

Transport Layer: TCP and UDP

Raj Jain

Washington University in Saint Louis

Saint Louis, MO 63130

Jain@wustl.edu

Audio/Video recordings of this lecture are available on-line at:

<http://www.cse.wustl.edu/~jain/cse4703-26/>

Student Questions



- ❑ Transport Layer Design Issues:
 - Multiplexing/Demultiplexing
 - Reliable Data Transfer
 - Flow control
 - Congestion control
- ❑ UDP
- ❑ TCP
 - Header format, connection management, checksum
 - Congestion Control
- ❑ **Note:** This class lecture is based on Chapter 3 of the textbook (Kurose and Ross) and the figures provided by the authors.

Student Questions

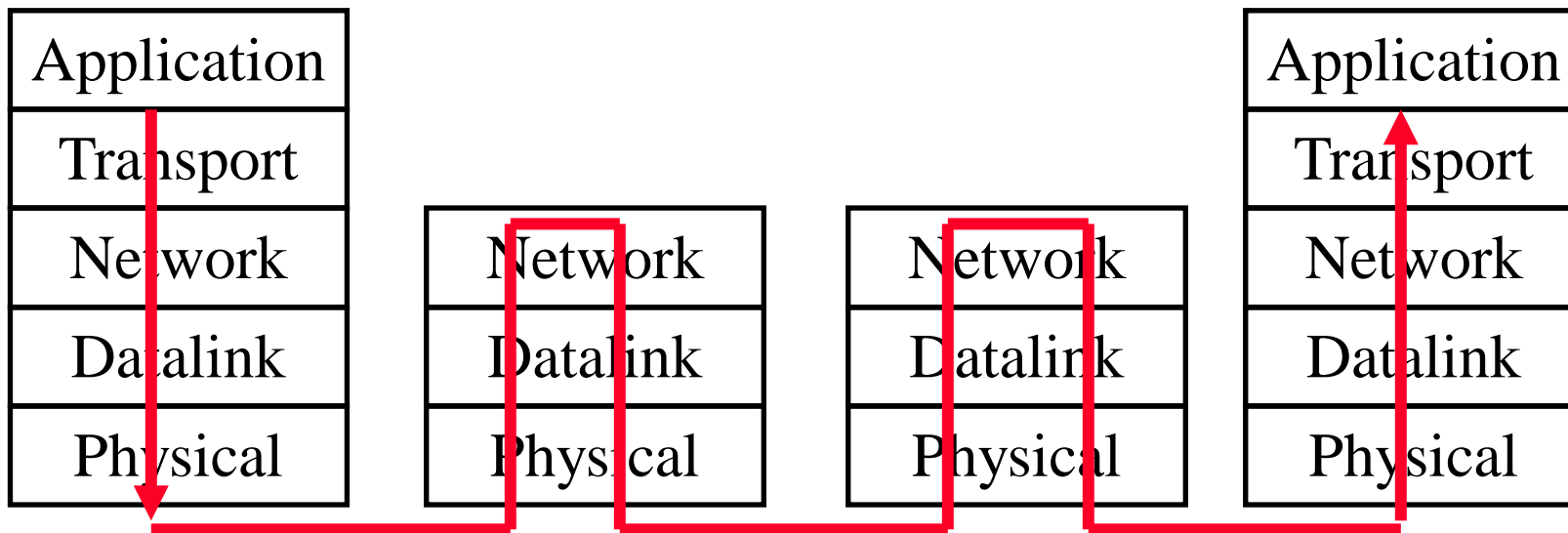
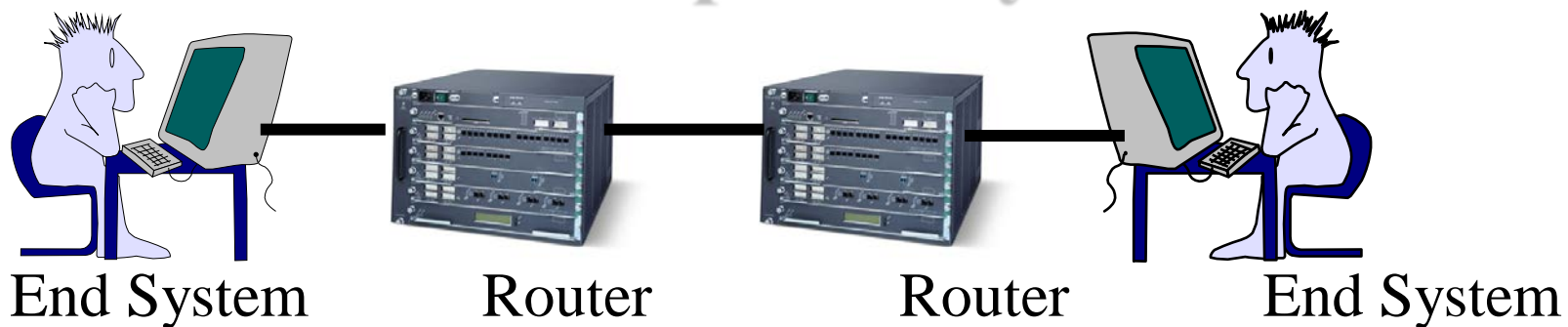


Transport Layer Design Issues

1. Transport Layer Functions
2. Multiplexing and Demultiplexing
3. Error Detection: Checksum
4. Flow Control
5. Efficiency Principle
6. Error Control: Retransmissions

Student Questions

Transport Layer



- ❑ Transport = End-to-End Services
Services required at source and destination systems
Not required on intermediate hops

Student Questions

Transport Layer Functions

1. **Multiplexing and demultiplexing:** Among applications and processes at end systems
2. **Error detection:** Bit errors
3. **Loss detection:** Lost packets due to buffer overflow at intermediate systems (Sequence numbers and acks)
4. **Error/loss recovery:** Retransmissions
5. **Flow control:** Ensuring the destination has buffers
6. **Congestion Control:** Ensuring the network has capacity

Not all transports provide all functions

Student Questions

Protocol Layers

- Top-Down approach

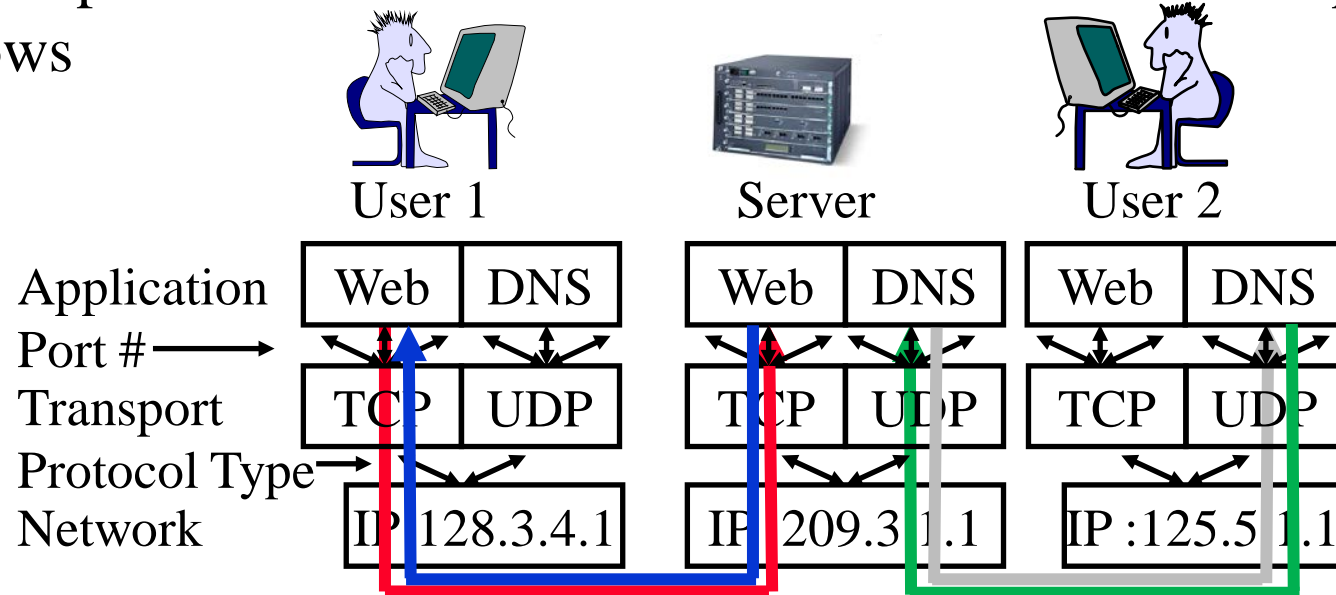
Application	HTTP	FTP	SMTP	P2P	DNS	Skype
Transport	TCP				UDP	
Internetwork	IP					
Host to Network	Ethernet	Point-to-Point		Wi-Fi		
Physical	Coax	Fiber	Wireless			

- Two other transport protocols, Stream Control Transport Protocols (**SCTP**), and Datagram Congestion Control Protocol (**DCCP**) are used but not covered in the textbook or this course. A variation called **Cubic** is discussed briefly later in this module.

Student Questions

Multiplexing and Demultiplexing

- Transport **Ports** and Network **addresses** are used to separate flows



HTTP Req. SP:3009 | DP:80 | SA: 128.3.4.1 | DA: 209.3.1.1 →

HTTP Resp. SP:80 | DP:3009 | SA:209.3.1.1 | DA:128.3.4.1 ←

DNS Req. SP:5009 | DP:53 | SA: 125.5.4.1 | DA: 209.3.1.1 ←

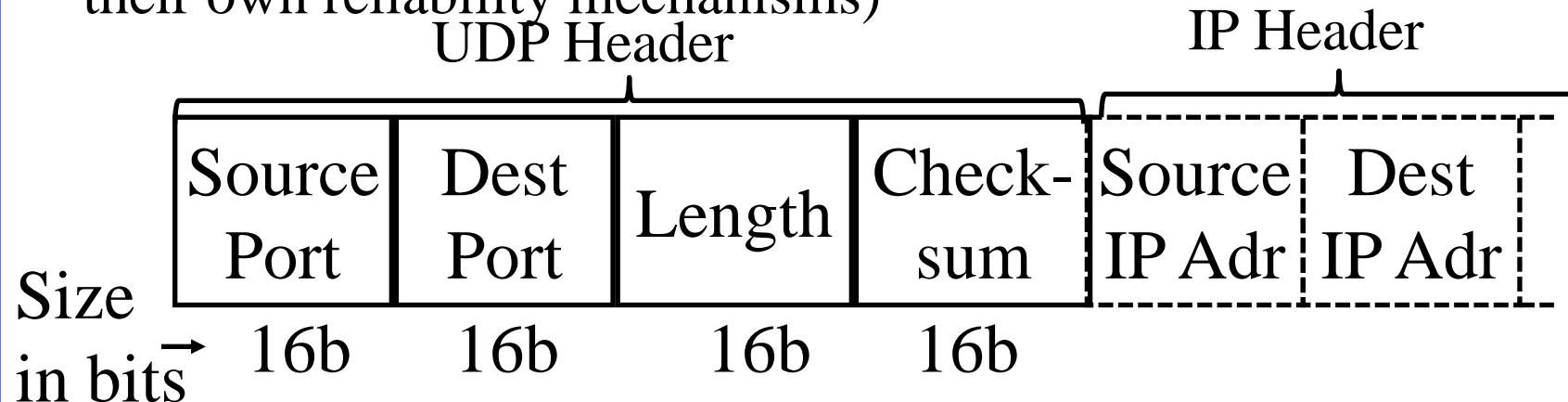
DNS Resp. SP:53 | DP:3009 | SA:209.3.1.1 | DA:125.5.4.1 →

Student Questions

Ref: http://en.wikipedia.org/wiki/List_of_TCP_and_UDP_port_numbers

User Datagram Protocol (UDP)

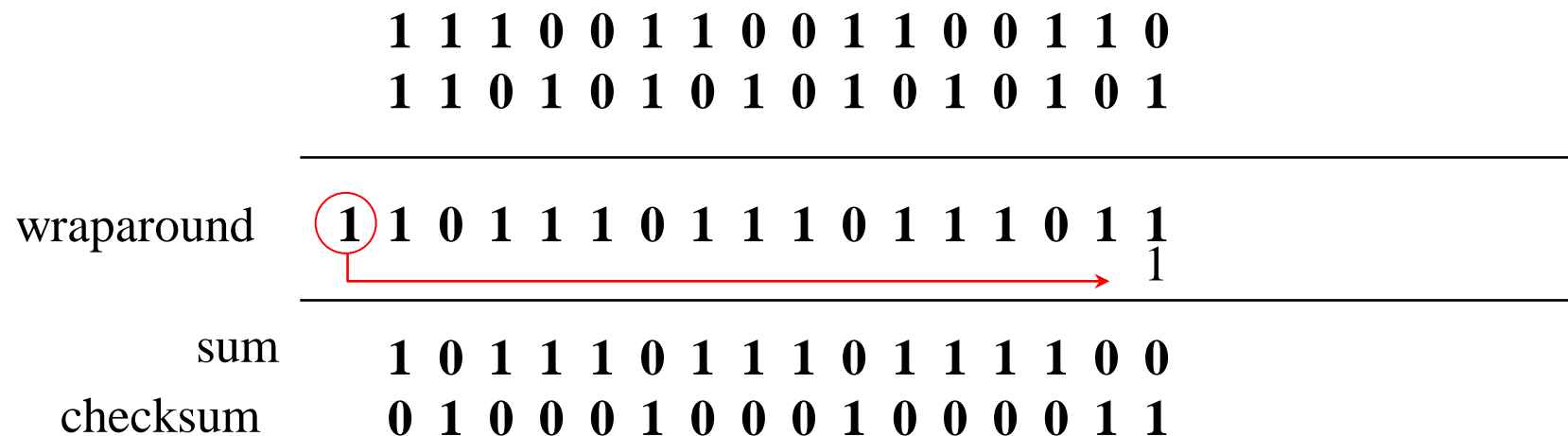
- ❑ Connectionless end-to-end service
- ❑ Provides multiplexing via ports
- ❑ Error detection (Checksum) is optional. Applies to **pseudo-header** (same as TCP) and UDP segment. If not used, it is set to zero.
- ❑ No error recovery (no acks). No retransmissions.
- ❑ Used by network management, DNS, and streamed multimedia (Applications that are loss-tolerant, delay-sensitive, or have their own reliability mechanisms)



Student Questions

Error Detection: Checksum

- ❑ **Cyclic Redundancy Check (CRC):** Powerful but generally requires hardware
- ❑ **Checksum:** Weak but easily done in software
 - **Example:** 1's complement of 1's complement sum of 16-bit words with overflow wrapped around



At the destination, the sum is all 1's, and the checksum is zero.

Student Questions

1's Complement

2's Complement: -ve of a number is complement+1

- $1 = 0001$ $-1 = 1111$
- $2 = 0010$ $-2 = 1110$
- $0 = 0000$ $-0 = 0000$

1's complement: -ve of a number is its complement

- $1 = 0001$ $-1 = 1110$
- $2 = 0010$ $-2 = 1101$
- $0 = 0000$ $-0 = 1111$

2's Complement sum: Add with carry. Drop the final carry, if any.

$$6-7 = 0110 + (-0111) = 0110 + 1001 = 1111 \Rightarrow -1$$

1's complement sum: Add with carry. Add end-around carry back to sum

$$\square 6-7 = 0110 + (-0111) = 0110+1000 = 1110 \Rightarrow -1$$

Complement of 1's complement sum: 0001

Checksum: At the transmitter: 0110 1000, append 0001

At the destination: 0110 1000 0001 compute checksum of the full packet
= complement of sum = complement of 1111 = 0000

Ref: https://en.wikipedia.org/wiki/Ones%27_complement

Washington University in St. Louis

<http://www.cse.wustl.edu/~jain/cse4703-26/>

©2026 Raj Jain

Student Questions

Homework 3A: Checksum

[6 points] Consider the following two 16-bit words: ABCD 1234

- A. What is the checksum as computed by the Source
- B. Add your answer of Part A to the end of the packet and show how the destination will compute the checksum of the received three 16-bit words and confirm that there are no errors.
- C. Now assume that the first bit of the packet is flipped due to an error. Repeat Part B at the destination. Is the error detected?

Student Questions

1's Complement: More Examples

- Fill in the missing numbers. Each decimal cell should have two numbers as shown in the bottom left.

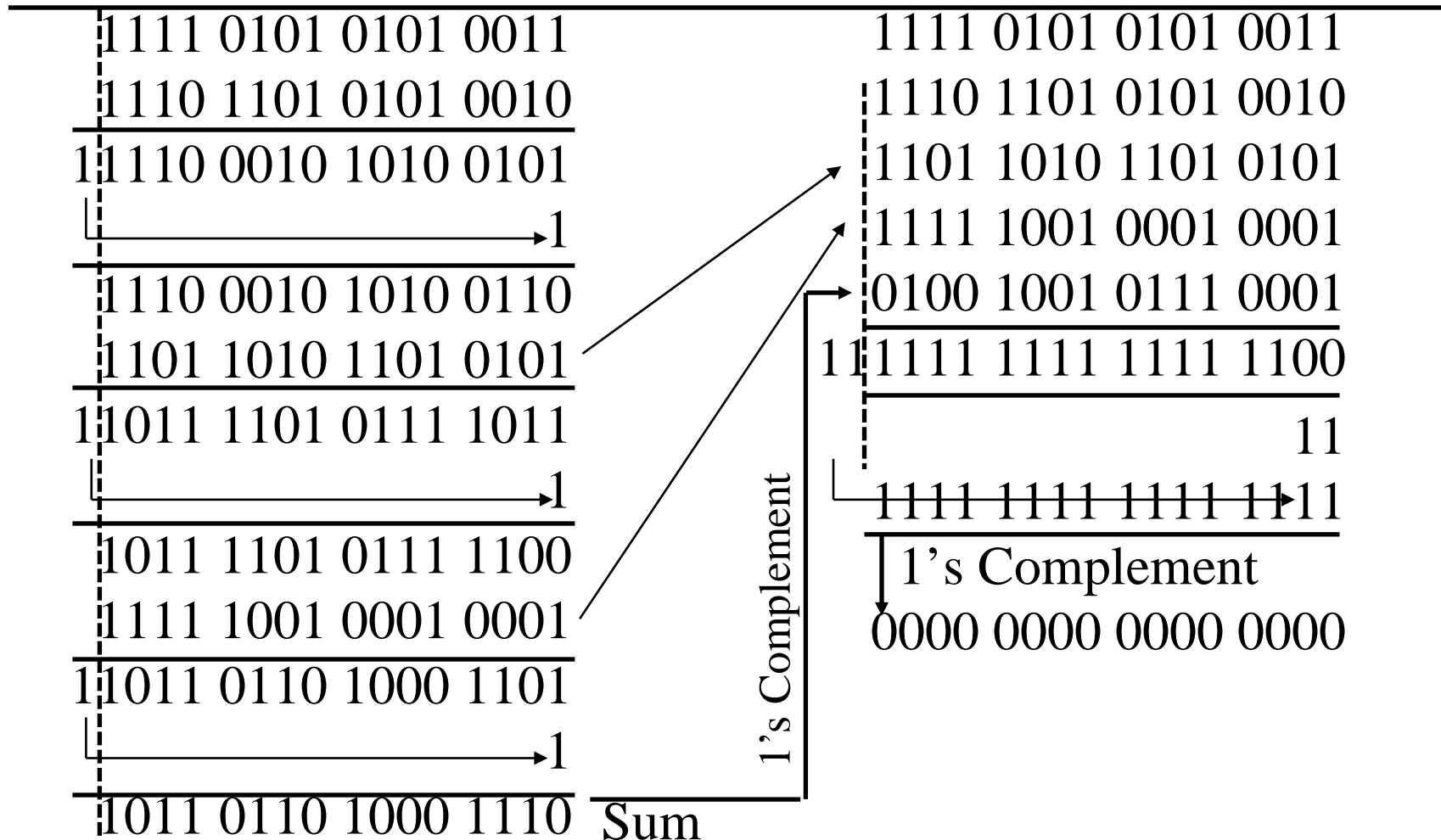
Decimal	1's Complement	Decimal	1's Complement
0	0000	-0	1111
1		-1	
2	0010	-2	1101
3		-3	
4		-4	
5	0101	-5	1010
6		-6	
7		-7	
8	1000	-8	0111
9		-9	
10		-10	
11		-11	
12	1100	-12	0011
13		-13	
14		-14	
15	1111	-15	0000
11 or -4	1011		1001

Student Questions

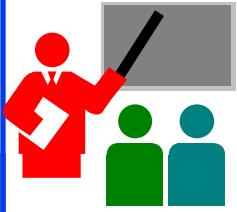
Checksum: Another Example

Sender: 4-word message

Destination



Student Questions



UDP: Summary

1. UDP provides flow multiplexing using port #s
2. UDP optionally provides error detection using the checksum
3. UDP does not have an error or loss recovery mechanism

Student Questions

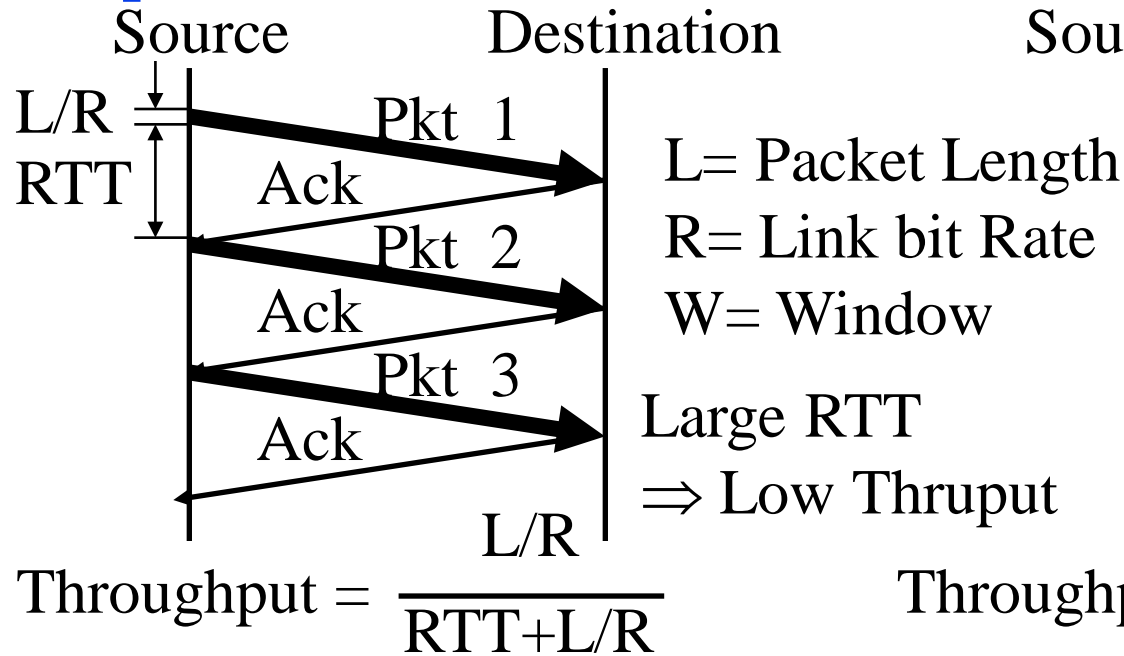


Flow Control

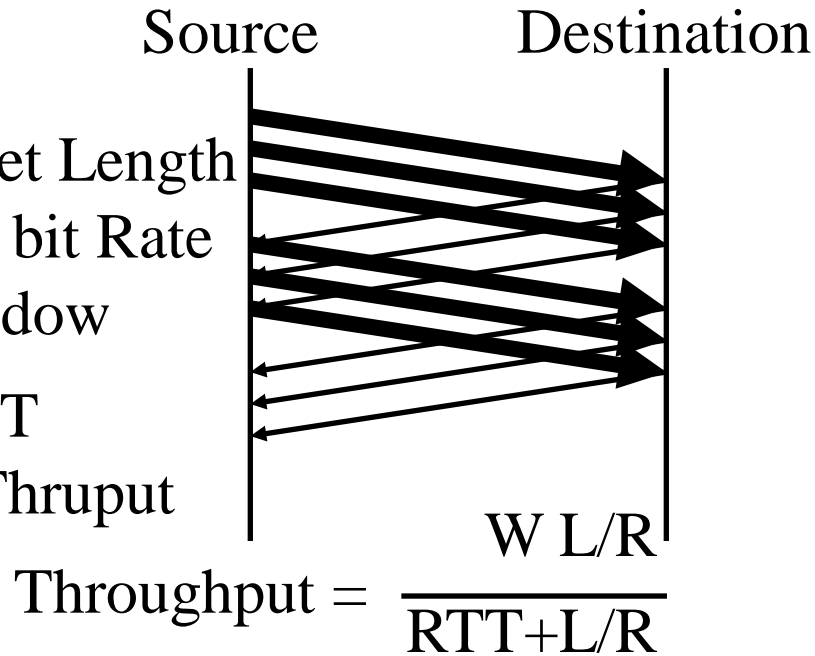
Flow Control Goals:

1. Source does not flood the destination,
2. Maximize throughput

Stop and Wait Flow Control

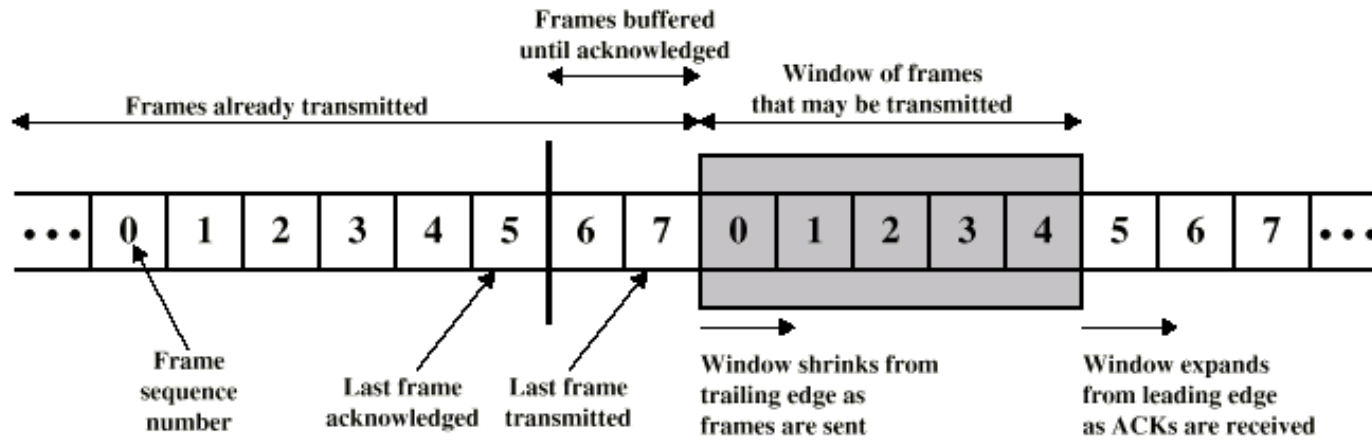


Window Flow Control

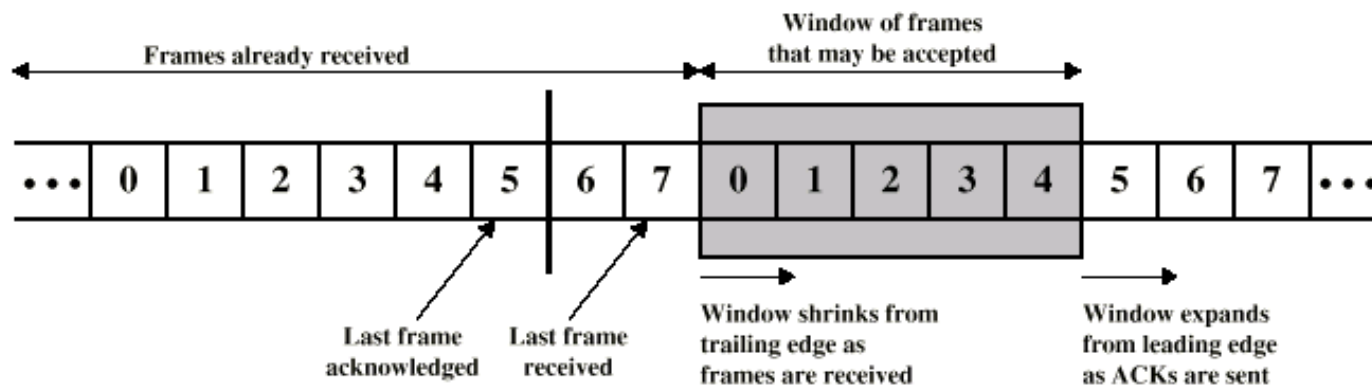


Student Questions

Sliding Window Diagram



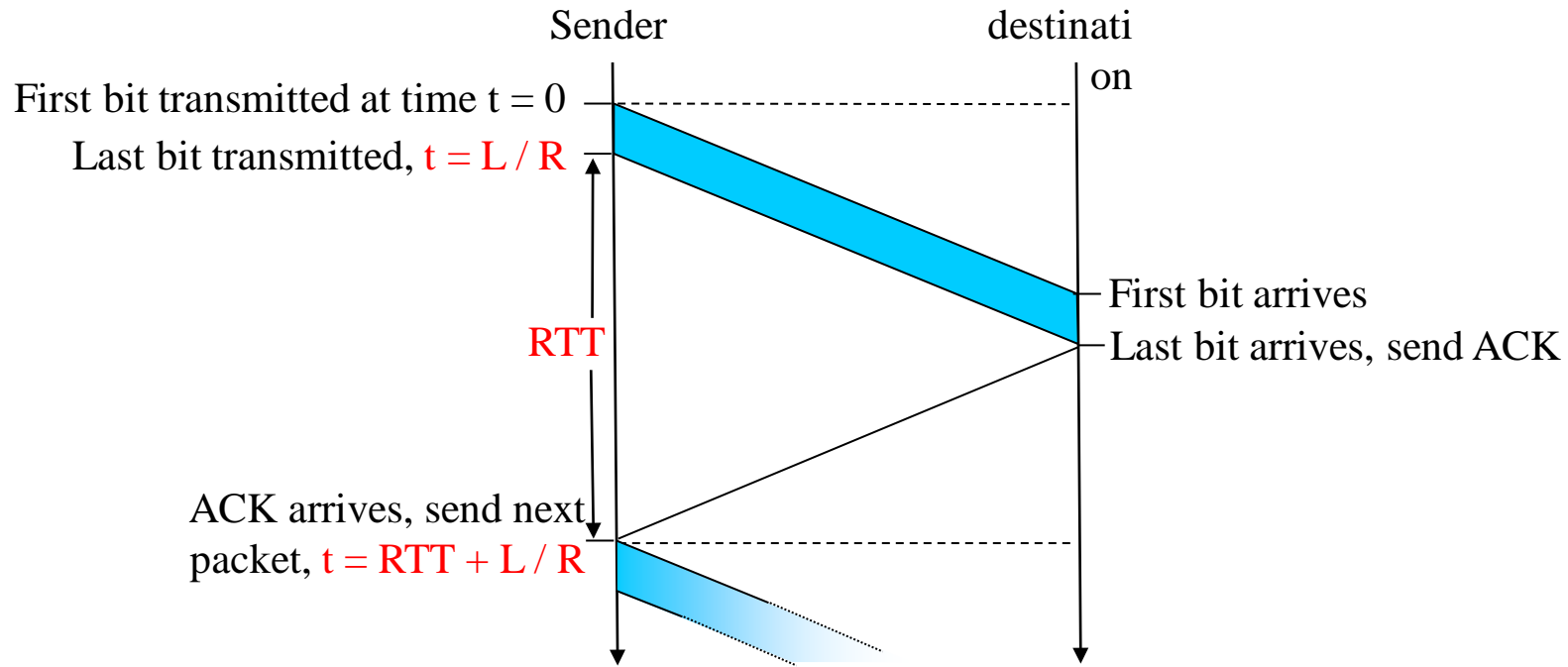
(a) Sender's perspective



(b) Receiver's perspective

Student Questions

Stop and Wait Flow Control

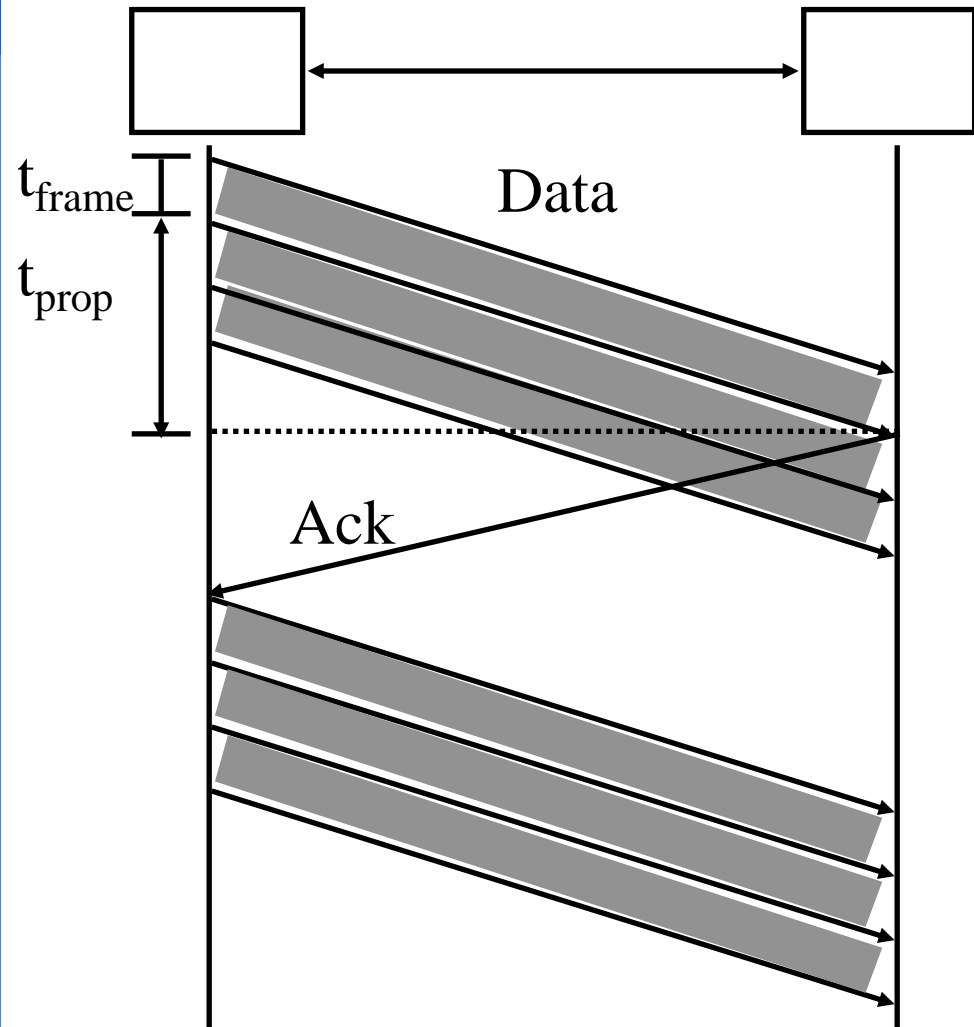


$$\text{Utilization } U = \frac{L / R}{RTT + L / R} = \frac{t_{\text{frame}}}{2t_{\text{prop}} + t_{\text{frame}}} = \frac{1}{2\alpha + 1}$$

Here, $\alpha = t_{\text{prop}} / t_{\text{frame}}$

Student Questions

Sliding Window Protocol Efficiency



$$U = \frac{W t_{\text{frame}}}{2t_{\text{prop}} + t_{\text{frame}}}$$

$$= \begin{cases} \frac{W}{2\alpha + 1} \\ 1 \text{ if } W > 2\alpha + 1 \end{cases}$$

Here, $\alpha = t_{\text{prop}}/t_{\text{frame}}$

$W=1 \Rightarrow$ Stop and Wait

Student Questions

Utilization: Examples

Satellite Link: One-way Propagation Delay = 270 ms

RTT=540 ms

Frame Size $L = 500$ Bytes = 4 kb

Data rate $R = 56$ kbps $\Rightarrow t_{\text{frame}} = L/R = 4\text{kb}/56\text{kbps} = 0.071\text{s} = 71$ ms

$\alpha = t_{\text{prop}}/t_{\text{frame}} = 270/71 = 3.8$

$U = 1/(2\alpha+1) = 0.12$

❑ Short Link: 1 km = 5 μs (Assuming Fiber 200 m/ μs),

Rate=10 Mbps,

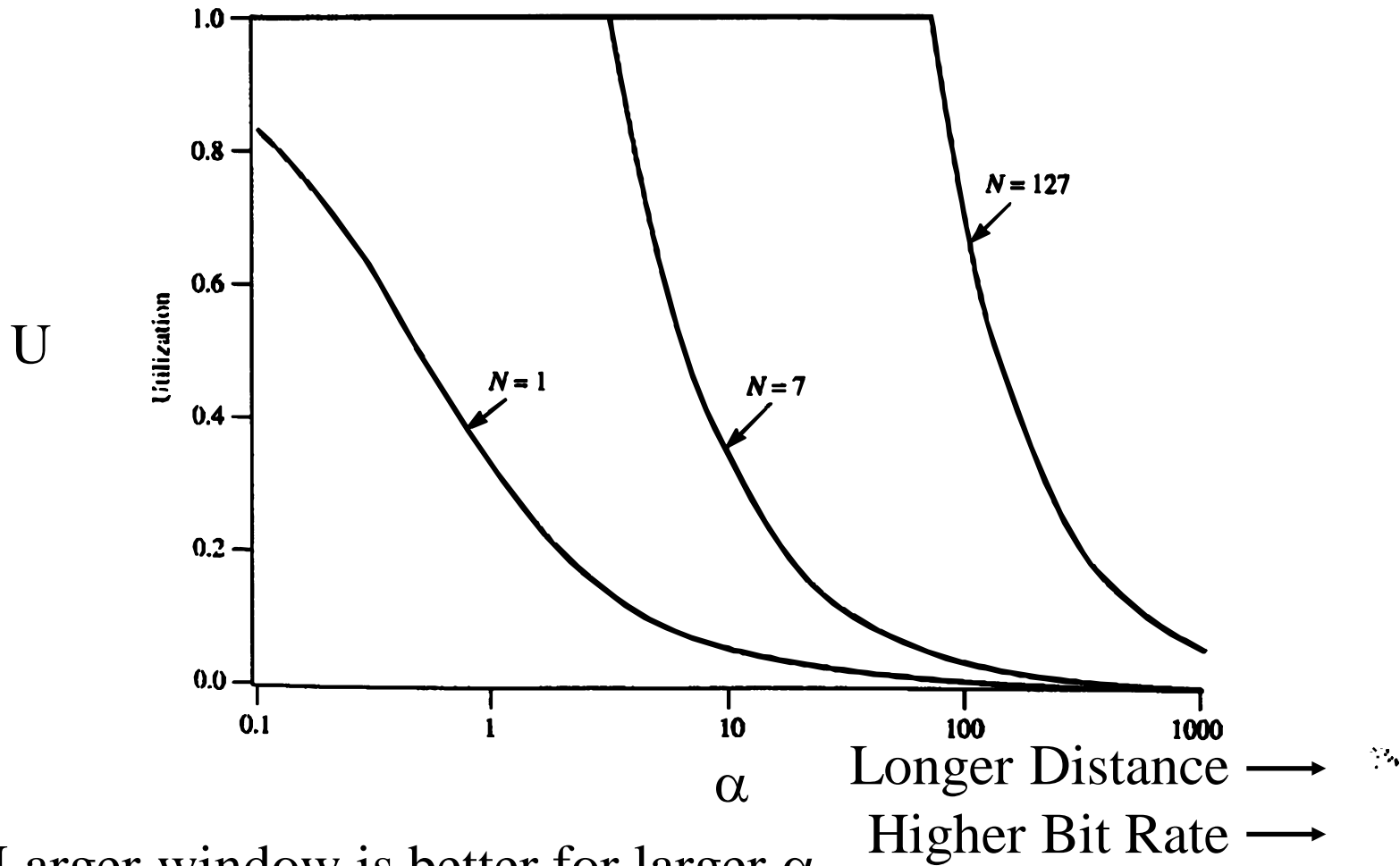
Frame=500 bytes $\Rightarrow t_{\text{frame}} = 4\text{k}/10\text{M} = 400 \mu\text{s}$

$\alpha = t_{\text{prop}}/t_{\text{frame}} = 5/400 = 0.012 \Rightarrow U = 1/(2\alpha+1) = 0.98$

Note: The textbook uses RTT in place of t_{prop} and L/R for t_{frame}

Student Questions

Effect of Window Size

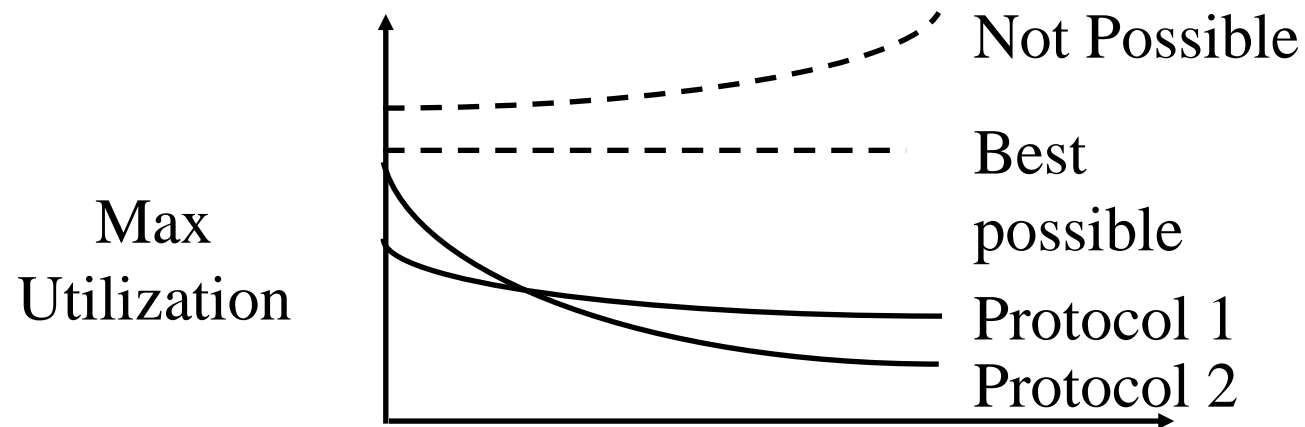


□ Larger window is better for larger α

Student Questions

Efficiency Principle

- For **all** protocols, the maximum utilization (efficiency) is a *non-increasing* function of α .



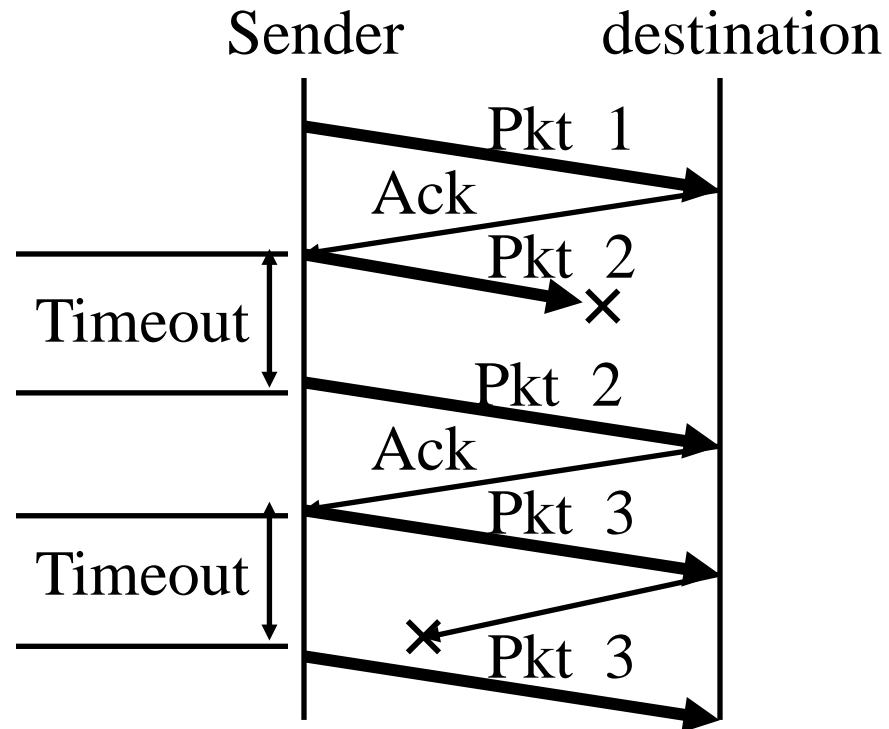
$$\alpha = \frac{t_{\text{prop}}}{t_{\text{frame}}} = \frac{\text{Distance/Speed of Signal}}{\text{Bits Transmitted /Bit rate}}$$
$$= \frac{\text{Distance} \times \text{Bit rate}}{\text{Bits Transmitted} \times \text{Speed of Signal}}$$

Student Questions

Error Control: Retransmissions

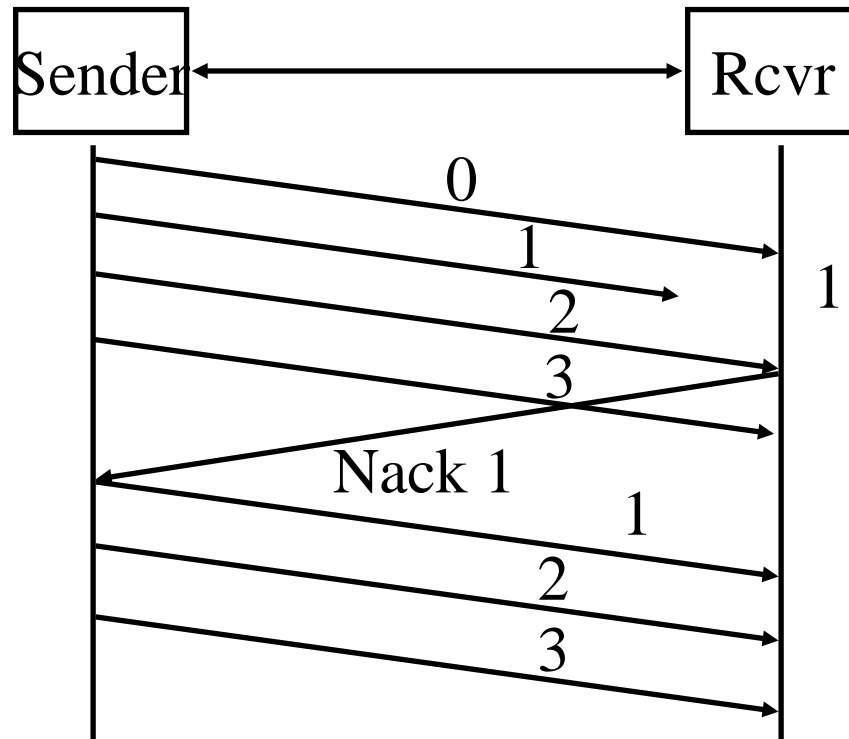
- ❑ Error Control = Error Recovery
- ❑ Retransmit lost packets \Rightarrow Automatic Repeat reQuest (ARQ)

Stop and Wait ARQ



Student Questions

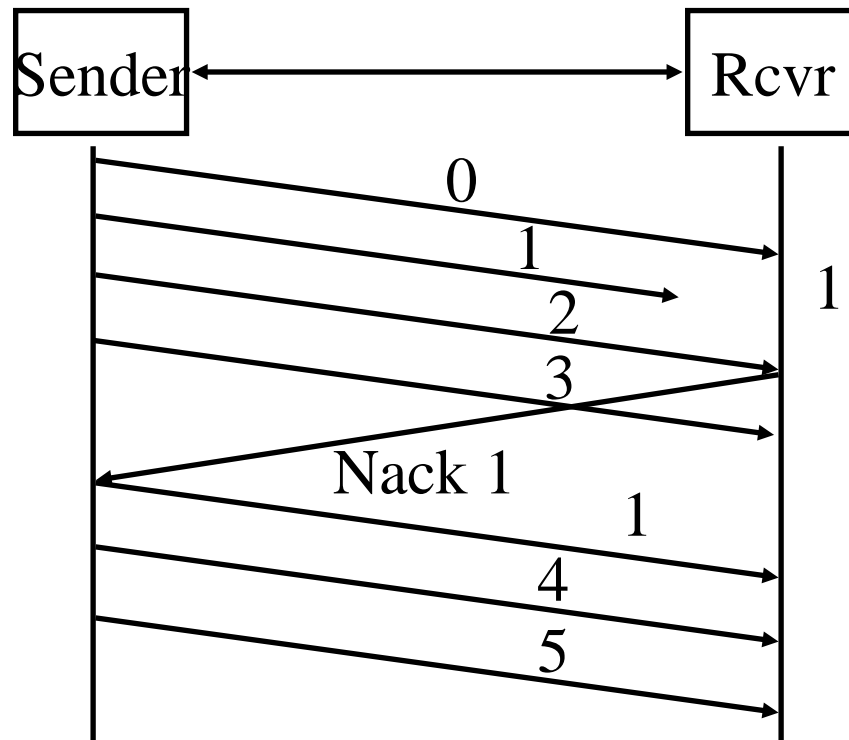
Go-Back-N ARQ



- ❑ The destination does not cache out-of-order frames
- ❑ The sender has to *go back* and retransmit all frames after the lost frame.

Student Questions

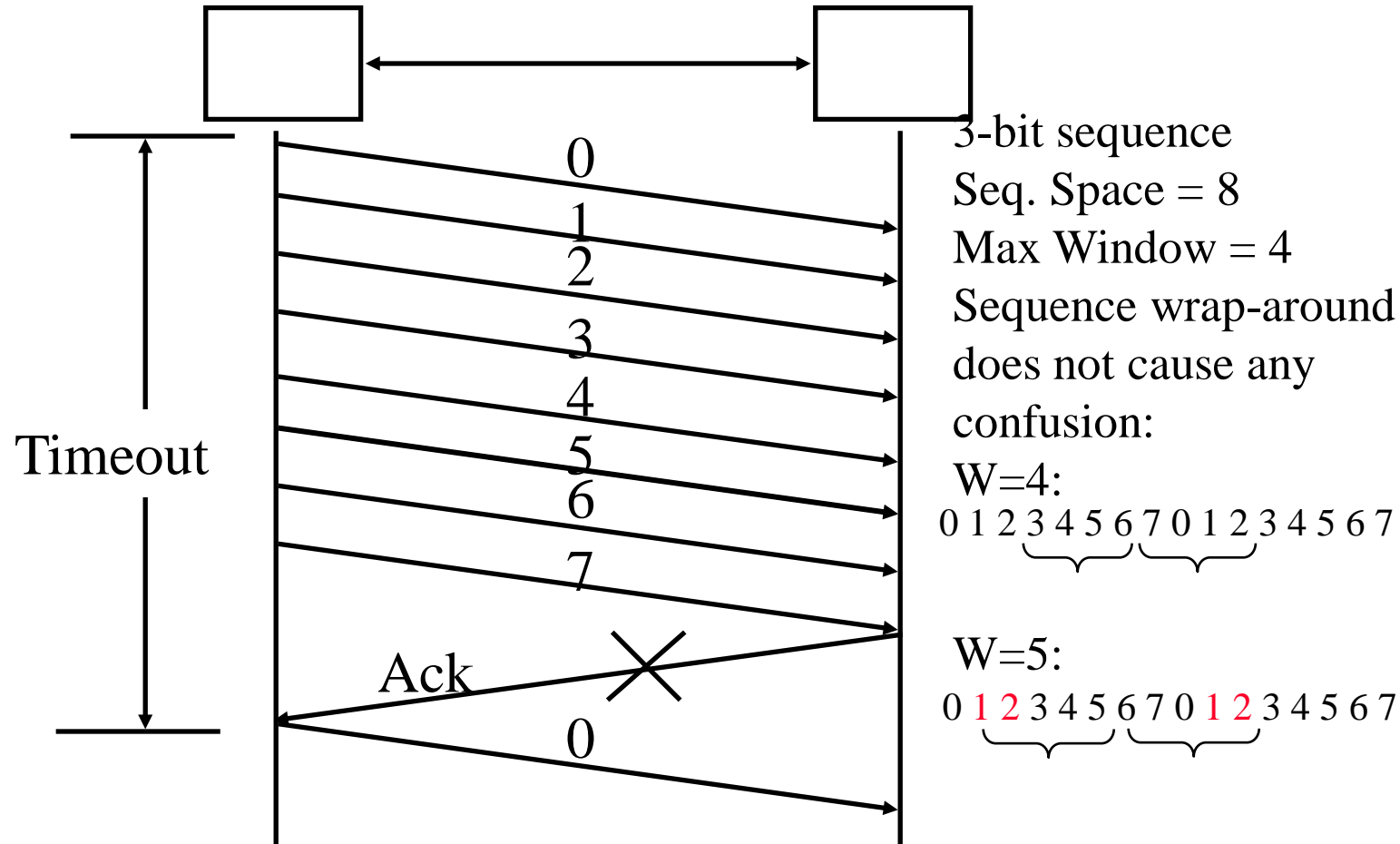
Selective Repeat ARQ



- ❑ The destination caches out-of-order frames
- ❑ The sender retransmits only the lost frame
- ❑ Also known as selective *reject* ARQ

Student Questions

Selective Repeat: Window Size



Sequence number space ≥ 2 window size

Window size $\leq 2^{n-1}$ with n bit sequence numbers

Student Questions

Performance: Maximum Utilization

□ **Stop and Wait Flow Control:** $U = 1/(1+2\alpha)$

□ **Window Flow Control:**

$$\alpha = \frac{\text{Propagation Time}}{\text{Frame Time}}$$

$$U = \begin{cases} 1 & W \geq 1+2\alpha \\ W/(1+2\alpha) & W < 1+2\alpha \end{cases}$$

□ **Stop and Wait ARQ:** $U = (1-P)/(1+2\alpha)$

□ **Go-back-N ARQ:** $P = \text{Probability of Loss}$

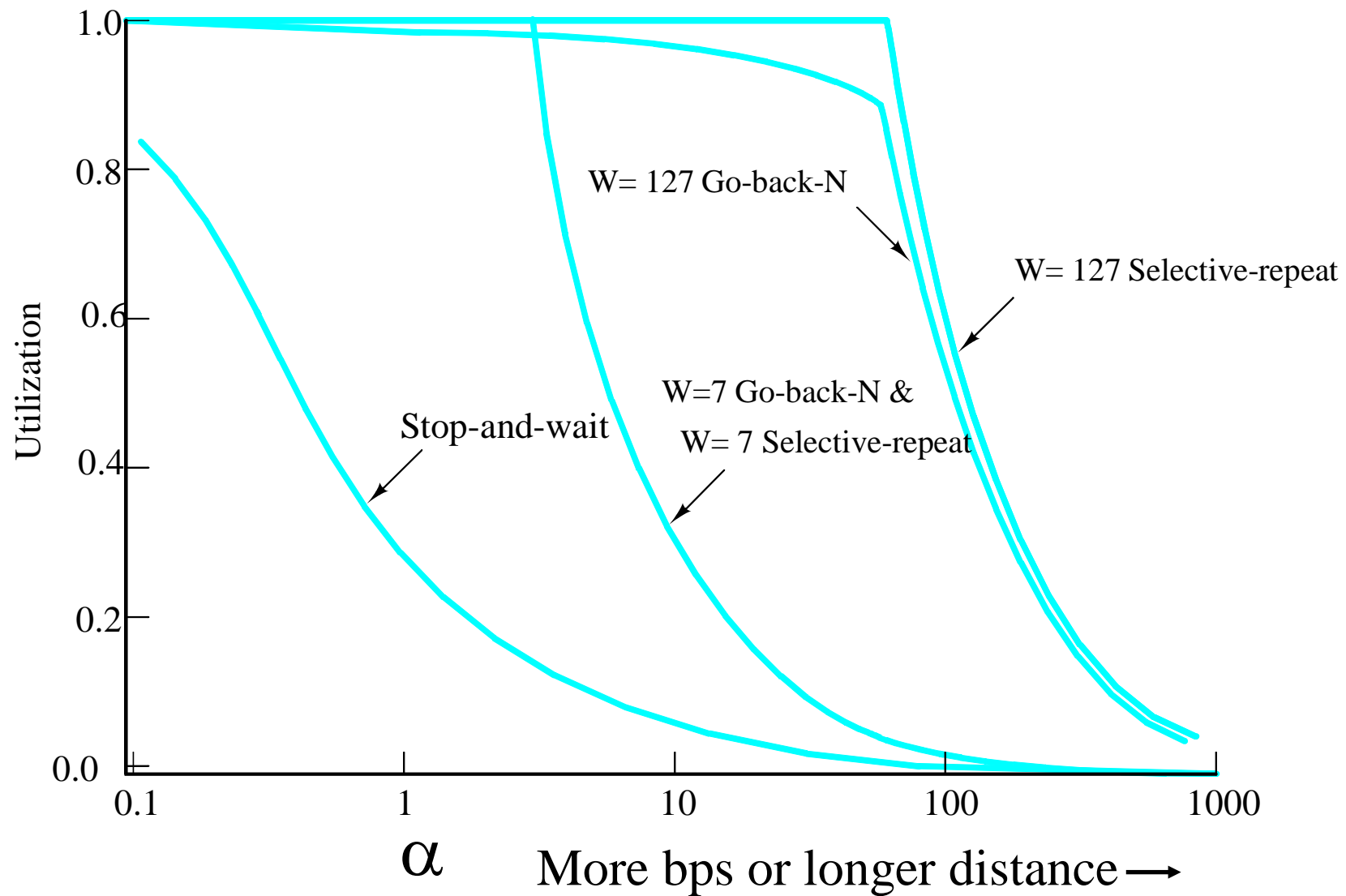
$$U = \begin{cases} (1-P)/(1+2\alpha P) & W \geq 1+2\alpha \\ W(1-P)/[(1+2\alpha)(1-P+WP)] & W < 1+2\alpha \end{cases}$$

□ **Selective Repeat ARQ:**

$$U = \begin{cases} (1-P) & W \geq 1+2\alpha \\ W(1-P)/(1+2\alpha) & W < 1+2\alpha \end{cases}$$

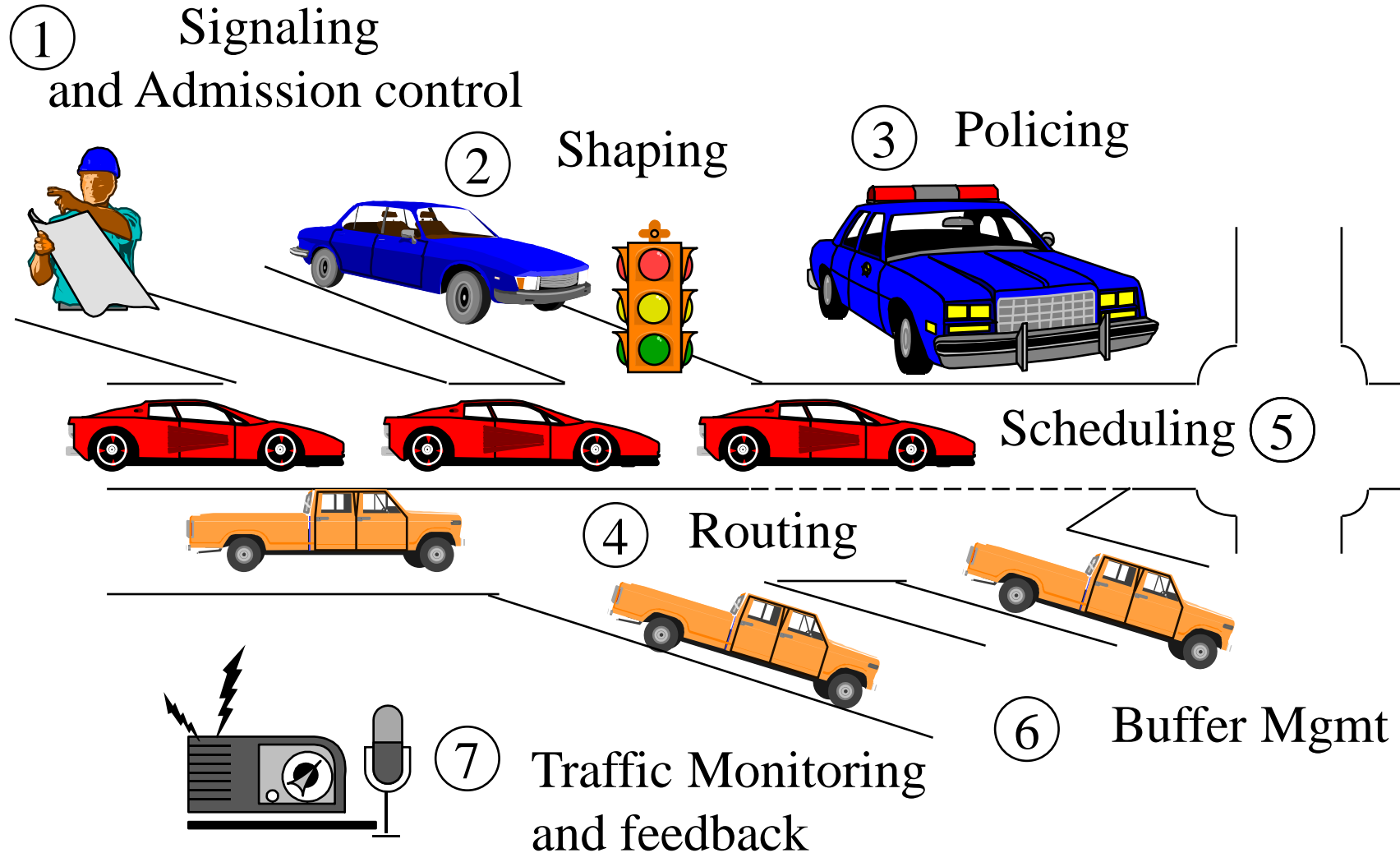
Student Questions

Performance Comparison

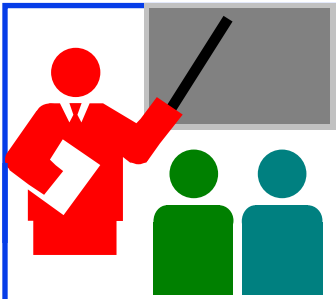


Student Questions

Traffic Management Methods



Student Questions



Transport Layer Design: Summary

1. Multiplexing/demultiplexing by a combination of source and destination IP addresses and port numbers.
2. Window flow control is better for long-distance or high-speed networks
3. Longer distance or higher speed
 \Rightarrow Larger $\alpha \Rightarrow$ Larger window is better.
4. Stop and and wait flow control is ok for short-distance or low-speed networks
5. Selective repeat is better than stop and wait ARQ
Only slightly better than go-back-N

Read Sections 3.4.2, 3.4.3, and 3.4.4. Do R11-13, P23, P24

Do Lab 3.

Student Questions

Homework 3B: Flow Control

[8 points] Similar to problem 22 on page 292 of the textbook:

Consider the GBN protocol with a sender window size of 3 and a sequence number range of 1,024. Suppose that at time t , the next in-order packet the destination expects has a sequence number k .

- A. If the source has received $k-2$ ack. What are the sequence numbers of the packets that the source may have sent?
- B. What are all possible values of the ACK field in all possible messages currently propagating back to the sender at time t ?

Window Flow Control:

- C. How big a window (in the number of packets) is required for the channel utilization to be greater than 70% on a cross-country fiber link of 3000 km running at 40 Mbps using 1 kByte packets? Round up your answer to the next kB.

Efficiency Principle:

- D. Ethernet V1 access protocol was designed to run at 10 Mbps over 2.5 km using 1500-byte packets. This same protocol must be used at 100 Mbps with the same efficiency. What distance in meters can it cover if the frame size is not changed?

Student Questions

Lab 3: Reliable Transport Protocol

[60 points] Overview

In this laboratory programming assignment, you will be writing the sending and receiving transport-level code for implementing a simple reliable data transfer protocol. Although there are two versions of this lab, the Alternating-Bit-Protocol version and the Go-Back-N version, you are supposed to do only ABP version. This lab should be **fun** since your implementation will differ very little from what would be required in a real-world situation.

Since you probably don't have standalone machines (with an OS that you can modify), your code will have to execute in a simulated hardware/software environment. However, the programming interface provided to your routines, i.e., the code that would call your entities from above and from below is very close to what is done in an actual UNIX environment. (Indeed, the software interfaces described in this programming assignment are much more realistic than the infinite loop senders and receivers that many texts describe). Stopping/starting of timers are also simulated, and timer interrupts will cause your timer handling routine to be activated.

The routines you will write

The procedures you will write are for the sending entity (A) and the receiving entity (B). Only unidirectional transfer of data (from A to B) is required. Of course, the B side will have to send packets to A to acknowledge (positively or negatively) receipt of data. Your routines are to be implemented in the form of the procedures described below. These procedures will be called by (and will call) procedures that I have written which emulate a network environment. The overall structure of the environment is shown in Figure Lab.3-1 (structure of the emulated environment):

The unit of data passed between the upper layers and your protocols is a *message*, which is declared as:

```
struct msg { char data[20];  
};
```

This declaration, and all other data structure and emulator routines, as well as stub routines (i.e., those you are to complete) are in the file, **prog2.c** (<http://gaia.cs.umass.edu/kurose/transport/prog2.c>). Your sending entity will thus receive data in 20-byte chunks from layer5; your receiving entity should deliver 20-byte chunks of correctly received data to layer5 at the receiving side.

Student Questions

Lab 3 (Cont)

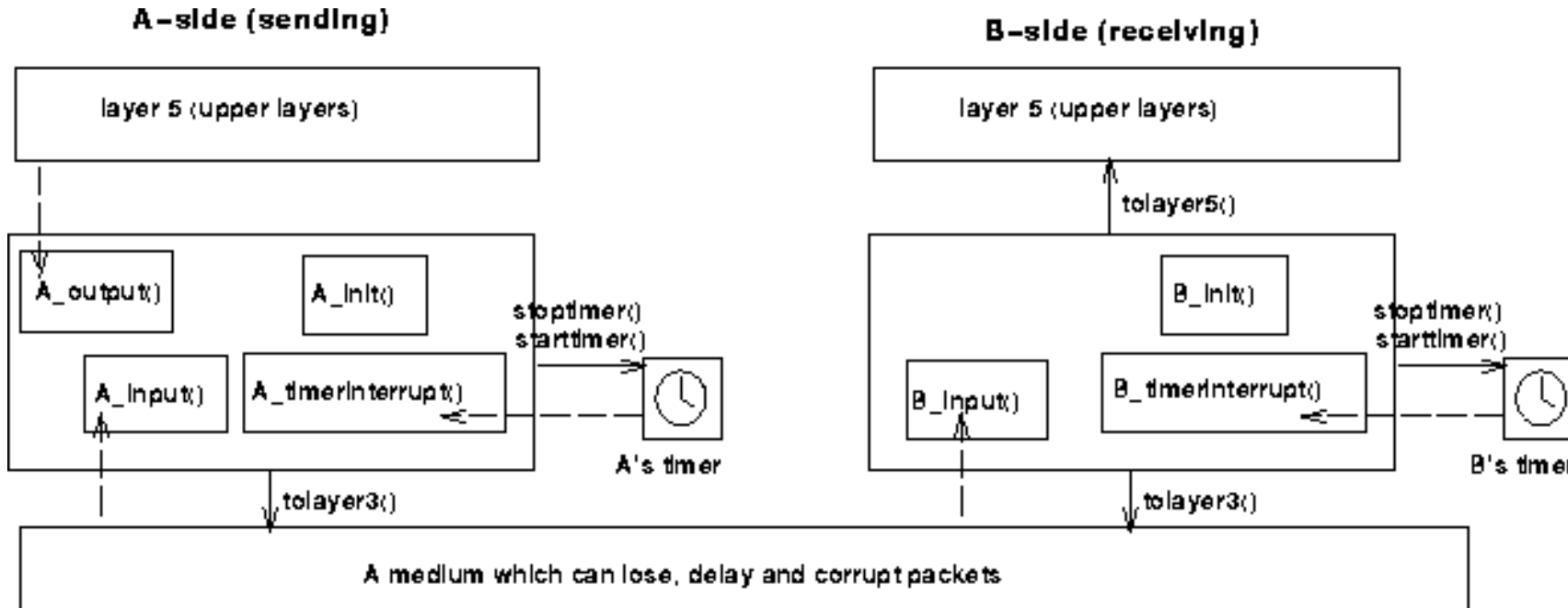


Figure Lab.3-1

Student Questions

Lab 3 (Cont)

The unit of data passed between your routines and the network layer is the *packet*, which is declared as:

```
struct pkt { int seqnum; int acknum;  
int checksum; char payload[20];  
};
```

Your routines will fill in the payload field from the message data passed down from layer5. The other packet fields will be used by your protocols to insure reliable delivery, as we've seen in class.

The routines you will write are detailed below. As noted above, such procedures in real-life would be part of the operating system, and would be called by other procedures in the operating system.

A_output(message), where message is a structure of type msg, containing data to be sent to the B-side. This routine will be called whenever the upper layer at the sending side (A) has a message to send. It is the job of your protocol to insure that the data in such a message is delivered in-order, and correctly, to the receiving side upper layer.

A_input(packet), where packet is a structure of type pkt. This routine will be called whenever a packet sent from the B-side (i.e., as a result of a tolayer3() being done by a B-side procedure) arrives at the A-side. packet is the (possibly corrupted) packet sent from the B-side.

A_timerinterrupt() This routine will be called when A's timer expires (thus generating a timer interrupt). You'll probably want to use this routine to control the retransmission of packets. See starttimer() and stoptimer() below for how the timer is started and stopped.

A_init() This routine will be called once, before any of your other A-side routines are called. It can be used to do any required initialization.

B_input(packet), where packet is a structure of type pkt. This routine will be called whenever a packet sent from the A-side (i.e., as a result of a tolayer3() being done by a A-side procedure) arrives at the B-side. packet is the (possibly corrupted) packet sent from the A-side.

B_init() This routine will be called once, before any of your other B-side routines are called. It can be used to do any required initialization.

Student Questions

Lab 3 (Cont)

Software Interfaces

The procedures described above are the ones that you will write. I have written the following routines which can be called by your routines:

starttimer(calling_entity,increment), where `calling_entity` is either 0 (for starting the A-side timer) or 1 (for starting the B side timer), and `increment` is a *float* value indicating the amount of time that will pass before the timer interrupts. A's timer should only be started (or stopped) by A-side routines, and similarly for the B-side timer. To give you an idea of the appropriate increment value to use: a packet sent into the network takes an average of 5 time units to arrive at the other side when there are no other messages in the medium.

stoptimer(calling_entity), where `calling_entity` is either 0 (for stopping the A-side timer) or 1 (for stopping the B side timer).

tolayer3(calling_entity,packet), where `calling_entity` is either 0 (for the A-side send) or 1 (for the B side send), and `packet` is a structure of type `pkt`. Calling this routine will cause the packet to be sent into the network, destined for the other entity.

tolayer5(calling_entity,message), where `calling_entity` is either 0 (for A-side delivery to layer 5) or 1 (for B-side delivery to layer 5), and `message` is a structure of type `msg`. With unidirectional data transfer, you would only be calling this with `calling_entity` equal to 1 (delivery to the B-side). Calling this routine will cause data to be passed up to layer 5.

Student Questions

Lab 3 (Cont)

The simulated network environment

A call to procedure `tolayer3()` sends packets into the medium (i.e., into the network layer). Your procedures `A_input()` and `B_input()` are called when a packet is to be delivered from the medium to your protocol layer.

The medium is capable of corrupting and losing packets. It will not reorder packets. When you compile your procedures and my procedures together and run the resulting program, you will be asked to specify values regarding the simulated network environment:

Number of messages to simulate. My emulator (and your routines) will stop as soon as this number of messages have been passed down from layer 5, regardless of whether or not all of the messages have been correctly delivered. Thus, you need **not** worry about undelivered or unACK'ed messages still in your sender when the emulator stops. Note that if you set this value to 1, your program will terminate immediately, before the message is delivered to the other side. Thus, this value should always be greater than 1.

Loss. You are asked to specify a packet loss probability. A value of 0.1 would mean that one in ten packets (on average) are lost.

Corruption. You are asked to specify a packet loss probability. A value of 0.2 would mean that one in five packets (on average) are corrupted. Note that the contents of payload, sequence, ack, or checksum fields can be corrupted. Your checksum should thus include the data, sequence, and ack fields.

Tracing. Setting a tracing value of 1 or 2 will print out useful information about what is going on inside the emulation (e.g., what's happening to packets and timers). A tracing value of 0 will turn this off. A tracing value greater than 2 will display all sorts of odd messages that are for my own emulator-debugging purposes. A tracing value of 2 may be helpful to you in debugging your code. You should keep in mind that *real* implementors do not have underlying networks that provide such nice information about what is going to happen to their packets!

Average time between messages from sender's layer5. You can set this value to any non-zero, positive value.

Note that the smaller the value you choose, the faster packets will be arriving to your sender.

Student Questions

Lab 3 (Cont)

The Alternating-Bit-Protocol Version of this lab.

You are to write the procedures, `A_output()`, `A_input()`, `A_timerinterrupt()`, `A_init()`, `B_input()`, and `B_init()` which together will implement a stop-and-wait (i.e., the alternating bit protocol, which we referred to as rdt3.0 in the text) unidirectional transfer of data from the A-side to the B-side. **Your protocol should use both ACK and NACK messages.**

You should choose a very large value for the average time between messages from sender's layer5, so that your sender is never called while it still has an outstanding, unacknowledged message it is trying to send to the receiver. I'd suggest you choose a value of 1000. You should also perform a check in your sender to make sure that when `A_output()` is called, there is no message currently in transit. If there is, you can simply ignore (drop) the data being passed to the `A_output()` routine.

You should put your procedures in a file called `prog2.c`. You will need the initial version of this file, containing the emulation routines we have written for you, and the stubs for your procedures. You can obtain this program from <http://gaia.cs.umass.edu/kurose/transport/prog2.c>.

This lab can be completed on any machine supporting C. It makes no use of UNIX features. (You can simply copy the `prog2.c` file to whatever machine and OS you choose).

We recommend that you should hand in a code listing, a screen shot of output, and an explanation of events in the output. For your sample output, your procedures might print out a message whenever an event occurs at your sender or receiver (a message/packet arrival, or a timer interrupt) as well as any action taken in response. You might want to hand in output for a run up to the point (approximately) when 10 messages have been ACK'ed correctly at the receiver, a loss probability of 0.1, and a corruption probability of 0.3, and a trace level of 2. You might want to annotate your printout showing how your protocol correctly recovered from packet loss and corruption.

Note 1: The code requires GCC 4.8.

Ubuntu 14.0.4 comes with GCC 4.8. So you may need to install Ubuntu 14.0.4 in a virtual machine.

Note 2: Some students have suggested to add the line “`#include <stdlib.h>`” and removing all instances of “`char *malloc()`.”

Student Questions

Lab 3 (Cont)

Helpful Hints and the like

Checksumming. You can use whatever approach for checksumming you want. Remember that the sequence number and ack field can also be corrupted. We would suggest a TCP-like checksum, which consists of the sum of the (integer) sequence and ack field values, added to a character-by-character sum of the payload field of the packet (i.e., treat each character as if it were an 8 bit integer and just add them together).

Note that any shared "state" among your routines needs to be in the form of global variables. Note also that any information that your procedures need to save from one invocation to the next must also be a global (or static) variable. For example, your routines will need to keep a copy of a packet for possible retransmission. It would probably be a good idea for such a data structure to be a global variable in your code. Note, however, that if one of your global variables is used by your sender side, that variable should **NOT** be accessed by the receiving side entity, since in real life, communicating entities connected only by a communication channel can not share global variables.

There is a float global variable called *time* that you can access from within your code to help you out with your diagnostics msgs.

START SIMPLE. Set the probabilities of loss and corruption to zero and test out your routines. Better yet, design and implement your procedures for the case of no loss and no corruption, and get them working first. Then handle the case of one of these probabilities being non-zero, and then finally both being non-zero.

Debugging. We'd recommend that you set the tracing level to 2 and put LOTS of printf's in your code while your debugging your procedures.

Random Numbers. The emulator generates packet loss and errors using a random number generator. Our past experience is that random number generators can vary widely from one machine to another. You may need to modify the random number generation code in the emulator we have supplied you. Our emulation routines have a test to see if the random number generator on your machine will work with our code. If you get an error message:

It is likely that random number generation on your machine is different from what this emulator expects. Please take a look at the routine `jimsrand()` in the emulator code. Sorry.

then you'll know you'll need to look at how random numbers are generated in the routine `jimsrand()`; see the comments in that routine.

Student Questions

Lab 3 Hints

- ❑ Some students received warning messages when trying to compile the C code. This can be fixed by adding the line
`#include <stdlib.h>`
and removing all instances of
`char *malloc();`
- ❑ The method `starttimer`, which schedules `timerinterrupt` to trigger, is the one to be careful of.
 - `starttimer` schedules `timerinterrupt` to trigger after some amount of time and have the following signature.
 - `starttimer(int calling_entity, float increment){ }`
Note that the 2nd parameter must be a **float**.
 - If `increment` is not cast to a float, it leads to unexpected behavior (triggering back to back to back to back...).
 - This is an issue of bit representation. C will take in the bits that represent an int and interpret them as if they were a float. This leads to a passing in a much smaller value than expected.

Student Questions

Lab 3 Hints (Cont)

- i.e., the following will trigger back to back to back...
- `starttimer(A_is_calling_entity, 2000);`
// Leads to lots of repeated timerinterrupts
- The following will behave as expected
- `starttimer(A_is_calling_entity, (float)2000);` // Casting to float will fix the issue
- If you have a Macbook with M chip and cannot install a AMD iso file, consider installing Ubuntu in a VirtualBox VM as follows:
- ✓ Go to VirtualBox Downloads and download the "macOS / Apple Silicon hosts" version that works with your M-chip MacBook.
- ✓ Install VirtualBox.
- ✓ Download Ubuntu from <https://ubuntu.com/download/server/arm> , and select "Download 24.04.1 LTS". But if you can access a Windows computer, that would be easier—you can directly download Ubuntu 14.04 from <https://releases.ubuntu.com/14.04/>)

Student Questions

Lab 3 Hints (Cont)

- ✓ Follow the steps in https://www.youtube.com/watch?v=LjL_N0OZxvY for Mac M chip (or <https://www.youtube.com/watch?v=DhVjgI57Ino> for Windows) to install Ubuntu in VirtualBox. (Note: You might not encounter the “installing Ubuntu 24.10 server” part in the video for Mac.)
- ✓ Once the setup is done, run `gcc --version` in the terminal to check the installed GCC version (it should be something like GCC 13.3 if you choose to use MacOS).
- ✓ In general, newer GCC versions (like 13.3) are backward-compatible, meaning if your code compiles on GCC 4.8, it should work fine on GCC 13.3 too.

Student Questions

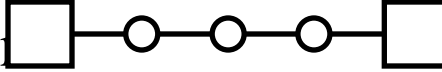


TCP

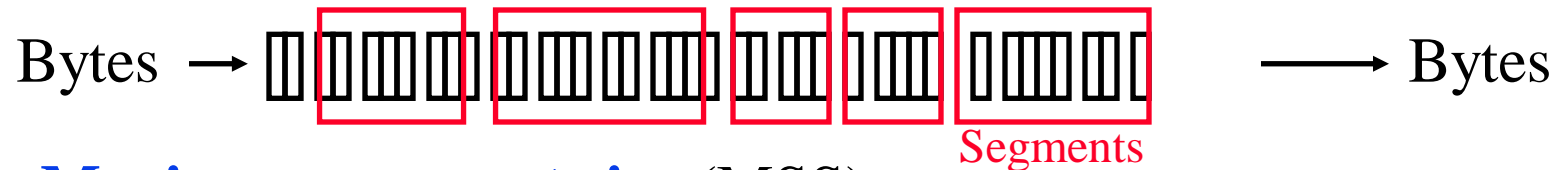
1. TCP Header Format, Options, Checksum
2. TCP Connection Management
3. Round Trip Time Estimation
4. Principles of Congestion Control
5. Slow Start Congestion Control

Student Questions

Key Features of TCP

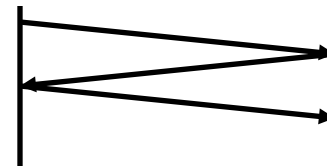
❑ **Point-to-Point:** One sender, one destination 

❑ **Byte Stream:** No message boundaries.
TCP makes “segments”

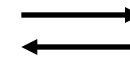


❑ **Maximum segment size (MSS)**

❑ **Connection Oriented:** Handshake to initialize states before data exchange

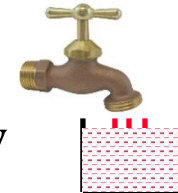


❑ **Full Duplex:** Bidirectional data flow in one connection



❑ **Reliable:** In-order byte delivery

❑ **Flow control:** To avoid destination buffer overflow



❑ **Congestion control:** To avoid network router buffer overflow

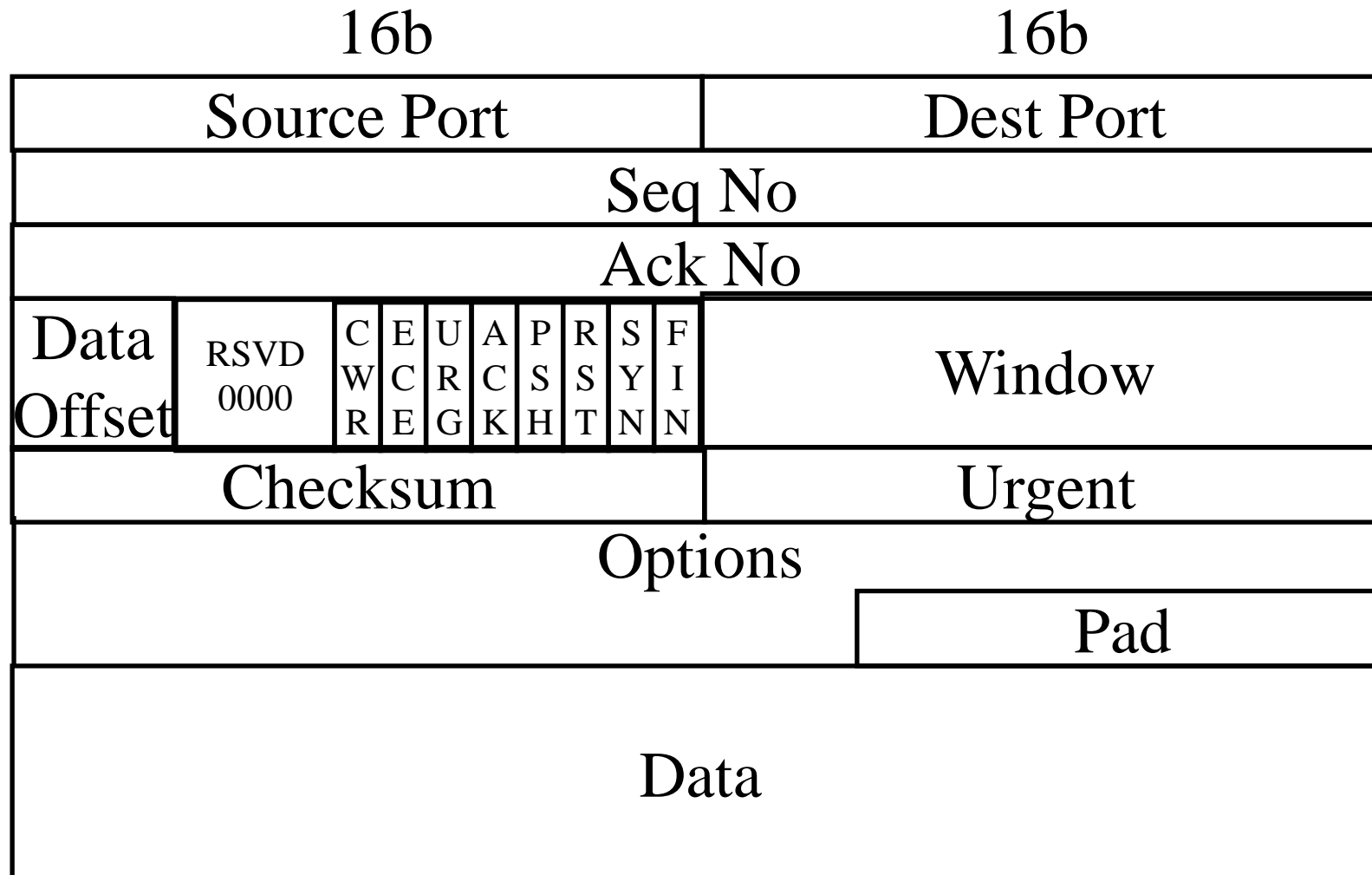
Student Questions

TCP

- ❑ Transmission Control Protocol
- ❑ Key Services:
 - **Send**: Please send when convenient
 - **Data stream push**: Source TCP, please send it now.
Set on the last packet of an application message.
 - **Urgent data signaling**: Destination TCP, please give this urgent data to the user out-of-band.
Generally used for CTRL-C.

Student Questions

TCP Segment Format (Cont)



Student Questions

TCP Header Fields

- ❑ **Source Port** (16 bits): Identifies source user process
- ❑ **Destination Port** (16 bits)
21 = FTP, 23 = Telnet, 53 = DNS, 80 = HTTP, ...
- ❑ **Sequence Number** (32 bits): The sequence number of the first data octet in this segment (**except when SYN is present**). If SYN is present the sequence number is the initial sequence number (ISN) and the first data octet is ISN+1.*
- ❑ **Ack number** (32 bits): Next byte expected
- ❑ **Data offset** (4 bits): Number of 32-bit words in the header
- ❑ **Reserved** (4 bits)

Student Questions

*Ref: IETF RFC 793, "Transmission Control Protocol," September 1981, 91 pp., Obsoleted by RFC 9293.

TCP Header (Cont)

- ❑ **Control** (8 bits): Congestion Window Reset
Explicit Congestion Experienced
Urgent pointer field significant,
Ack field significant,
Push function,
Reset the connection,
Synchronize the sequence numbers,
No more data from sender



- ❑ **Window** (16 bits):
Will accept [Ack] to [Ack]+[window]-1

Student Questions

TCP Header (Cont)

- ❑ **Checksum** (16 bits): covers the segment plus a pseudo-header. Includes the following fields from the IP header: source and dest adr, protocol, and segment length. Protects from IP misdelivery.
- ❑ **Urgent pointer** (16 bits): Points to the byte following urgent data. It Lets the destination knows how much data it should deliver right away out-of-band.
- ❑ **Options** (variable):
Max segment size (does not include TCP header, default 536 bytes), Window scale factor, Selective Ack permitted, Timestamp, No-Op, End-of-options.

Student Questions

TCP Options

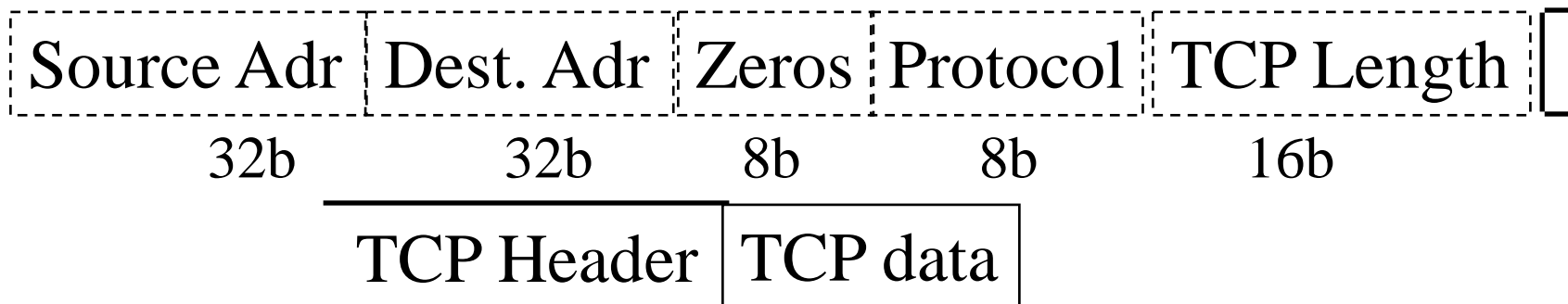
Kind	Length	Meaning
0	1	End of Valid options in header
1	1	No-op
2	4	Maximum Segment Size
3	3	Window Scale Factor
8	10	Timestamp

- ❑ **End of Options:** Stop looking for further option
- ❑ **No-op:** Ignore this byte. Used to align the next option on a 4-byte word boundary
- ❑ **Max Segment Size (MSS):** Does not include TCP header

Student Questions

TCP Checksum

- ❑ The checksum is the 16-bit one's complement of the one's complement sum of a pseudo header of information from the IP header, the TCP header, and the data, padded with zero octets at the end (if necessary) to make a multiple of two octets.
- ❑ The checksum field is filled with zeros initially.
- ❑ TCP length (in octet) is not transmitted but used in calculations.
- ❑ Efficient implementation in RFC1071.

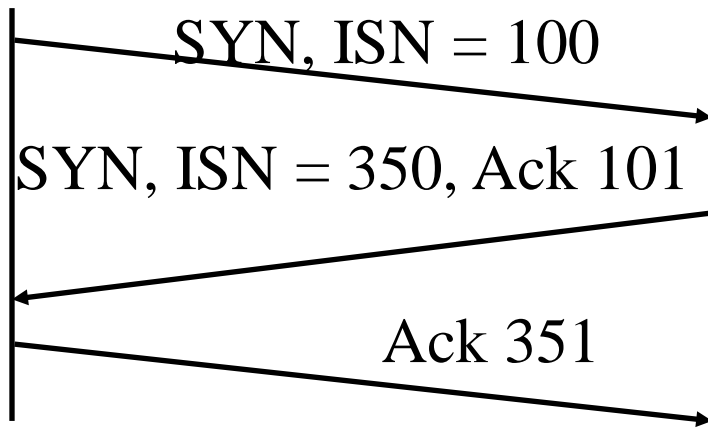


Student Questions

TCP Connection Management

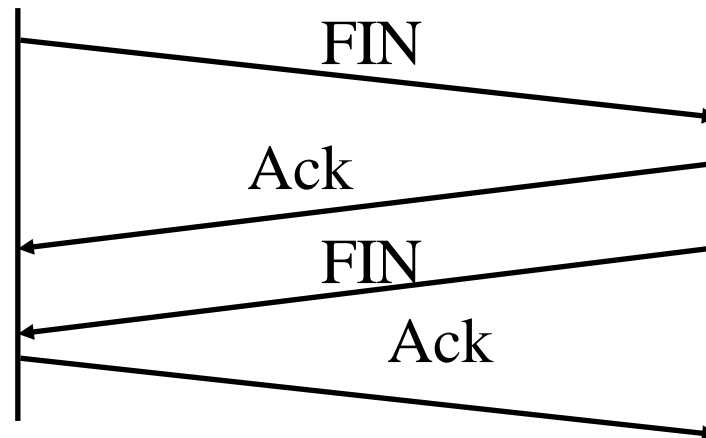
❑ Connection Establishment

- Three-way handshake
- SYN flag set
⇒ Request for connection



❑ Connection Termination

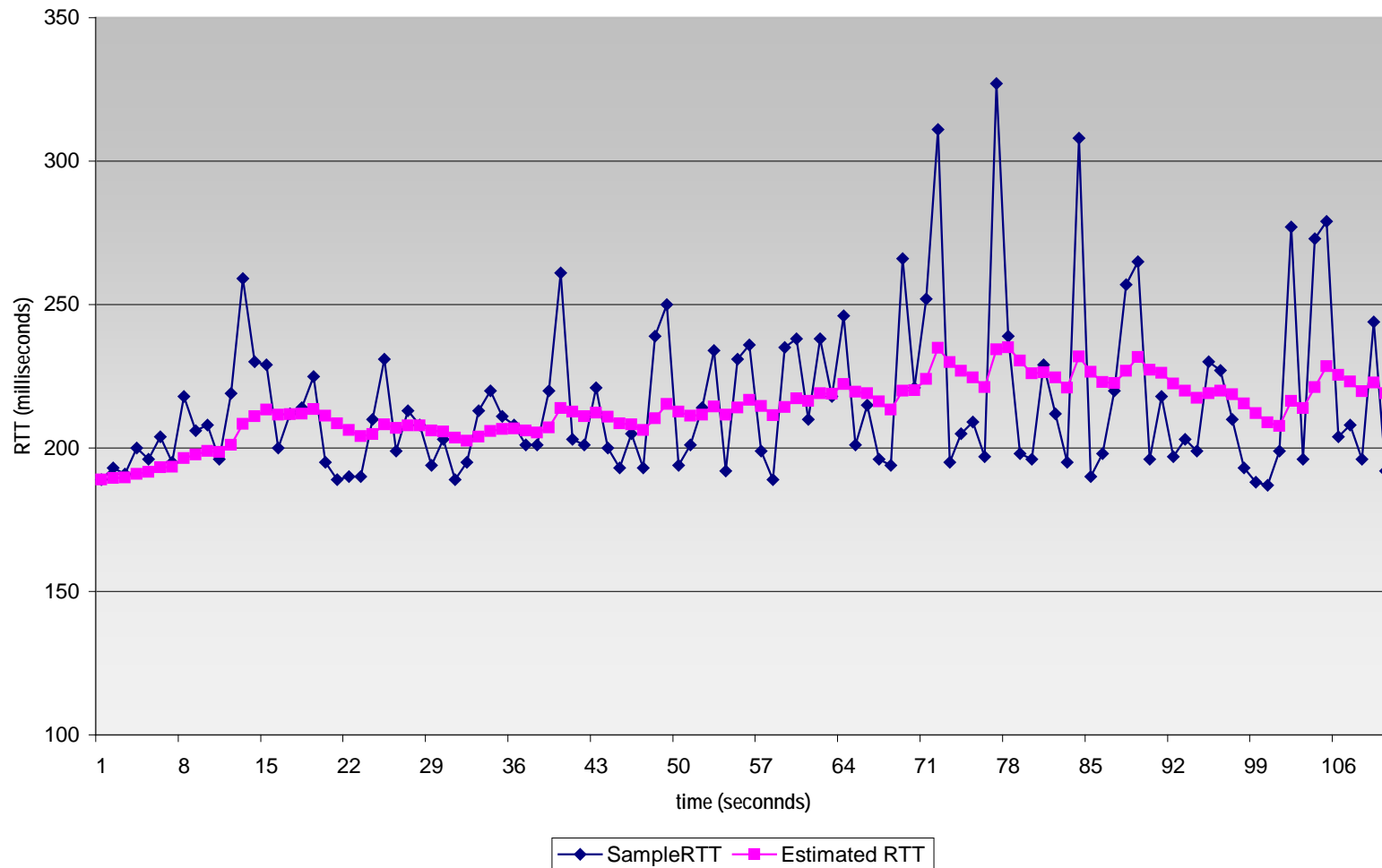
- Close with a FIN flag set
- Abort



Student Questions

Example RTT estimation:

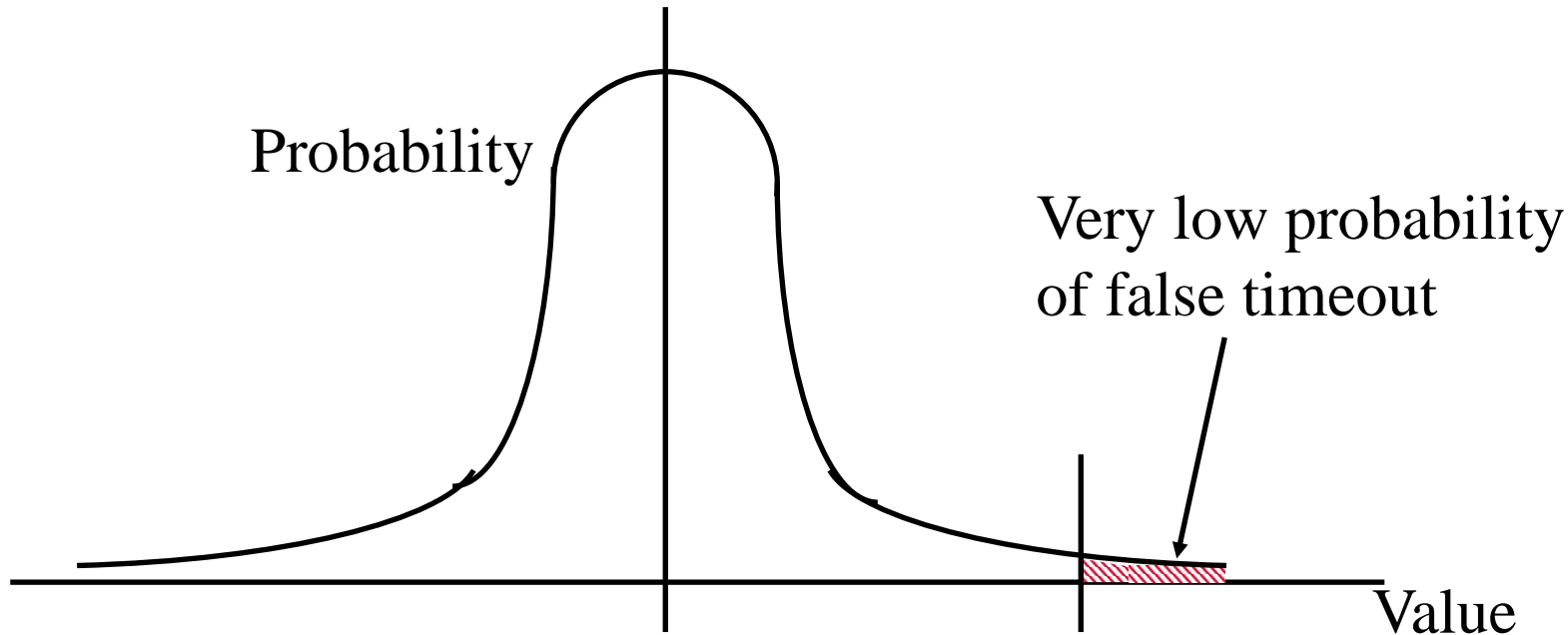
RTT: gaia.cs.umass.edu to fantasia.eurecom.fr



Student Questions

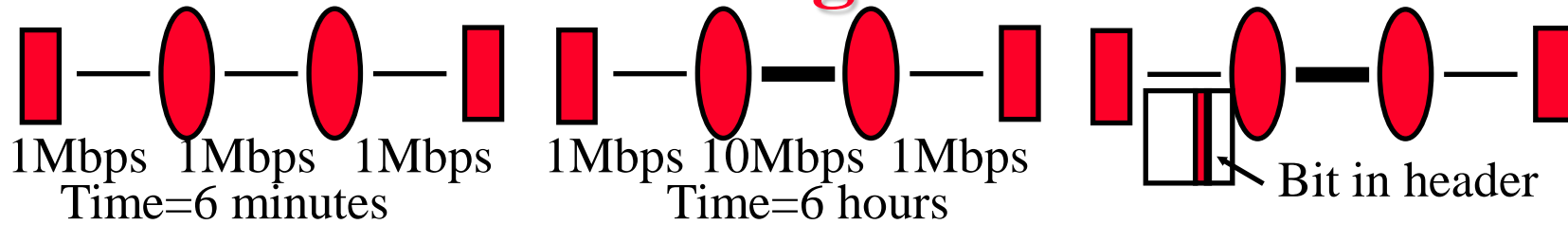
Round Trip Time Estimation

- ❑ Measured round trip time (SampleRTT) is very random.
- ❑ $\text{EstimatedRTT} = (1 - \alpha)\text{EstimatedRTT} + \alpha \text{ SampleRTT}$
- ❑ $\text{DevRTT} = (1 - \beta)\text{DevRTT} + \beta |\text{SampleRTT} - \text{EstimatedRTT}|$
- ❑ $\text{TimeoutInterval} = \text{EstimatedRTT} + 4 \text{ DevRTT}$



Student Questions

Our Research on Congestion Control



- ❑ Early 1980s, Digital Equipment Corporation (DEC) introduced Ethernet products
- ❑ Noticed that throughput goes down with a higher-speed link in the middle (because there are no congestion mechanisms in TCP)
- ❑ Results:
 1. Timeout \Rightarrow Congestion
 \Rightarrow Reduce the TCP window to one on a timeout [Jain 1986]
 2. Routers should set a bit when congested (DECbit).
[Jain, Ramakrishnan, Chiu 1988]
 3. Introduced the term “Congestion Avoidance.”
 4. Additive increase and multiplicative decrease (AIMD principle)
[Chiu and Jain 1989]
- ❑ There were presented to IETF in 1986.
 \Rightarrow Slow-start based on Timeout and AIMD [Van Jacobson 1988]

Student Questions

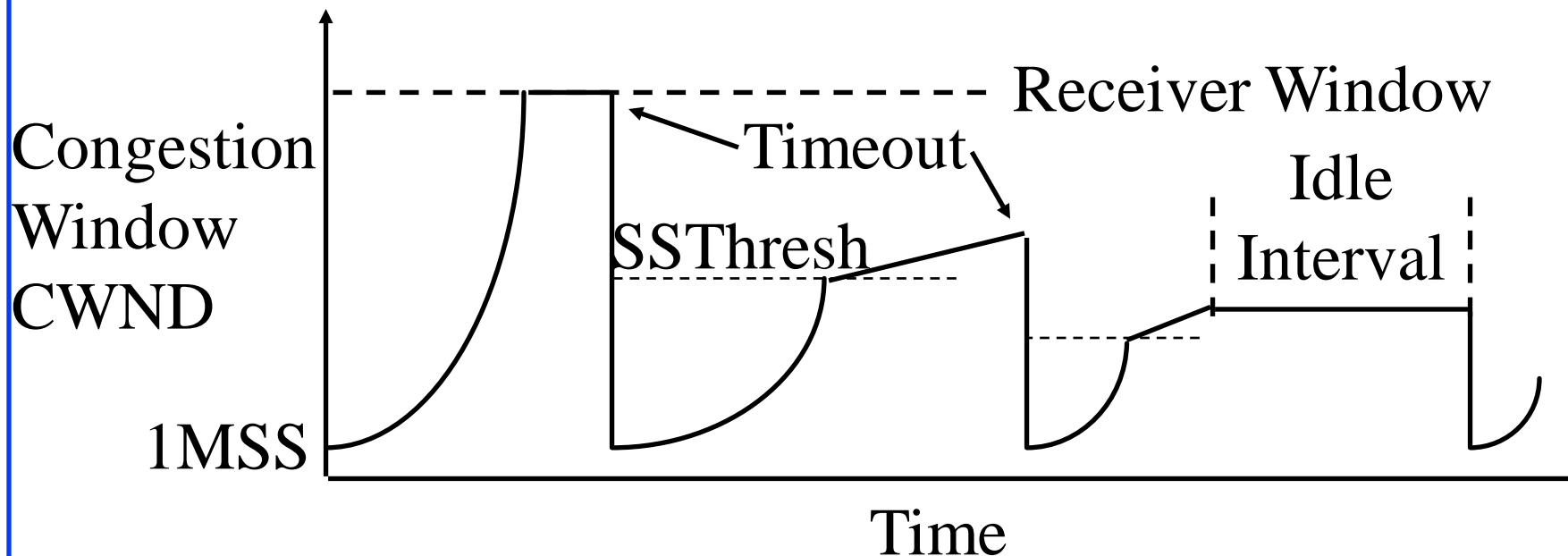
Slow Start Congestion Control

- ❑ Window = Flow control avoids destination overrun
- ❑ Need congestion control to avoid network overrun
- ❑ The sender maintains two windows:
 - Credits from the receiver
 - Congestion window from the network
 - The congestion window is always less than the receiver window
- ❑ Starts with a congestion window (CWND) of 1 max segment size (MSS)
 - ⇒ Do not disturb existing connections too much.
- ❑ Increase CWND by 1 MSS every time an ack is received
- ❑ Assume CWND is in bytes

Student Questions

Slow Start (Cont)

- If segments lost, remember slow start threshold (SSThresh) to $CWND/2$
Set $CWND$ to 1 MSS
Increment by 1MSS per ack until SSThresh
Increment by $1 \text{ MSS} * \text{MSS} / CWND$ per ack afterwards



Student Questions

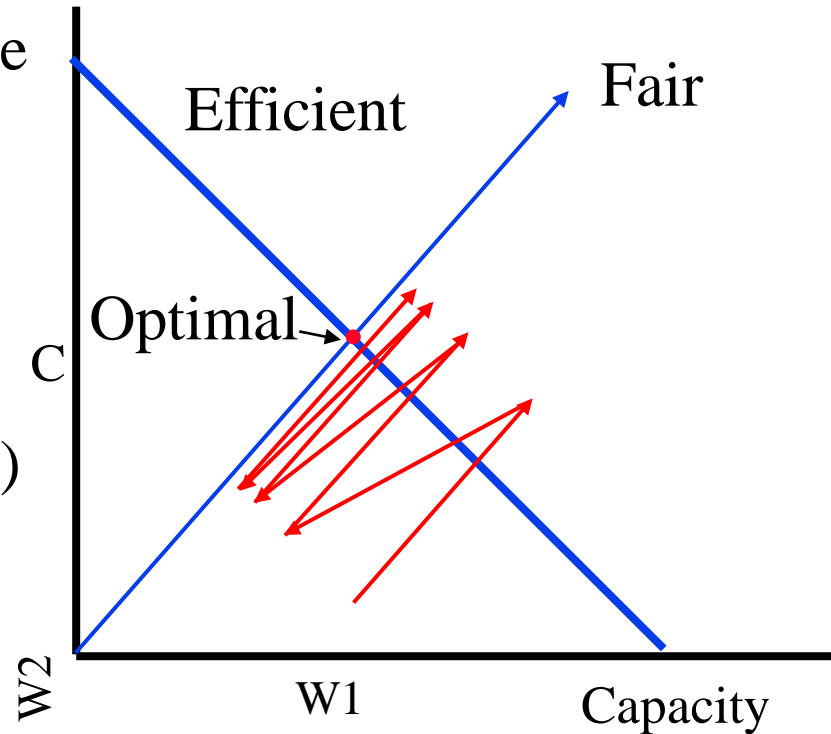
Slow Start (Cont)

- ❑ At the beginning, SSThresh = **Arbitrarily high value**
- ❑ After a long idle period (exceeding one round-trip time), reset the congestion window to one.
- ❑ If CWND is W MSS, W acks are received in one round trip.
- ❑ Below SSThresh, CWND is increased by 1MSS on every ack
 - ⇒ CWND increases to $2W$ MSS in one round trip
 - ⇒ CWND increases exponentially with timeExponential growth phase is also known as “*Slow start*” phase
- ❑ Above SSThresh, CWND is increased by $MSS/CWND$ on every ack
 - ⇒ CWND increases by 1 MSS in one round trip.
 - ⇒ CWND increases linearly with time.The linear growth phase is known as “*congestion avoidance*” phase

Student Questions

AIMD Principle

- ❑ Additive Increase, Multiplicative Decrease
- ❑ $W1+W2 = \text{Capacity}$
 \Rightarrow Efficiency,
 $W1=W2 \Rightarrow$ Fairness
- ❑ $(W1, W2)$ to $(W1+\Delta W, W2+\Delta W)$
 \Rightarrow Linear increase (45° line)
- ❑ $(W1, W2)$ to $(kW1, kW2)$
 \Rightarrow Multiplicative decrease
(line through origin)



Student Questions

Ref: D. Chiu and Raj Jain, "Analysis of the Increase/Decrease Algorithms for Congestion Avoidance in Computer Networks," Journal of Computer Networks and ISDN, Vol. 17, No. 1, June 1989, pp. 1-14,

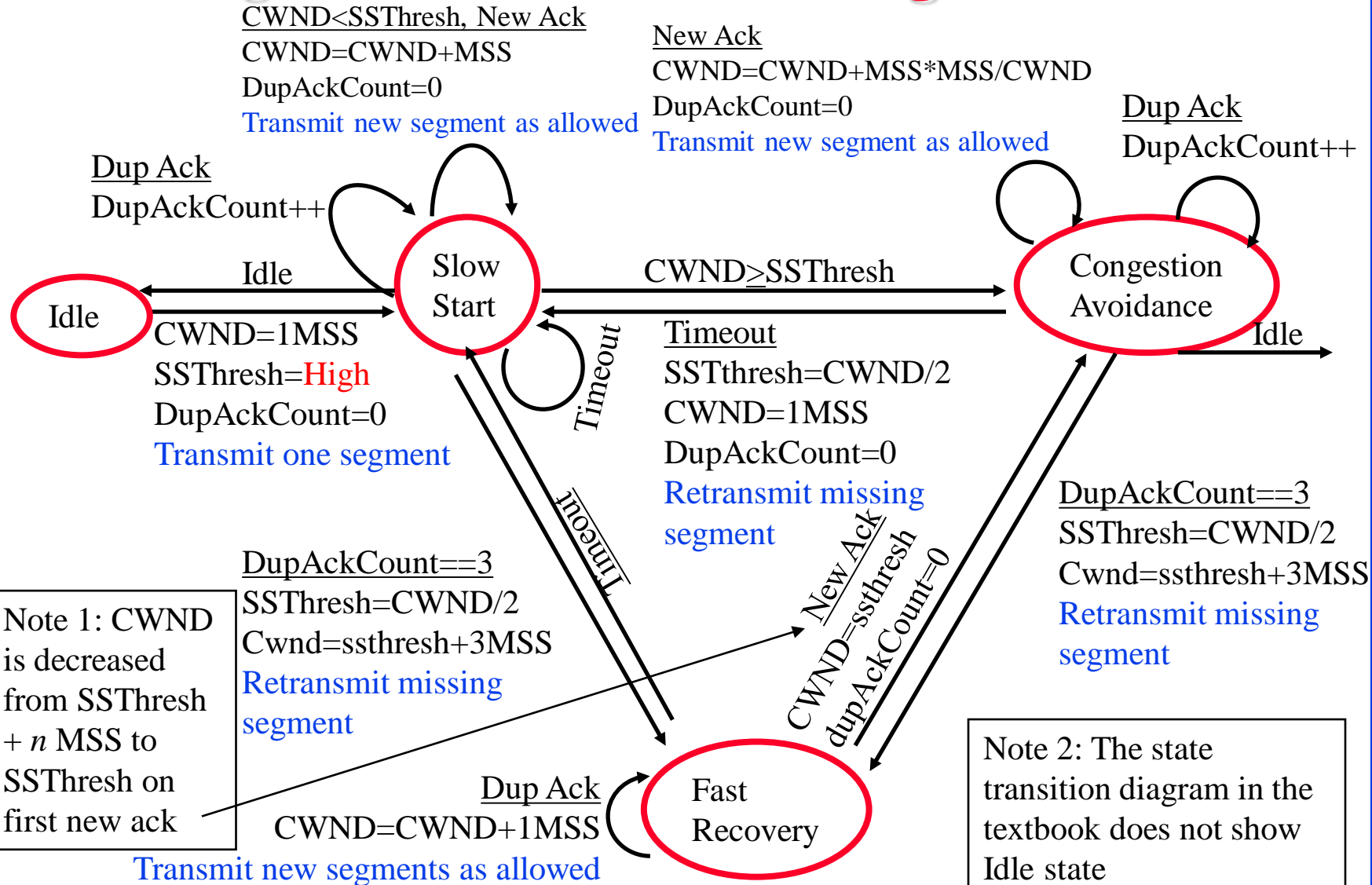
http://www.cse.wustl.edu/~jain/papers/cong_av.htm

Fast Retransmit

- ❑ Optional – implemented in TCP Reno (Earlier version was TCP Tahoe)
- ❑ Duplicate Ack indicates a lost/out-of-order segment
- ❑ On receiving 3 duplicate acks (4th ack for the same segment):
 - Enter Fast Recovery mode
 - ❑ Retransmit missing segment
 - ❑ Set $SSThresh = CWND/2$
 - ❑ Set $CWND = SSThresh + 3 \text{ MSS}$ (**Note: CWND is inflated**)
 - ❑ Every subsequent duplicate ack: $CWND = CWND + 1 \text{ MSS}$
 - When a new ack (not a duplicate ack) is received
 - ❑ Exit fast recovery
 - ❑ Set $CWND = SSTHRESH$ (**Note: CWND is deflated back**)

Student Questions

TCP Congestion Control State Diagram



Note 1: CWND is decreased from SSThresh + n MSS to SSThresh on first new ack

Note 2: The state transition diagram in the textbook does not show Idle state

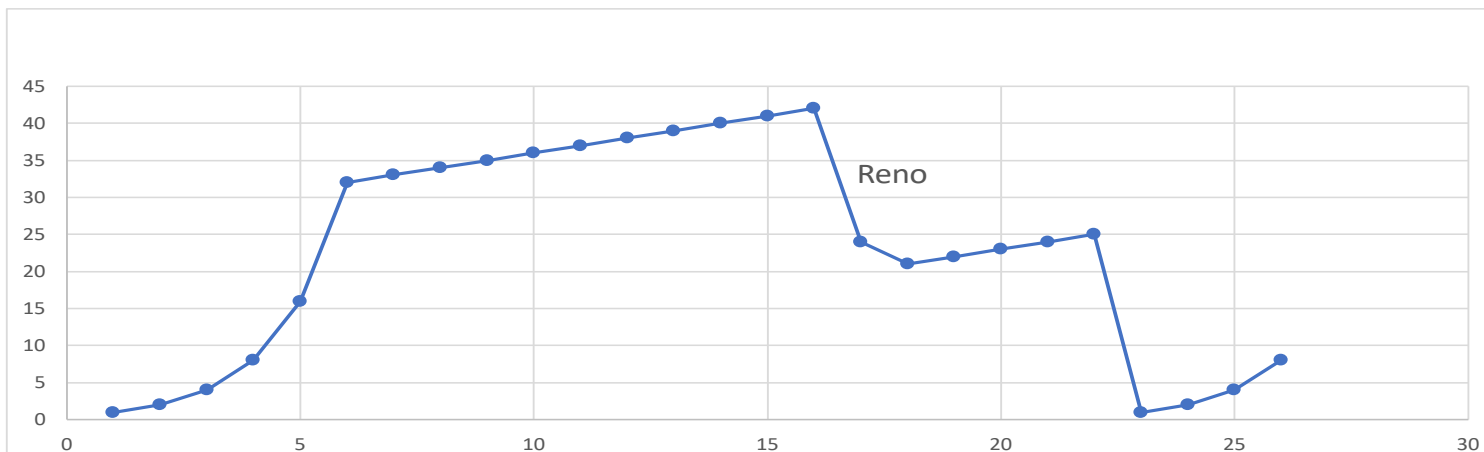
Student Questions

Homework 3C

[30 points] Consider the Figure below. Assuming TCP Reno is the protocol experiencing the behavior shown above, answer the following questions.

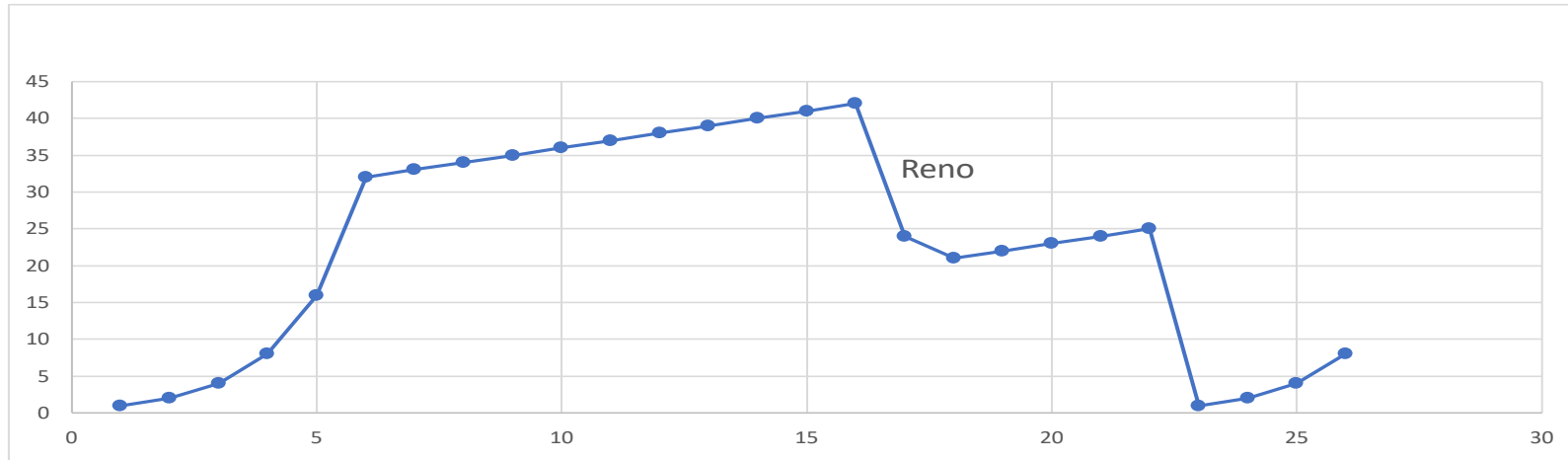
- A. Identify the first interval of time when TCP slow start is operating.
- B. Identify the 2nd interval of time when TCP slow start is operating.
- C. Identify the first interval of time when TCP congestion avoidance is operating.
- D. Identify the 2nd interval of time when TCP congestion avoidance is operating.

Round	CWND Reno
1	1
2	2
3	4
4	8
5	16
6	32
7	33
8	34
9	35
10	36
11	37
12	38
13	39
14	40
15	41
16	42
17	24
18	21
19	22
20	23
21	24
22	25
23	1
24	2
25	4
26	8



Student Questions

Homework 3C (Cont)



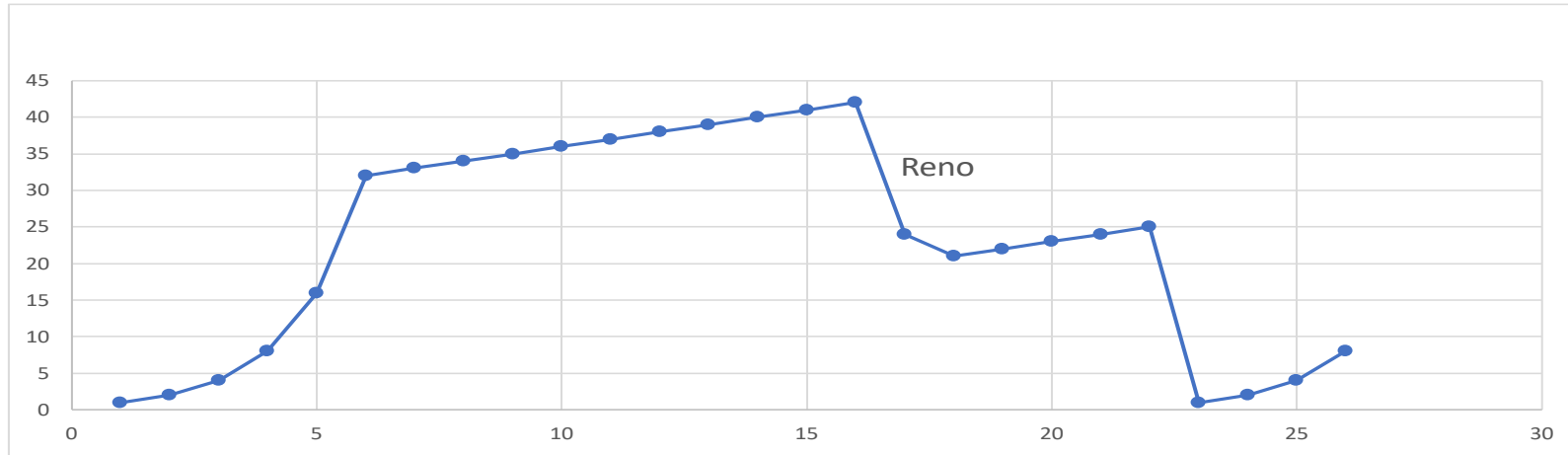
Student Questions

E. After the 16th transmission round, is segment loss detected by a timeout?

F. After the 22nd transmission round, is segment loss detected by a triple duplicate ACK or by a timeout?

G. What is the initial value of *ssthresh* at the first transmission round?

Homework 3C (Cont)



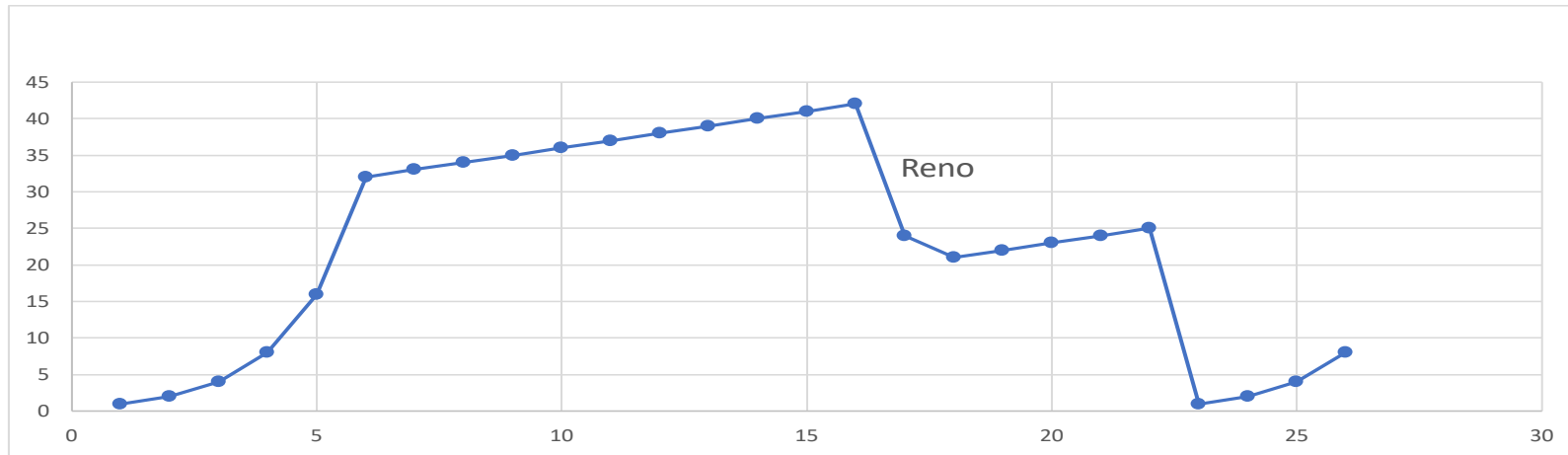
H. What is the value of $ssthresh$ at the 18th transmission round?

I. What is the value of $ssthresh$ at the 24th transmission round?

J. During what transmission round is the 70th segment sent?

Student Questions

Homework 3C (Cont)

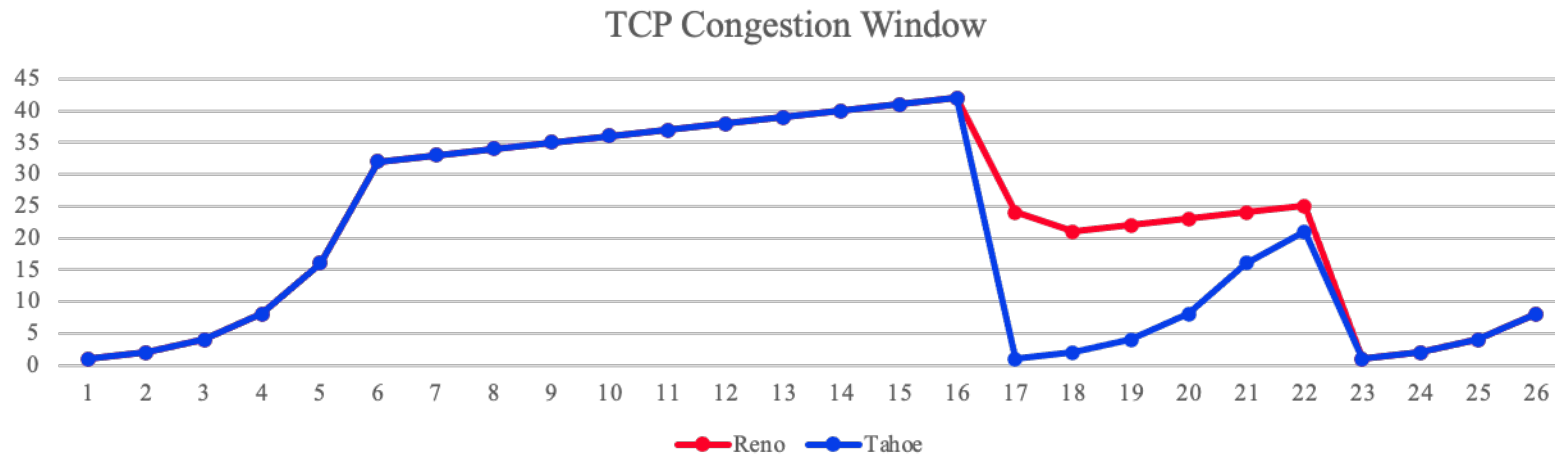


K. Assuming a packet loss is detected after the 26th round by the receipt of a triple duplicate ACK, what will be the value of the threshold?

L. What will be the congestion window size in K above?

Student Questions

Homework 3C (Cont)



Student Questions

M. Suppose TCP Tahoe is used (instead of TCP Reno), and assume that triple duplicate ACKs are received at the 16th round. What would the ssthresh be in the 19th round? (Hint: You need to calculate CWND in 17-18th rounds first. It will be different from that shown for Reno.)

N. What would be the congestion window size at the 19th round in the above case?

O. Again, suppose TCP Tahoe is used, and there is a timeout at the end of the 22nd round. How many packets have been sent out from the 17th round to the 22nd round, inclusive? (Hint: You need to calculate CWND in 17-22nd rounds first.)

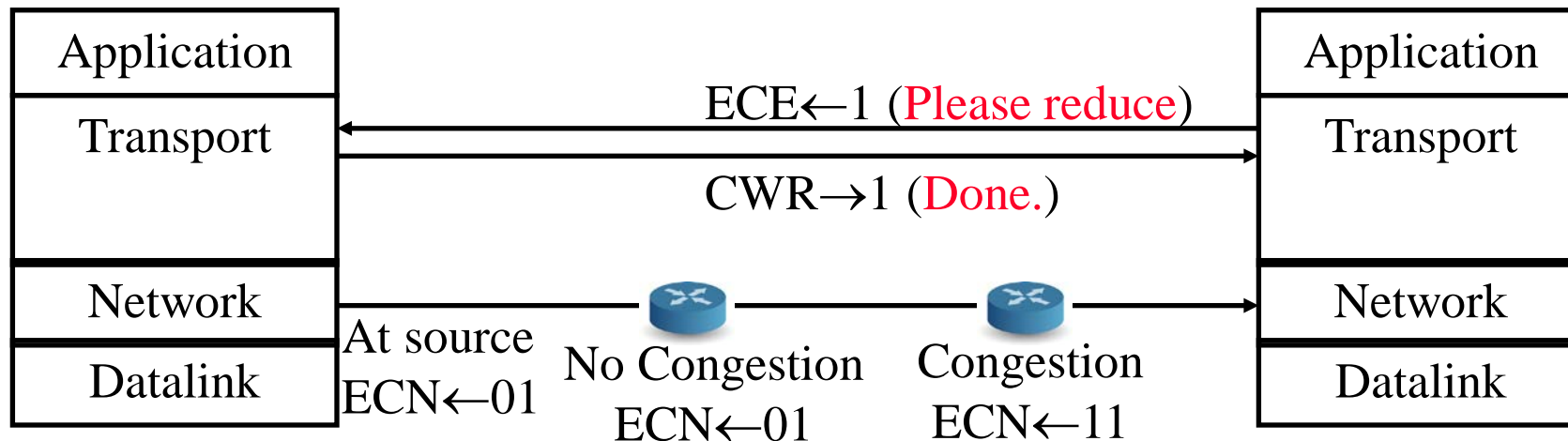
TCP Average Throughput

- Average Throughput = $\frac{1.22 \text{ MSS}}{\text{RTT} \sqrt{P}}$
- Here, P = Probability of Packet loss.
- Note 1: The formula is an approximation that does not apply at P=0 or P=1. At P=1, the throughput is zero. At P=0, the throughput is $\min\{1, (\text{receiver Window}/\text{RTT})\}$
- Note 2: The textbook has a different formula. Numerous such formulas are in literature. All under different assumptions and some empirical ones. This formula is not exact or universally agreed.

Student Questions

Explicit Congestion Notification (ECN)

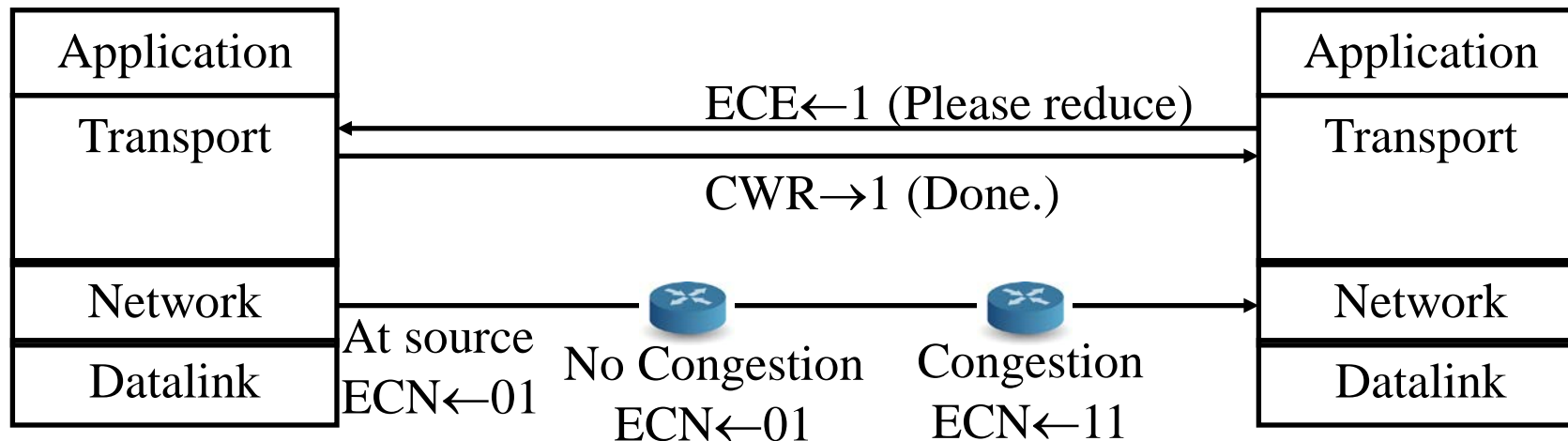
- Explicit congestion notification (ECN) is based on our DECbit research.
 - Two bits in IP Header: Last two bits of traffic class (Next chapter)
 - Two bits in the TCP header: ECE and CWR
- IP Bits:
 - 00: Transport is not capable of ECN (e.g., UDP)
 - 01 or 10: ECN capable transport
 - 11: Congestion Experienced
- TCP Bits:
 - 01: **Reduce** cong Window
 - 10: Cong window **reduced**
 - 11: Both
- When a router encounters congestion, instead of dropping the datagram, it marks the two bits as “11” congestion experienced.



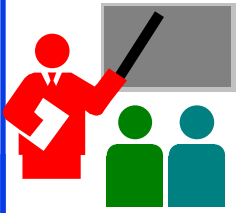
Student Questions

ECN (Cont)

- ❑ ECN uses two bits in the TCP header: ECE and CWR
- ❑ On receiving “CE” code point, the receiver sends “ECN Echo (ECE)” flag in the TCP header
- ❑ On seeing the ECE flag, the source reduces its congestion window, and sets “Congestion Window Reduced (CWR) flag in the outgoing segment
- ❑ On receiving “CWR” flag, the receiver, stops setting ECE bit



Student Questions



TCP: Summary

1. TCP uses **port numbers** for multiplexing
2. TCP provides reliable **full-duplex** connections.
3. TCP is **stream** based and has **window flow control**
4. **Slow-start congestion control** works on timeout
5. **Explicit congestion notification** works using ECN bits

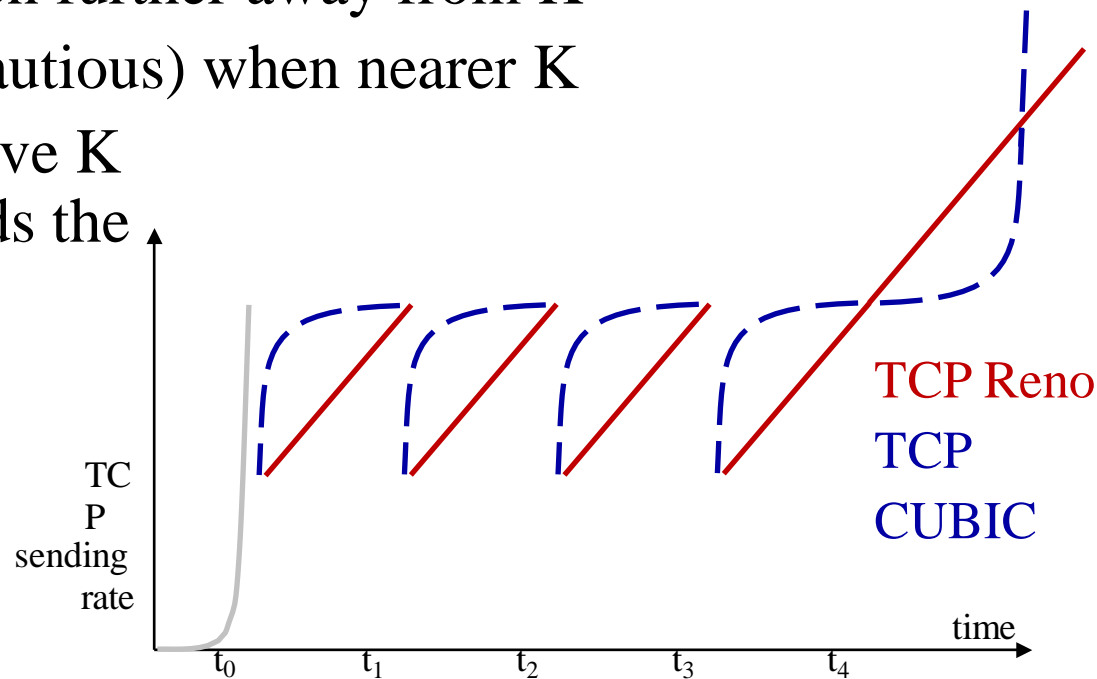
Student Questions

Read Sections 3.5, 3.6, and 3.7. Do R14-R19, P25, P26, P31, P32, P33, P40

TCP CUBIC

- ❑ K: Estimate of time when TCP window size will reach W_{\max} . K is a tunable parameter.
- ❑ Increase W as a function of the *cube* of the distance between the current time and K
- ❑ Larger increases when further away from K
- ❑ Smaller increases (cautious) when nearer K
- ❑ Increases slowly above K and then quickly finds the new limit
- ❑ TCP QUBIC is default in Linux
- ❑ Most popular TCP for Web servers

Read textbook p. 271-273



Student Questions

Ref: L. Xu, S. Ha, A. Zimmermann, L. Eggert, R. Scheffenger, "CUBIC for Fast Long Distance Networks," RFC 8312, Feb 2018,

updated by RFC 9438.

Washington University in St. Louis

<http://www.cse.wustl.edu/~jain/cse4703-26/>

©2026 Raj Jain

Summary



1. **Multiplexing/demultiplexing** by a combination of source and destination IP addresses and port numbers.
2. **Longer distance or higher speed**
⇒ A larger α ⇒ Larger window is better.
3. Window flow control is better for long-distance or high-speed networks
4. UDP is connectionless and simple.
No flow/error control. Has error **detection.**
5. TCP provides **full-duplex** connections with flow/error/**congestion** control.

Student Questions

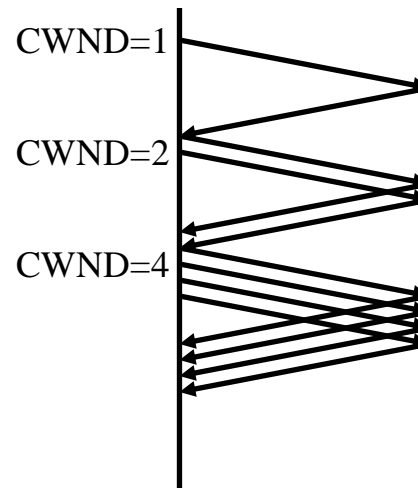
Optional Homework 3D

Try but do not submit.

A TCP entity opens a connection and uses slow start.

Approximately how many round-trip times are required before TCP can send N segments. Assume no timeout.

Hint:



Student Questions

Acronyms

- ❑ ACK ACKnowledgement
- ❑ AIMD Additive increase and multiplicative decrease
- ❑ ARQ Automatic Repeat Request
- ❑ CE Congestion Experienced
- ❑ CRC Cyclic Redundancy Check
- ❑ CWND Congestion Window
- ❑ CWR Congestion Window Reduced
- ❑ DA Destination Address
- ❑ DEC Digital Equipment Corporation
- ❑ DECbit DEC's bit based congestion scheme
- ❑ DevRTT Deviation of RTT
- ❑ DNS Domain Name System
- ❑ DP Destination Port
- ❑ ECE Explicit Congestion Experienced
- ❑ ECN Explicit Congestion Notification
- ❑ FIN Final

Student Questions

Acronyms (Cont)

- ❑ FTP File Transfer Protocol
- ❑ GBN Go-Back N
- ❑ HTTP Hyper-Text Transfer Protocol
- ❑ IETF Internet Engineering Task Force
- ❑ IP Internet Protocol
- ❑ ISN Initial Sequence Number
- ❑ kB Kilo-Byte
- ❑ MSS Maximum segment size
- ❑ PBX Private Branch Exchange
- ❑ PSH Push
- ❑ RFC Request for Comments
- ❑ RM Resource Management
- ❑ RST Reset
- ❑ RTT Round-Trip Time
- ❑ SA Source Address
- ❑ SACK Selective Acknowledgement

Student Questions

Acronyms (Cont)

- ❑ SMTP Simple Mail Transfer Protocol
- ❑ SP Source Port
- ❑ SStresh Slow Start Threshold
- ❑ SYN Synchronization
- ❑ SYNACK SYN Acknowledgement
- ❑ TCP Transmission Control Protocol
- ❑ UDP User Datagram Protocol
- ❑ URG Urgent
- ❑ VCI Virtual Circuit Identifiers

Student Questions

Scan This to Download These Slides



Raj Jain

<http://rajjain.com>

Student Questions

http://www.cse.wustl.edu/~jain/cse4703-26/i_3tcp.htm

Related Modules



CSE 567: The Art of Computer Systems Performance Analysis

https://www.youtube.com/playlist?list=PLjGG94etKypJEKjNAa1n_1X0bWWNyZcof

CSE473S: Introduction to Computer Networks (Fall 2011),

https://www.youtube.com/playlist?list=PLjGG94etKypJWOSPMh8Azcg5e_10TiDw



CSE 570: Recent Advances in Networking (Spring 2013)

<https://www.youtube.com/playlist?list=PLjGG94etKypLHyBN8mOgwJLHD2FFIMGq5>

CSE571S: Network Security (Spring 2011),

<https://www.youtube.com/playlist?list=PLjGG94etKypKvzfVtutHcPFJXumyyg93u>



Video Podcasts of Prof. Raj Jain's Lectures,

<https://www.youtube.com/channel/UCN4-5wzNP9-ruOzQMs-8NUw>

Student Questions