Unit 3: Transport (UDP, TCP)

Sources: L. Cerdà, J. Rexford, ISOC, wikipedia, etc.

IP Protocol Stack: Key Abstractions

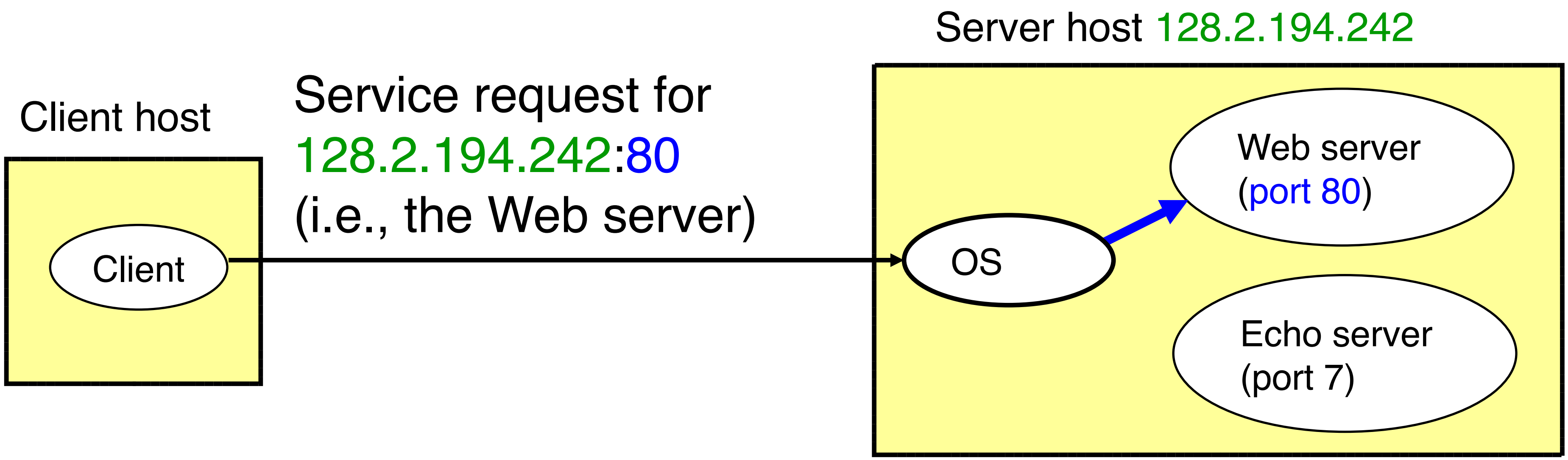
- Transport layer is where we “pay the piper”
  - Provide applications with good abstractions
  - Without support or feedback from the network

Transport Protocols

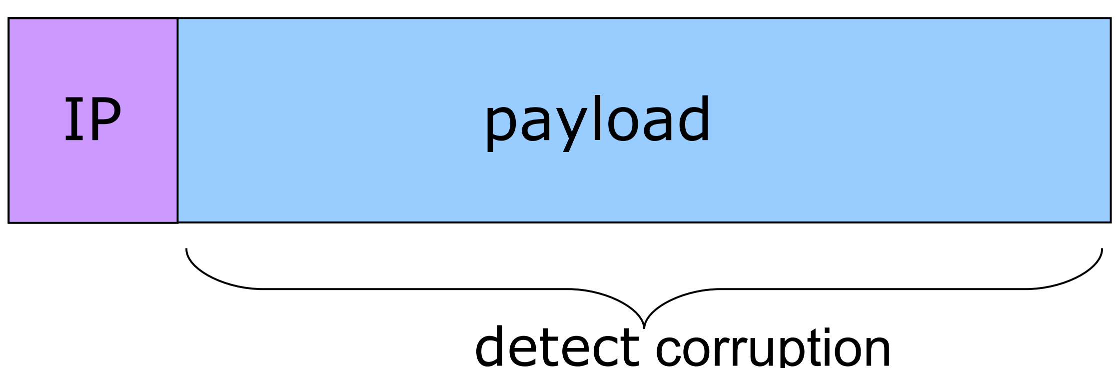
- Logical communication between processes
  - Sender divides a message into segments
  - Receiver reassembles segments into message
- Transport services
  - (De)multiplexing packets
  - Detecting corrupted data
  - Optionally: reliable delivery, flow control, …

Two Basic Transport Features

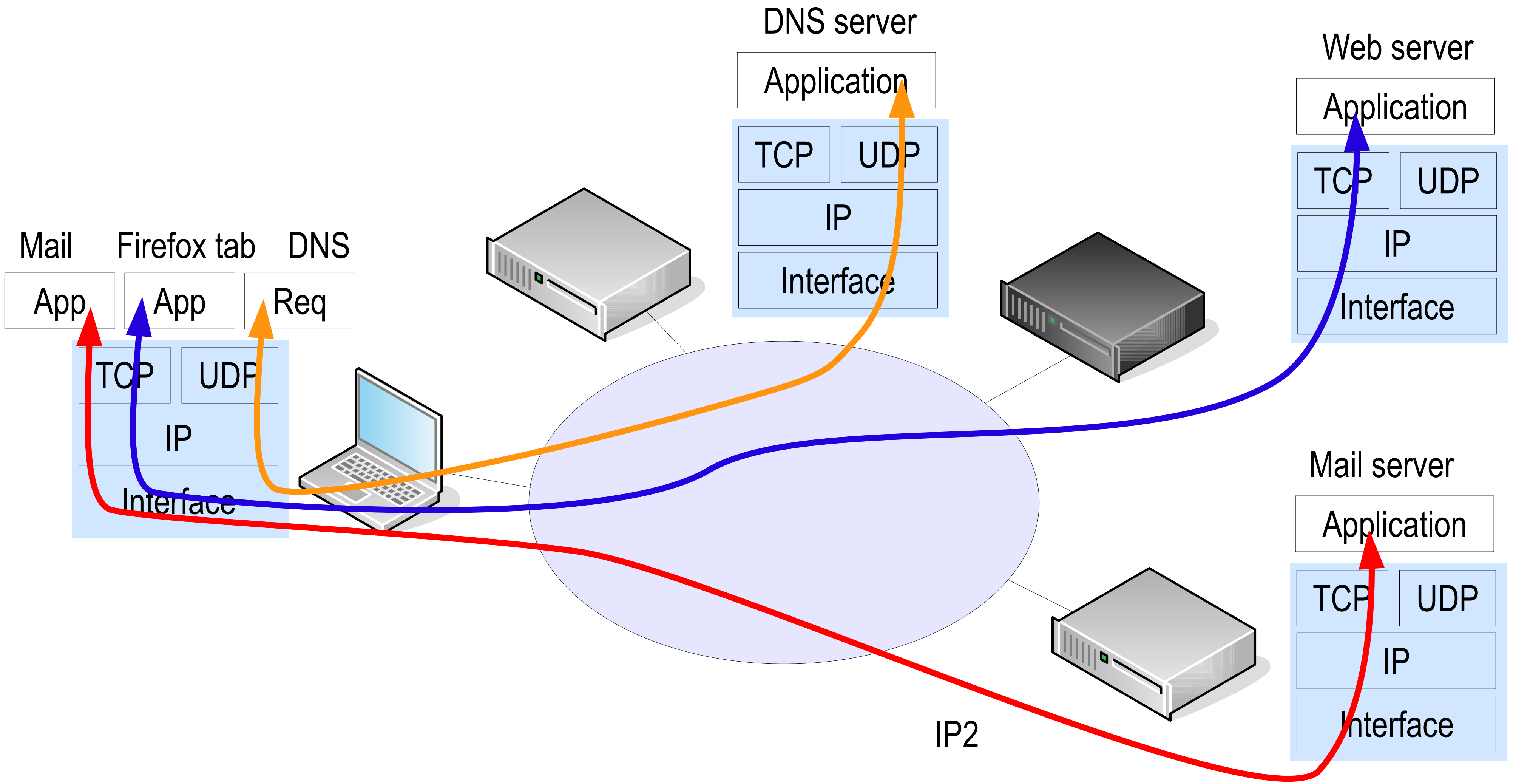
- **Demultiplexing:** port numbers



- **Error detection:** checksums

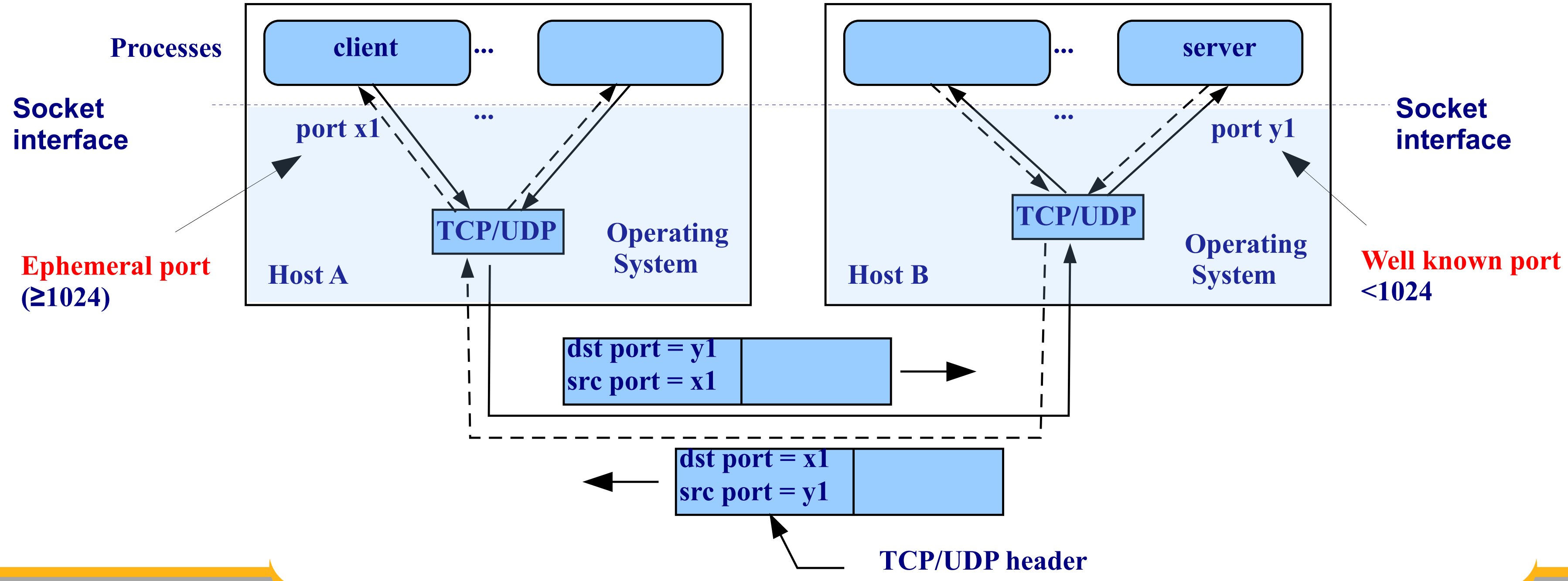


Application processes



Client-server in detail

- Client always initiates connection to an IP address, and *well known* port (<1024), in the TCP/UDP header
- Well known ports agreed by IANA/ICANN as assigned numbers /etc/services
- Server: *daemon* waiting for client requests



Port numbers

- 16 bits: 0 – 65535
- Identifies an application process
- 0-1023: well know TCP/IP ports (servers)
- 1024-65535: ephemeral ports (clients)
- Packets: source, destination

Services, well known ports

- FTP 20, 21
- SSH 22
- Telnet 23
- SMTP 25
- DNS 53
- DHCP (BOOTP) 67, 68
- HTTP 80
- POP3 110
- IMAP 143
- HTTPS 443
- RIP 520

```
rtmp              1/ddp    #Routing Table Maintenance P
tcpmux            1/udp     # TCP Port Service Multiple
tcpmux            1/tcp     # TCP Port Service Multiple
nbp               2/ddp    #Name Binding Protocol
compressnet       2/udp     # Management Utility
compressnet       2/tcp     # Management Utility
compressnet       3/udp     # Compression Process
compressnet       3/tcp     # Compression Process
echo              4/ddp    #AppleTalk Echo Protocol
rje               5/udp     # Remote Job Entry
rje               5/tcp     # Remote Job Entry
zip               6/ddp    #Zone Information Protocol
echo              7/udp     # Echo
echo              7/tcp     # Echo
discard           9/udp     # Discard
discard           9/tcp     # Discard
systat           11/udp     # Active Users
systat           11/tcp     # Active Users
daytime          13/udp     # Daytime (RFC 867)
daytime          13/tcp     # Daytime (RFC 867)
qotd             17/udp     # Quote of the Day
qotd             17/tcp     # Quote of the Day
msp              18/udp     # Message Send Protocol
msp              18/tcp     # Message Send Protocol
chargen          19/udp     # Character Generator
chargen          19/tcp     # Character Generator
ftp-data         20/udp     # File Transfer [Default Da
ftp-data         20/tcp     # File Transfer [Default Da
ftp              21/udp     # File Transfer [Control]
ftp              21/tcp     # File Transfer [Control]
ssh              22/udp     # SSH Remote Login Protocol
ssh              22/tcp     # SSH Remote Login Protocol
telnet           23/udp     # Telnet
telnet           23/tcp     # Telnet
                 24/udp     # any private mail system
                 24/tcp     # any private mail system
smtp             25/udp     # Simple Mail Transfer
Smtp             25/tcp     # Simple Mail Transfer
```

User Datagram Protocol (UDP)

- Datagram messaging service
  - Demultiplexing: port numbers
  - Detecting corruption: checksum for <u>data integrity</u>
- Lightweight communication between processes
  - Send and receive messages
  - Avoid overhead of ordered, reliable delivery
  - When dropping Ok, better than waiting retransmission

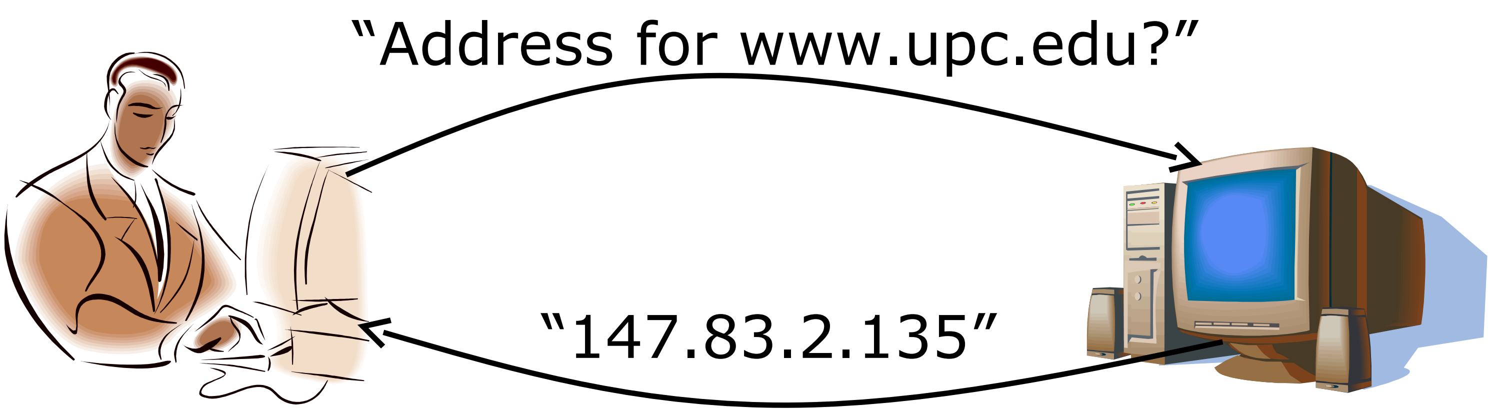
SRC port	DST port
checksum	length
DATA	

Advantages of UDP

- Fine-grain control
  - UDP sends as soon as the application writes
- No connection set-up delay
  - UDP sends without establishing a connection
- No connection state
  - No buffers, parameters, sequence #s, etc.
- Small header overhead
  - UDP header is only eight-bytes long

Popular Applications That Use UDP

- Multimedia streaming
  - Retransmitting packets is not always worthwhile
  - E.g., phone calls, video conferencing, gaming, IPTV
- Simple query-response protocols
  - Overhead of connection establishment is overkill
  - E.g., Domain Name System (DNS), DHCP, etc.



Socket API

- Client:

type: SOCK_STREAM (TCP) / SOCK_DGRAM (UDP)

```
int sock = socket(AF_INET, type, 0;
struct sockaddr_in s; s.sin_port = htons(PORT);
connect(sock, &s, sizeof(s));
send(sock, "Hello", sizeof("Hello")); Generates a UDP packet
read(sock, buffer, sizeof(buffer));    Reads a UDP packet
```

- Server:

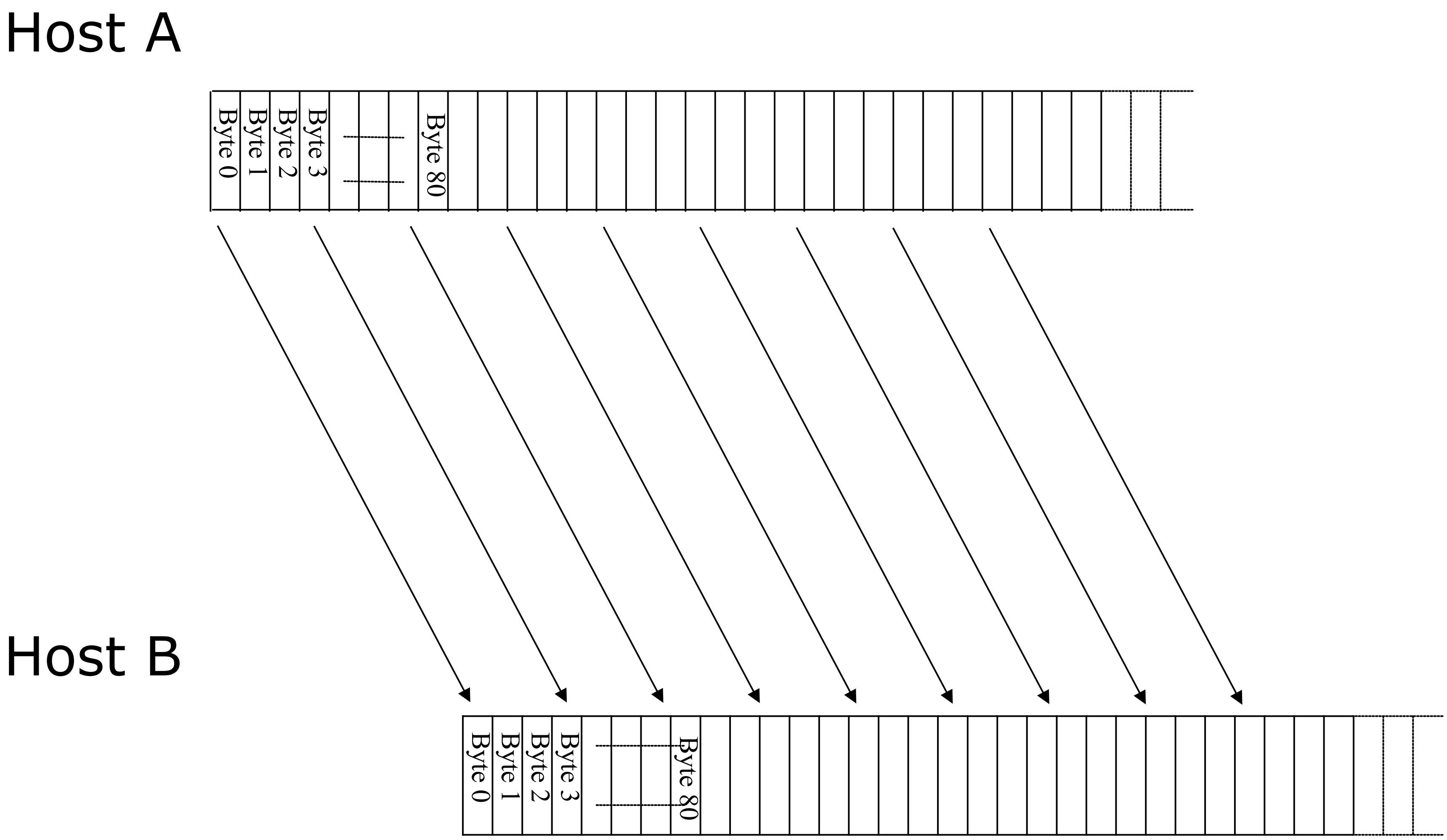
```
int sock = socket(AF_INET, type, 0)              Set option attachment
setsockopt(sock, SOL_SOCKET, SO_REUSEPORT, …)    socket to port
struct sockaddr_in s; s.sin_port = htons(PORT);  Port selection
bind(sock, &s, sizeof(s)); Attachment of socket to port
listen(sock, 4)
accept(sock, &s, sizeof(s));
read(sock, buffer, sizeof(buffer));
send(sock, "Hello", sizeof("Hello"));
```

Transmission Control Protocol (TCP)

- Stream-of-bytes service
  - Sends and receives a stream of bytes
- Reliable, in-order delivery
  - Corruption: checksums
  - Detect loss/reordering: sequence numbers
  - Reliable delivery: acknowledgments and retransmissions
- Connection oriented
  - Explicit set-up and tear-down of TCP connection
- Flow control
  - Prevent overflow of the receiver's buffer space
- Congestion control
  - Adapt to network congestion for the greater good

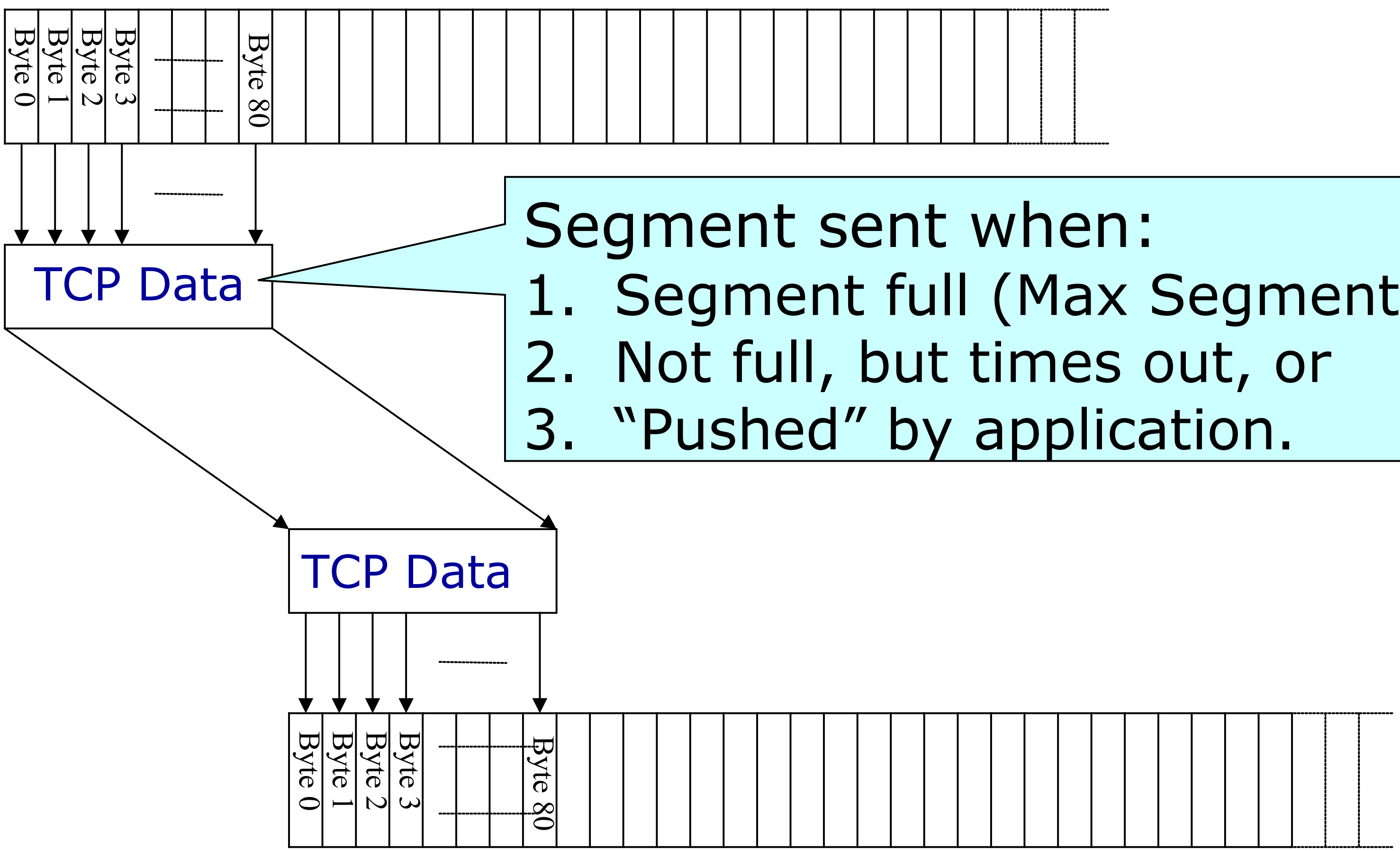
Breaking a Stream of Bytes into TCP Segments

TCP "Stream of Bytes" Service



…Emulated Using TCP "Segments"

Host A



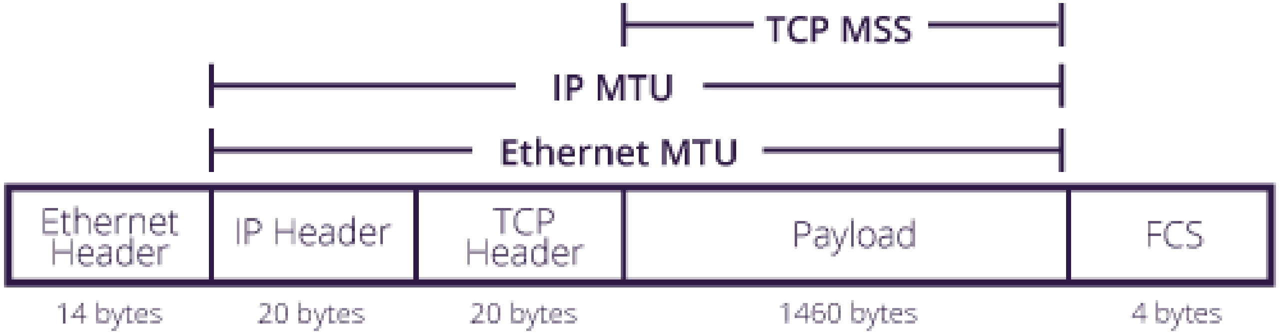
Segment sent when:

1. Segment full (Max Segment Size),
2. Not full, but times out, or
3. "Pushed" by application.

Host B

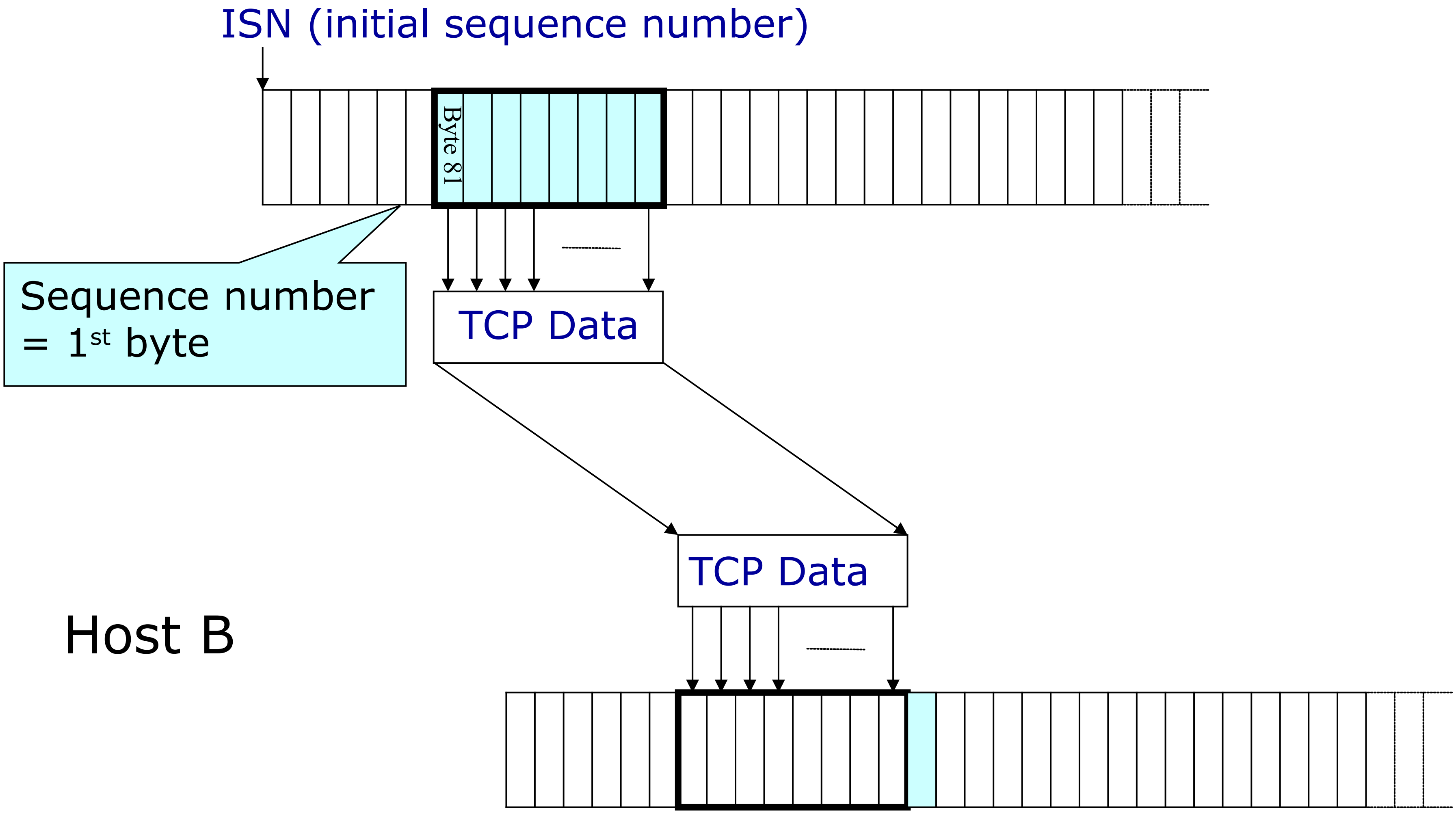
TCP Segment

- IP packet
  - No bigger than Maximum Transmission Unit (MTU)
  - E.g., up to 1500 bytes on an Ethernet link
- TCP packet
  - IP packet with a TCP header and data inside
  - TCP header is typically 20 bytes long
- TCP segment
  - No more than Maximum Segment Size (MSS) bytes
  - E.g., up to 1460 consecutive bytes from the stream



Sequence Number

Host A



Host B

Initial Sequence Number (ISN)

- Sequence number for the very first byte
  - E.g., Why not a de facto ISN of 0?
- Practical issue: reuse of port numbers
  - Port numbers must (eventually) get used again
  - … and an old packet may still be in flight
  - … and associated with the new connection
- So, TCP must change the ISN over time
  - Set from a 32-bit clock that ticks every 4 microsec
  - … which wraps around once every 4.55 hours!

Reliable Delivery on a Lossy Channel With Bit Errors

20

Challenges of Reliable Data Transfer

- Over a perfectly reliable channel
  - Easy: sender sends, and receiver receives
- Over a channel with bit errors
  - Receiver detects errors and requests retransmission
- Over a lossy channel with bit errors
  - Some data are missing, and others corrupted
  - Receiver cannot always detect loss
- Over a channel that may reorder packets
  - Receiver cannot distinguish loss from out-of-order

An Analogy

- Alice and Bob are talking
  - What if Alice couldn’t understand Bob?
  - Bob asks Alice to repeat what she said
- What if Bob hasn’t heard Alice for a while?
  - Is Alice just being quiet? Has she lost reception?
  - How long should Bob just keep on talking?
  - Maybe Alice should periodically say “uh huh”
  - … or Bob should ask “Can you hear me now?” ☺

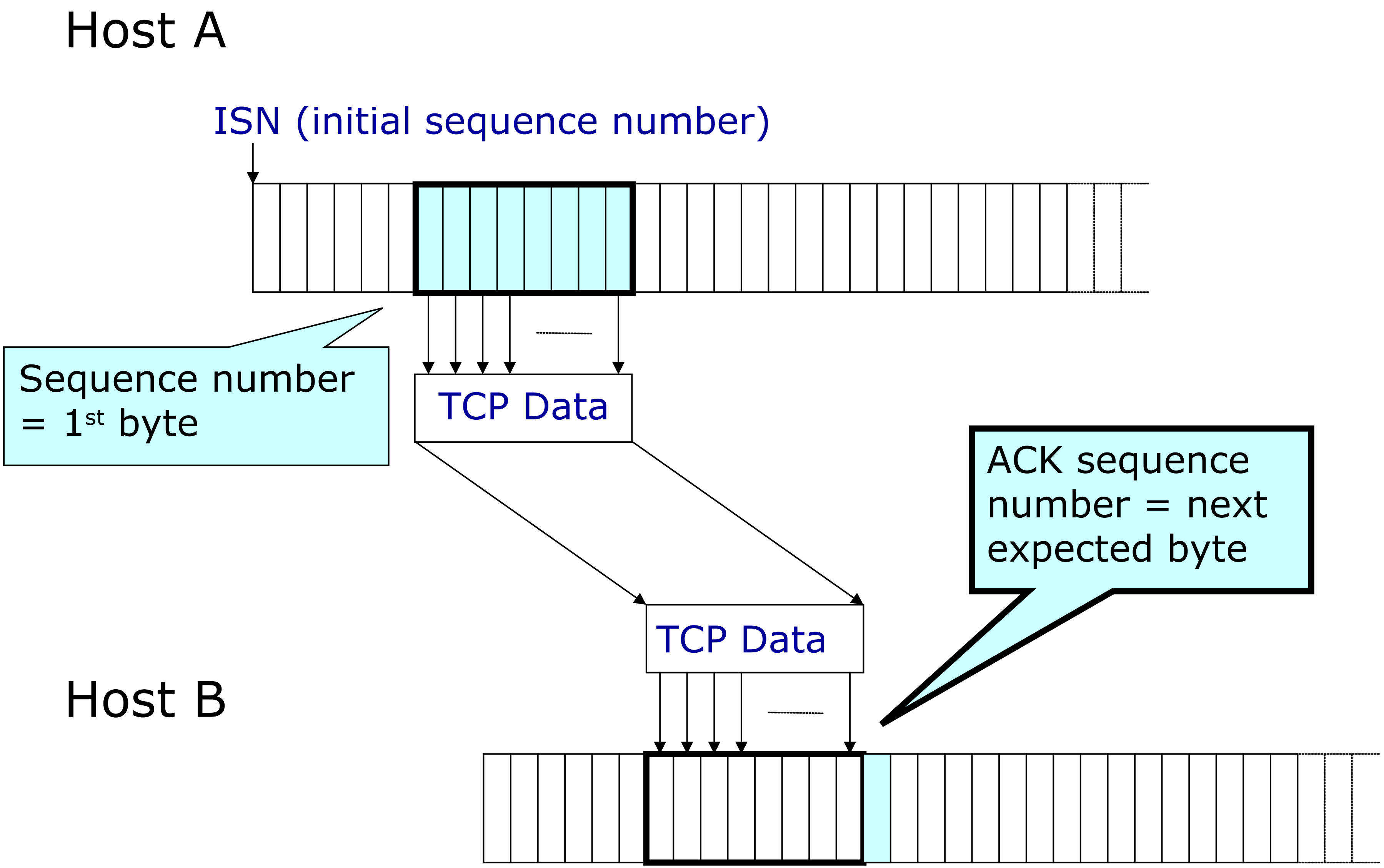
Take-Aways from the Example

- Acknowledgments from receiver
  - Positive: "okay" or "uh huh" or "ACK" or "+1"
  - Negative: "please repeat that" or "NACK"
- Retransmission by the sender
  - After *not* receiving an "ACK"
  - After receiving a "NACK"
- Timeout by the sender ("stop and wait")
  - Don't wait forever without some acknowledgment

TCP Support for Reliable Delivery

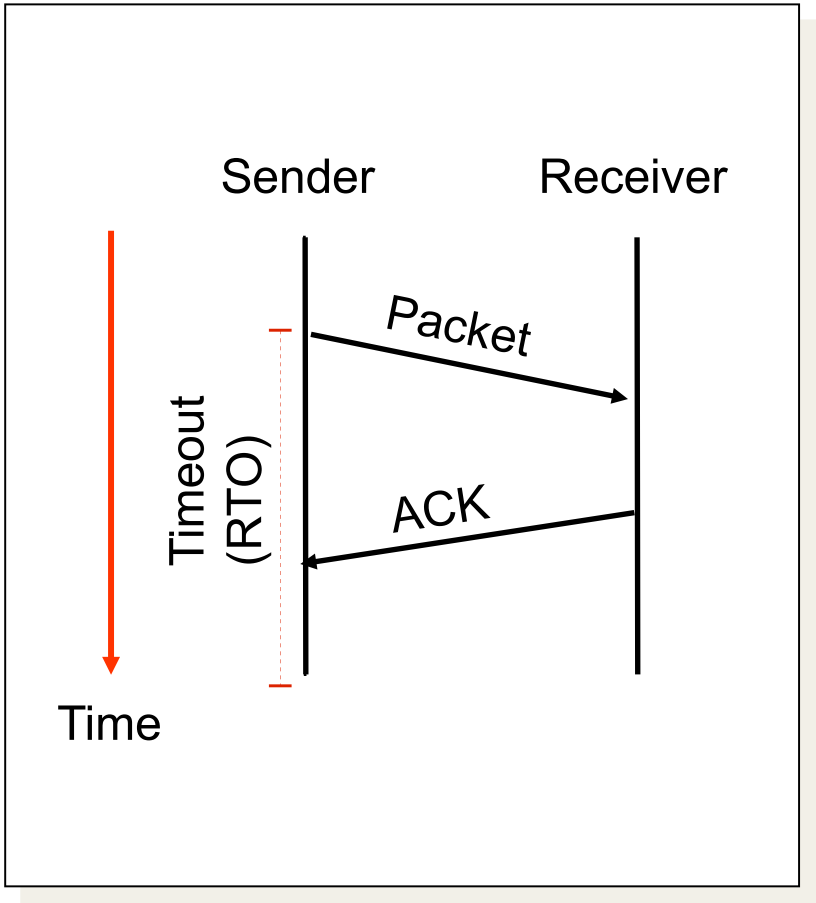
- **Detect bit errors:** checksum of data
  - Used to detect corrupted data at the receiver
  - …leading the receiver to drop the packet
- **Detect missing data:** sequence number
  - Used to detect a gap in the stream of bytes
  - ... and for putting the data back in order
- **Recover from lost data:** retransmission
  - Sender retransmits lost or corrupted data
  - Two main ways to detect lost packets

TCP Acknowledgments



Automatic Repeat reQuest (ARQ)

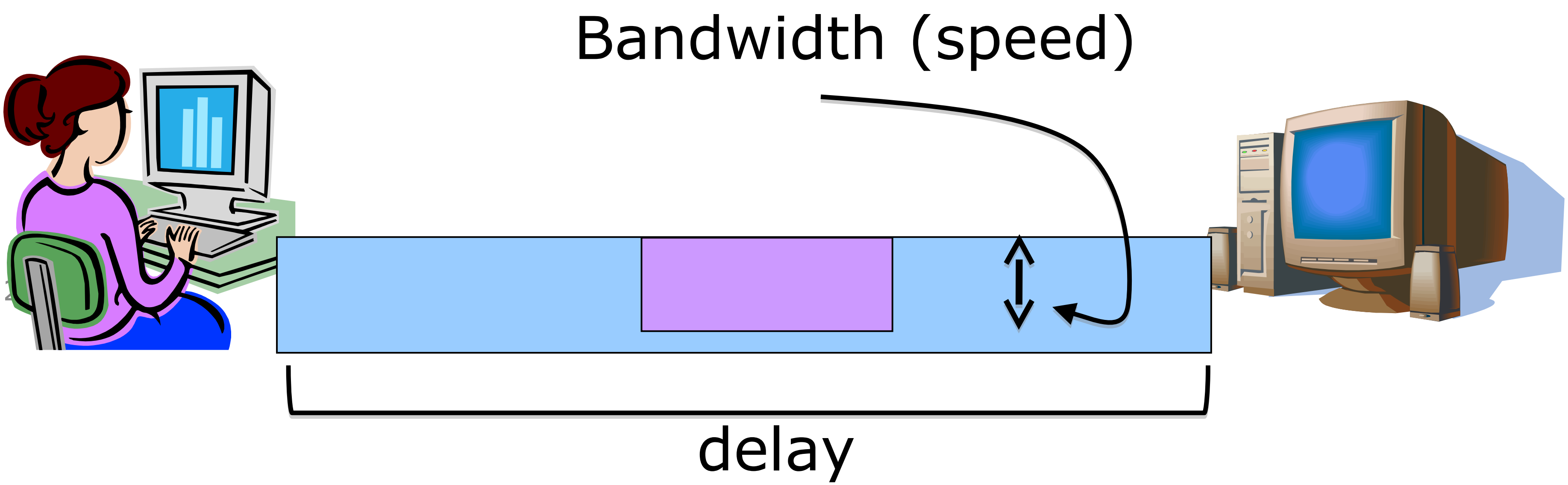
- ACK and timeouts
  - Receiver sends ACK when it receives packet
  - Sender waits for ACK and times out
- Simplest ARQ protocol
  - Stop and wait
  - Send a packet, stop and wait until ACK arrives

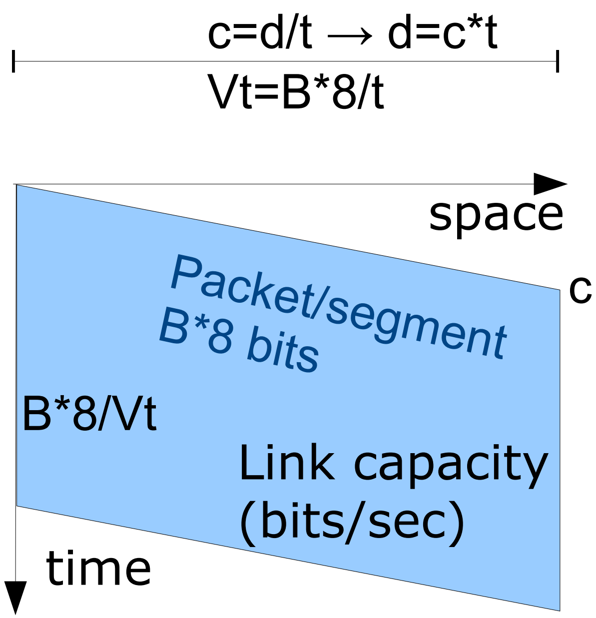


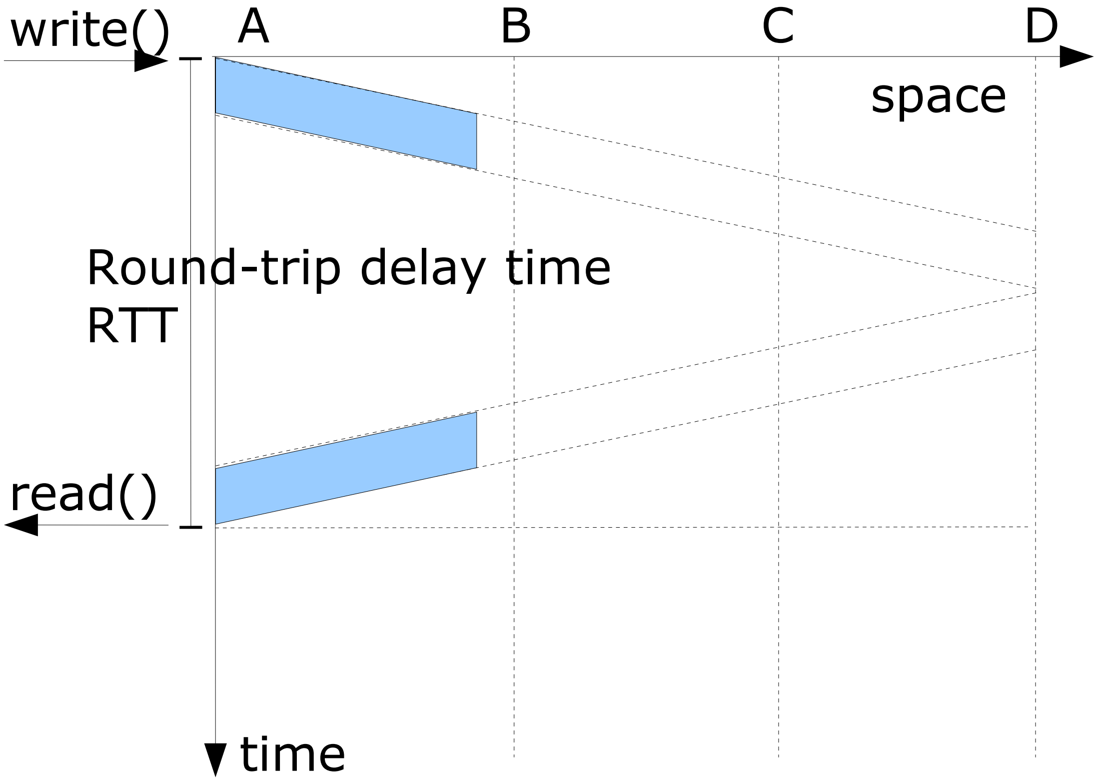
Flow Control:
TCP Sliding Window

Motivation for Sliding Window

- Stop-and-wait is inefficient
  - Only one TCP segment is “in flight” at a time
  - Especially bad for high “bandwidth-delay product”







Speed = data *(bits)* / time *(s)* → data = speed*time

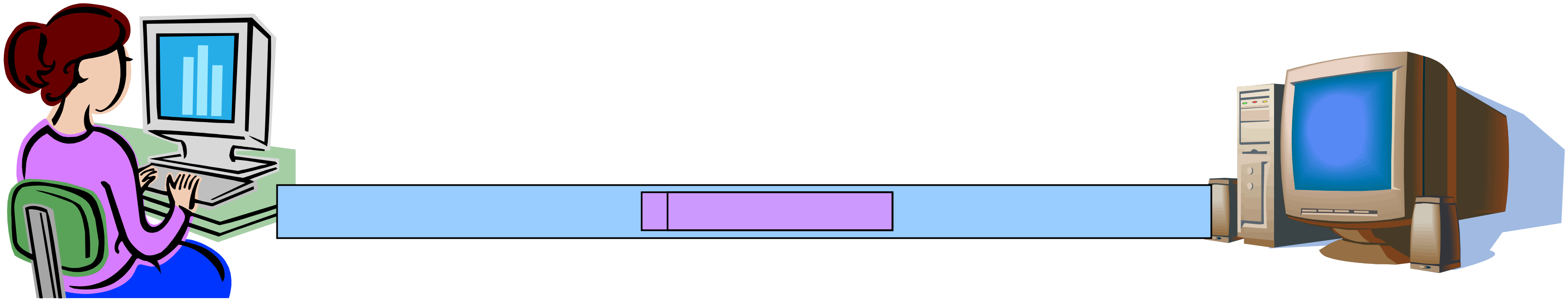
Write continuously until read returns: time = Delay (RTT)

Bandwidth-delay product (bits) = RTT * speed (bits/sec)

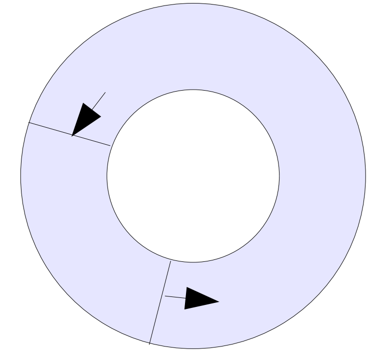
An ideal buffer: the maximum amount of data on network transmitted but not yet acknowledged

Numerical Example

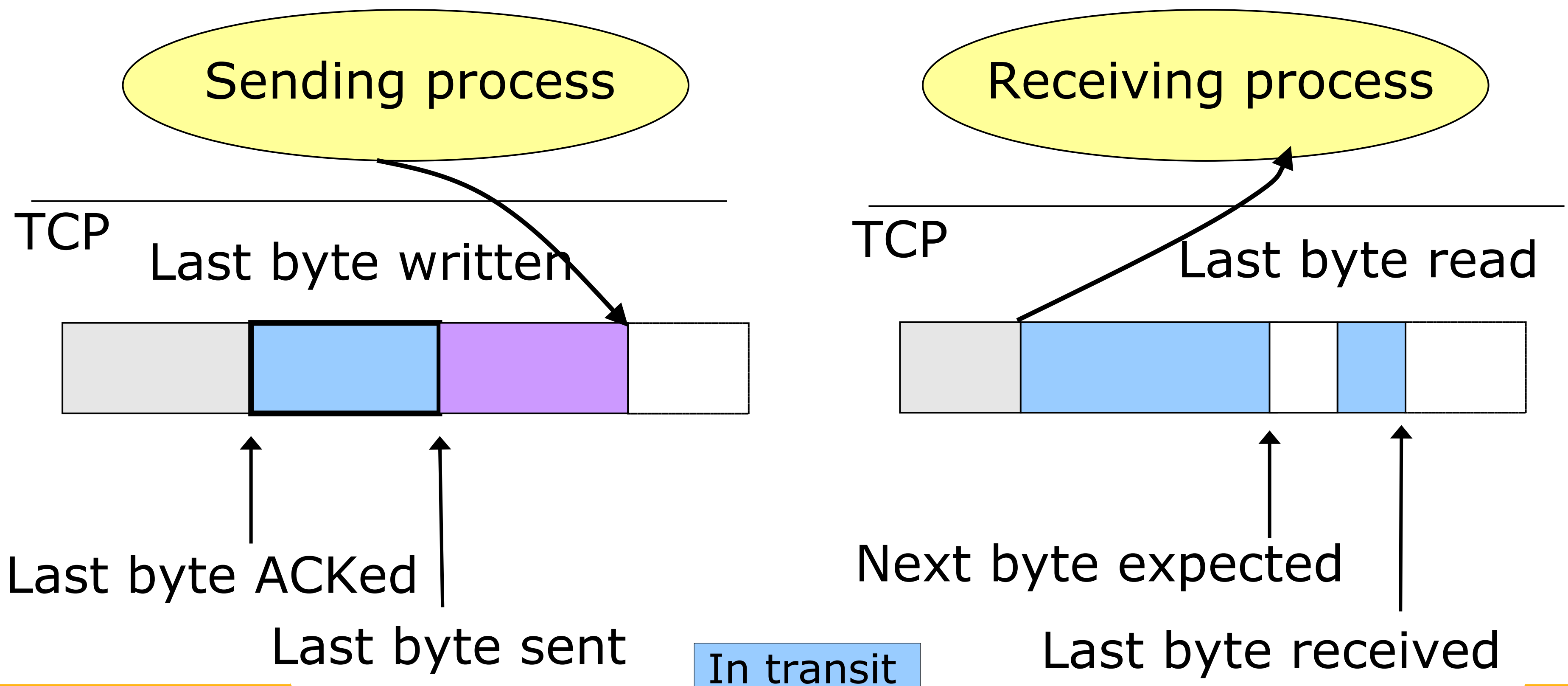
- 1.5 Mbps link with 45 msec round-trip time (RTT)
  - Bandwidth-delay product is 67.5 Kbits (or 8 KBytes)
- Sender can send at most one packet per RTT
  - Assuming a segment size of 1 KB (8 Kbits)
  - 8 Kbits/segment at 45 msec/segment ➔ 182 Kbps
  - That's just *one-eighth* of the 1.5 Mbps link capacity
- 20 Mbps with 50 msec RTT:
  - What buffer? (125KB) Many 1KB segments in 1 RTT!



Sliding Window

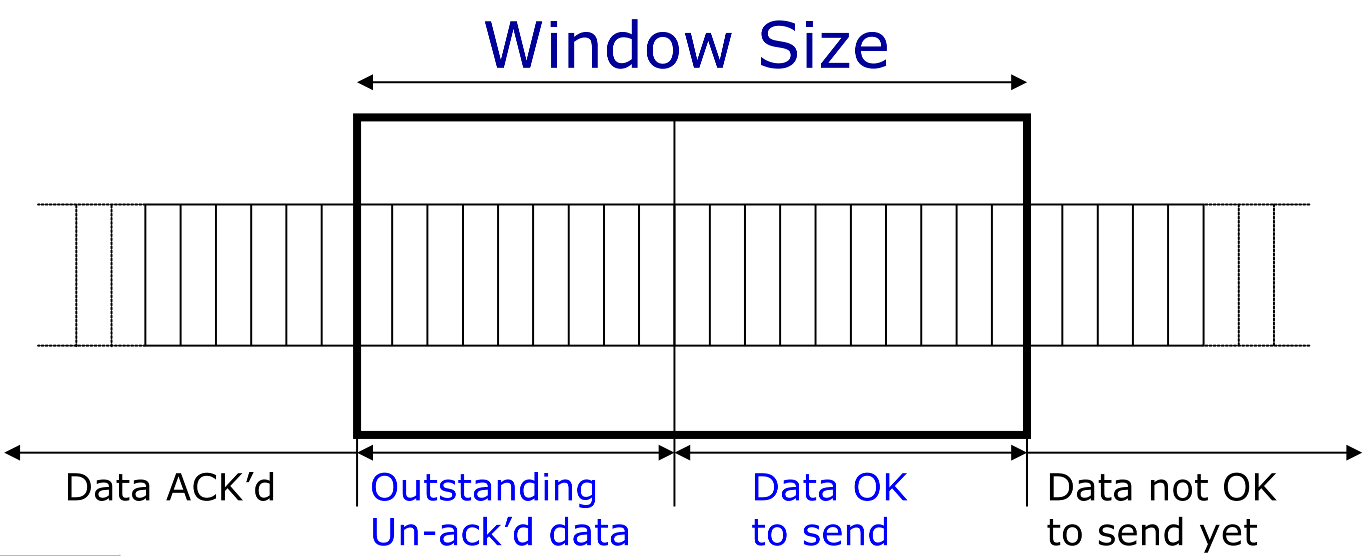


- Allow a larger amount of data "in flight"
  - Allow sender to get ahead of the receiver
  - … though not too far ahead



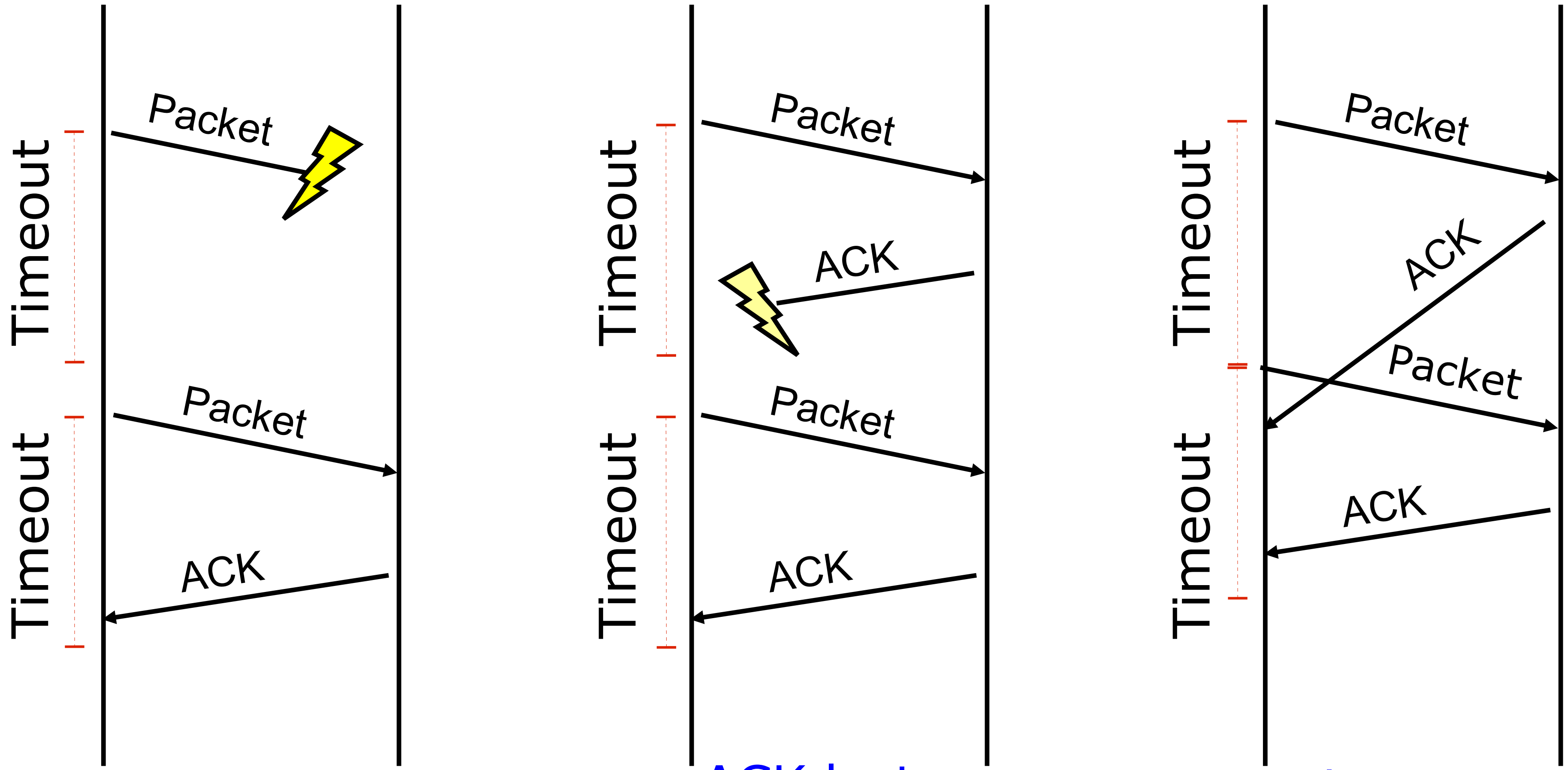
Receiver Buffering

- Receive window size
  - Amount that can be sent without acknowledgment
  - Receiver must be able to store this amount of data
- Receiver tells the sender the window
  - Tells the sender the amount of free space left



Optimizing Retransmissions

Reasons for Retransmission



Packet lost

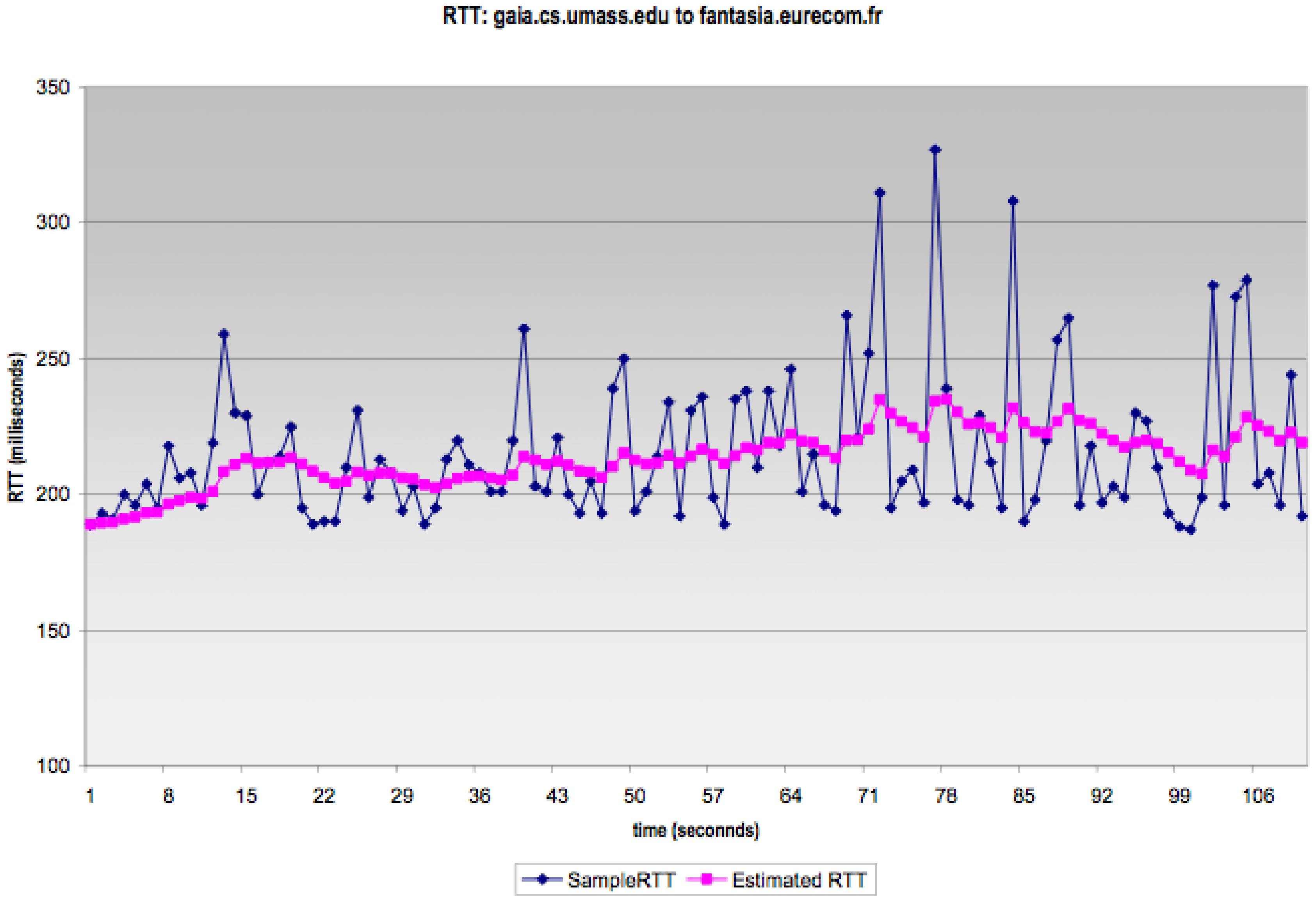
ACK lost
DUPLICATE PACKET

Early timeout
DUPLICATE PACKETS

How Long Should Sender Wait?

- Sender sets a timeout to wait for an ACK
  - Too short: wasted retransmissions
  - Too long: excessive delays when packet lost
- TCP sets timeout as a function of the RTT (RTO)
  - Expect ACK to arrive after an "round-trip time"
  - … plus a fudge factor to account for queuing
- But, how does the sender know the RTT?
  - Running average of delay to receive an ACK

Example RTT Estimation



Step 1: A's Initial SYN Packet

Flags: SYN
FIN
RST
PSH
URG
ACK

<table>
<tr><td colspan="3">A's port</td><td>B's port</td></tr>
<tr><td colspan="4">A's Initial Sequence Number</td></tr>
<tr><td colspan="4">Acknowledgment</td></tr>
<tr><td>20</td><td>0</td><td>Flags</td><td>Advertised window</td></tr>
<tr><td colspan="3">Checksum</td><td>Urgent pointer</td></tr>
<tr><td colspan="4">Options (variable)</td></tr>
</table>

A tells B it wants to open a connection…

Step 2: B's SYN-ACK Packet

Flags: SYN
FIN
RST
PSH
URG
ACK

B's port		A's port	
B's Initial Sequence Number			
A's ISN plus 1			
20	0	Flags	Advertised window
Checksum		Urgent pointer	
Options (variable)			

B tells A it accepts, and is ready to hear the next byte…

… upon receiving this packet, A can start sending data

Step 3: A's ACK of the SYN-ACK

Flags:
- SYN
- FIN
- RST
- PSH
- URG
- ACK

<table>
<tr><td colspan="3">A's port</td><td>B's port</td></tr>
<tr><td colspan="4">Sequence number</td></tr>
<tr><td colspan="4">B's ISN plus 1</td></tr>
<tr><td>20</td><td>0</td><td>Flags</td><td>Advertised window</td></tr>
<tr><td colspan="3">Checksum</td><td>Urgent pointer</td></tr>
<tr><td colspan="4">Options (variable)</td></tr>
</table>

A tells B it is okay to start sending

… upon receiving this packet, B can start sending data

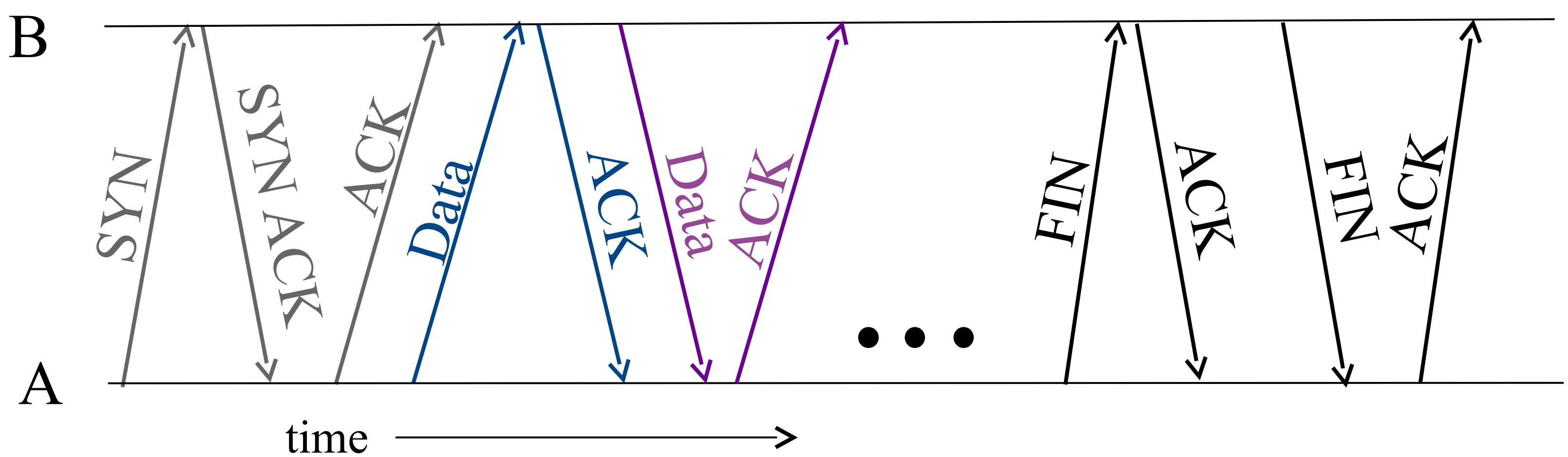
What if the SYN Packet Gets Lost?

- Suppose the SYN packet gets lost
  - Packet is lost inside the network, or
  - Server rejects the packet (e.g., full listen queue)
- Eventually, no SYN-ACK arrives
  - Sender sets a timer and wait for the SYN-ACK
  - … and retransmits the SYN if needed
- How should the TCP sender set the timer?
  - Sender has no idea how far away the receiver is
  - Some TCPs use a default of 3 or 6 seconds

SYN Loss and Web Downloads

- User clicks on a hypertext link
  - Browser creates a socket and does a “connect”
  - The “connect” triggers the OS to transmit a SYN
- If the SYN is lost…
  - The 3-6 seconds of delay is very long
  - The impatient user may click “reload”
- User triggers an “abort” of the “connect”
  - Browser “connects” on a new socket
  - Essentially, forces a fast send of a new SYN!

Tearing Down the Connection

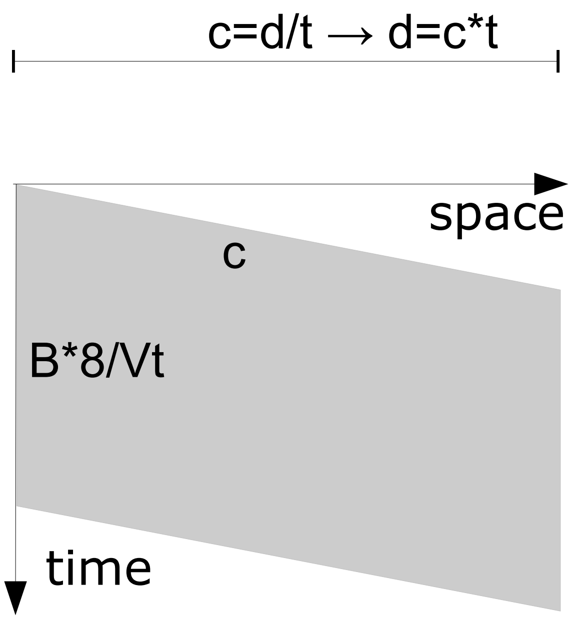


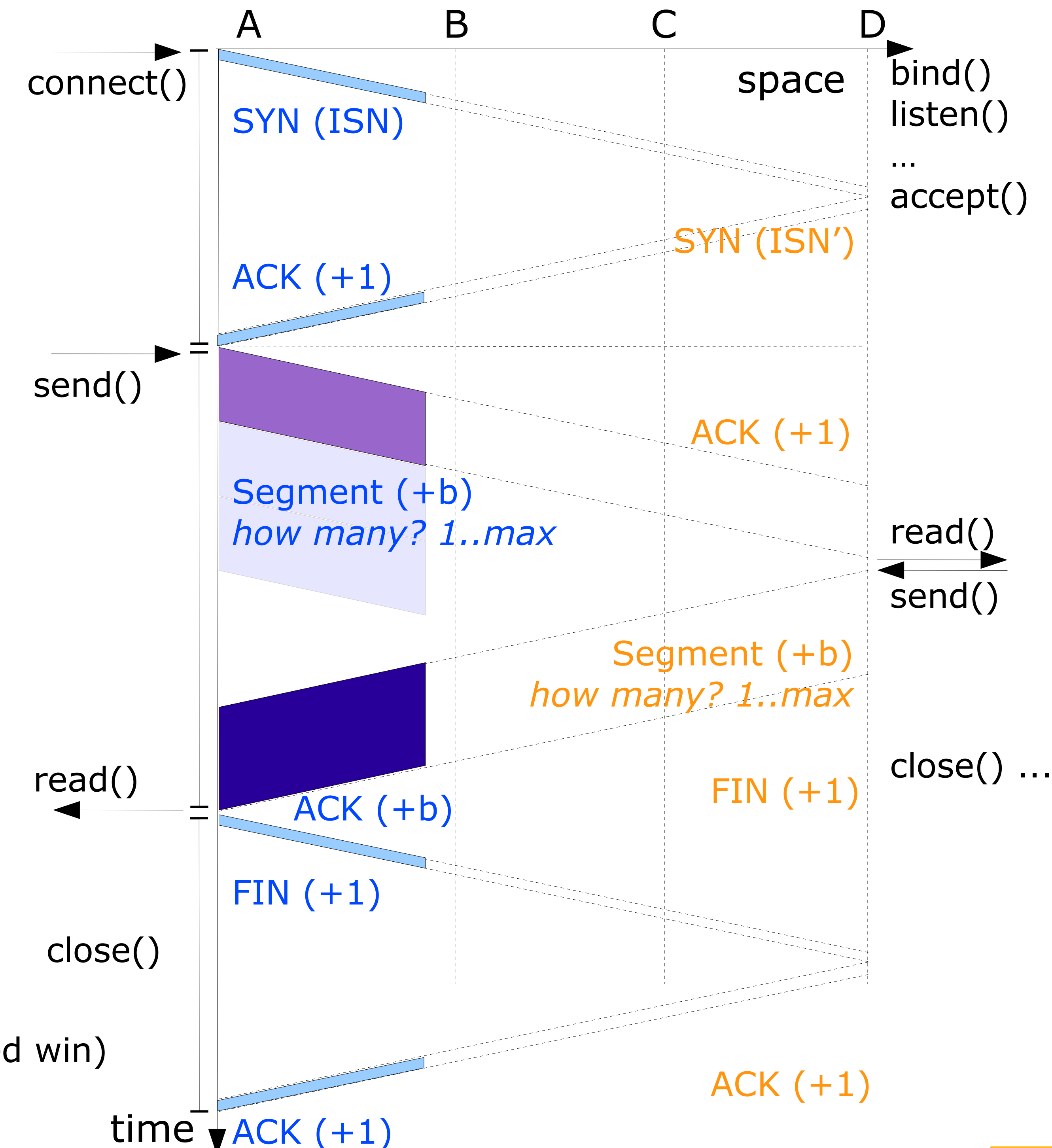
- Closing (each end of) the connection
  - Finish (FIN) to close and receive remaining bytes
  - And other host sends a FIN ACK (+1) to acknowledge
  - Reset (RST) to close and not receive remaining bytes

Sending/Receiving the FIN Packet

- Sending a FIN: close()
  - Process is done sending data via the socket
  - Process invokes “close()” to close the socket
  - Once TCP has sent all the outstanding bytes…
  - … then TCP sends a FIN

- Receiving a FIN: EOF
  - Process is reading data from the socket
  - Eventually, the attempt to read returns an EOF





Max?

BW-delay: max usage

Max buffer receiver (advertised win)

Max *buffer along path?*

Conclusions

- Transport protocols
  - Multiplexing and demultiplexing
  - Checksum-based error detection
  - Sequence numbers
  - Retransmission
  - Window-based flow control