndpkt =
  make_pkt(expectedseqnum,ACK,chksum)

Wait

extract(rcvpkt,data)
deliver_data(data)
sndpkt = make_pkt(expectedseqnum,ACK,chksum)
udt_send(sndpkt)
expectedseqnum++

ACK-only: always send ACK for correctly-received packet with highest *in-order* seq #
- may generate duplicate ACKs
- need only remember `expectedseqnum`

▪ out-of-order packet:
- discard (don't buffer): *no receiver buffering!*
- re-ACK pkt with highest in-order seq #