Module 2
Transport Layer Protocols
Please note: Most of these slides come from this book. Note their copyright notice below…

A note on the use of these ppt slides:
We’re making these slides freely available to all (faculty, students, readers). They’re in PowerPoint form so you 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) in substantially unaltered form, that you mention their source (after all, we’d like people to use our book!)
❖ If you post any slides in substantially unaltered form 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.

Thanks and enjoy! JFK/KWR

All material copyright 1996-2010
J.F Kurose and K.W. Ross, All Rights Reserved
Transport services & protocols

- provide *logical communication* between app processes running on different hosts
- transport protocols run in end systems
  - send side: breaks app messages into *segments*, passes to network layer
  - rcv side: reassembles segments into messages, passes to app layer
- more than one transport protocol available to apps
  - Internet: TCP and UDP
Transport vs. network layer

- **network layer**: enables logical communication between hosts
  - your laptop and UWaterloo’s computer
- **transport layer**: enables logical communication between processes
  - your laptop’s Web browser and UWaterloo’s Web server
  - your laptop’s video player and CBC’s video server
  - relies on, enhances, network layer services
    - Enhancements depend on the particular protocol
Transport Layer is End-to-End
# Internet Protocols

<table>
<thead>
<tr>
<th>Application</th>
<th>FTP</th>
<th>Telnet</th>
<th>NFS</th>
<th>SMTP</th>
<th>HTTP</th>
<th>…</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transport</td>
<td>TCP</td>
<td>UDP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Network</td>
<td>IP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Link</td>
<td>X.25</td>
<td>Ethernet</td>
<td>Packet</td>
<td>ATM</td>
<td>FDDI</td>
<td>…</td>
</tr>
<tr>
<td>Physical</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Internet Apps: Their Protocols & Transport Protocols

<table>
<thead>
<tr>
<th>Applications</th>
<th>Data Loss</th>
<th>Throughput</th>
<th>Time Sensitive</th>
<th>Application Layer Protocol</th>
<th>Transport Protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Email</td>
<td>No loss</td>
<td>Elastic</td>
<td>No</td>
<td>smtp</td>
<td>TCP</td>
</tr>
<tr>
<td>remote terminal access</td>
<td>No loss</td>
<td>Elastic</td>
<td>Yes</td>
<td>telnet</td>
<td>TCP</td>
</tr>
<tr>
<td>Web</td>
<td>No loss</td>
<td>Elastic</td>
<td>No</td>
<td>http</td>
<td>TCP</td>
</tr>
<tr>
<td>File transfer</td>
<td>No loss</td>
<td>Elastic</td>
<td>No</td>
<td>ftp</td>
<td>TCP</td>
</tr>
<tr>
<td>streaming multimedia</td>
<td>Loss tolerant</td>
<td>audio: 5kbps-1Mbps video: 10kbps-5Mbps</td>
<td>Yes, 100’s msec</td>
<td>Proprietary</td>
<td>TCP or UDP</td>
</tr>
<tr>
<td>Remote file server</td>
<td>No loss</td>
<td>Elastic</td>
<td>No</td>
<td>NFS</td>
<td>TCP or UDP (typically UDP)</td>
</tr>
<tr>
<td>Internet telephony</td>
<td>Loss tolerant</td>
<td>Depends on encoding, 32kbps typically</td>
<td>Yes, few secs</td>
<td>SIP, RIP, Prorietary</td>
<td>TCP or UDP (typically UDP)</td>
</tr>
</tbody>
</table>
Internet transport-layer protocols

- **UDP**: unreliable, unordered delivery
  - Process-to-process data delivery
    - Multiplexing/demultiplexing
  - End-to-end error checking
- **TCP**: reliable, in-order delivery
  - Congestion control
  - Flow control
  - Connection setup
- **Services not available:**
  - Delay guarantees
  - Bandwidth guarantees
The programmer's conceptual view of a TCP/IP Internet

© Addison-Wesley Publishers 2000
How Apps Access Transport Services

- Through sockets

- **Socket**: a *host-local, application-created, OS-controlled* interface (a “door”) into which application process can both send and receive messages to/from another application process

- **Socket API**
  - introduced in BSD4.1 UNIX, 1981
  - explicitly created, used, released by apps
  - client/server paradigm
  - two types of transport service via socket API:
    - unreliable datagram
    - reliable, byte stream-oriented
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 delay)
- simple: no connection state at sender, receiver
- small segment header
- no congestion control: UDP can blast away as fast as desired
Client/server socket interaction: UDP

Server (running on hostid)

create socket, port= x.
serverSocket = DatagramSocket()

read datagram from serverSocket

write reply to serverSocket specifying client address, port number

Client

create socket, clientSocket = DatagramSocket()

Create datagram with server IP and port=x; send datagram via clientSocket

read datagram from clientSocket

close clientSocket
Example: Java client (UDP)

Example client-server app:

1) client reads line from standard input (\texttt{inFromUser} stream), sends to server via socket (\texttt{outToServer} stream)

2) server reads line from socket

3) server converts line to uppercase, sends back to client

4) client reads, prints modified line from socket (\texttt{inFromServer} stream)
Example: UDP Java client

```java
import java.io.*;
import java.net.*;

class UDPClient {
    public static void main(String args[]) throws Exception {
        BufferedReader inFromUser =
            new BufferedReader(new InputStreamReader(System.in));

        DatagramSocket clientSocket = new DatagramSocket();

        InetAddress IPAddress = InetAddress.getByName("hostname");

        byte[] sendData = new byte[1024];
        byte[] receiveData = new byte[1024];

        String sentence = inFromUser.readLine();
        sendData = sentence.getBytes();

        // More code here...
    }
}
```
Example: UDP Java client (cont’d)

```java
DatagramPacket sendPacket =
    new DatagramPacket(sendData, sendData.length, IPAddress, 9876);
clientSocket.send(sendPacket);

DatagramPacket receivePacket =
    new DatagramPacket(receiveData, receiveData.length);
clientSocket.receive(receivePacket);

String modifiedSentence =
    new String(receivePacket.getData);

System.out.println("FROM SERVER:" + modifiedSentence);
clientSocket.close();
```
Example: UDP Java server

```java
import java.io.*;
import java.net.*;

class UDPServer {
    public static void main(String args[]) throws Exception {
       DatagramSocket serverSocket = new DatagramSocket(9876);
        byte[] receiveData = new byte[1024];
        byte[] sendData = new byte[1024];

        while(true) {
            DatagramPacket receivePacket =
                new DatagramPacket(receiveData, receiveData.length);
            serverSocket.receive(receivePacket);
        }
    }
}
```
Example: UDP Java server
(cont’d)

```java
String sentence = new String(receivePacket.getData());

InetAddress IPAddress = receivePacket.getAddress();
int port = receivePacket.getPort();

String capitalizedSentence = sentence.toUpperCase();

sendData = capitalizedSentence.getBytes();

DatagramPacket sendPacket = new DatagramPacket(sendData, sendData.length, IPAddress, port);

serverSocket.send(sendPacket);
```

get IP addr
port #, of
sender

create datagram
to send to client

write out
datagram
to socket

end of while loop,
loop back and wait for
another datagram
UDP Use and Format

- often used for streaming multimedia apps
  - loss tolerant
  - rate sensitive
- other UDP uses
  - DNS
  - SNMP
- reliable transfer over UDP: add reliability at application layer
  - application-specific error recovery!

UDP segment format:
- Source port #
- Destination port #
- Length, in bytes of UDP segment, including header
- Checksum

Length, in bytes of UDP segment, including header

Application data (message)

UDP segment format
Multiplexing/demultiplexing

Demultiplexing at rcv host:
delivering received segments to correct socket

Multiplexing at send host:
gathering data from multiple sockets, enveloping data with header (later used for demultiplexing)

= socket       = process

<table>
<thead>
<tr>
<th>application</th>
<th>transport</th>
<th>network</th>
<th>link</th>
<th>physical</th>
</tr>
</thead>
<tbody>
<tr>
<td>P3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>application</th>
<th>transport</th>
<th>network</th>
<th>link</th>
<th>physical</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>application</th>
<th>transport</th>
<th>network</th>
<th>link</th>
<th>physical</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>application</th>
<th>transport</th>
<th>network</th>
<th>link</th>
<th>physical</th>
</tr>
</thead>
<tbody>
<tr>
<td>P4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

host 1

host 2

host 3
How demultiplexing works

- Host receives IP datagrams
  - Each datagram has source IP address, destination IP address
  - Each datagram carries 1 transport-layer segment
  - Each segment has source, destination port number

- Host uses IP addresses & port numbers to direct segment to appropriate socket

TCP/UDP segment format:
- 32 bits
- Source port #
- Destination port #
- Other header fields
- Application data (message)
Connectionless demultiplexing

• **recall**: create sockets with host-local port numbers:
  
  ```java
  DatagramSocket mySocket1 = new DatagramSocket(12534);
  DatagramSocket mySocket2 = new DatagramSocket(12535);
  ```

• **recall**: when creating datagram to send into UDP socket, must specify
  
  (dest IP address, dest port number)

• **when host receives UDP segment:**
  - checks destination port number in segment
  - directs UDP segment to socket with that port number

• **IP datagrams with different source IP addresses and/or source port numbers directed to same socket**
DatagramSocket serverSocket = new DatagramSocket(6428);

SP provides “return address”
UDP checksum

**Goal:** detect “errors” (e.g., flipped bits) in transmitted segment

**Sender:**
- treat segment contents as sequence of 16-bit integers
- checksum: addition (1’s complement sum) of segment contents
- sender puts checksum value into UDP checksum field

**Receiver:**
- compute checksum of received segment
- check if computed checksum equals checksum field value:
  - NO - error detected
  - YES - no error detected.
TCP: Transport Control Protocol

- **point-to-point:**
  - one sender, one receiver
- **reliable, in-order byte steam:**
  - no “message boundaries”
- **pipelined:**
  - TCP congestion and flow control set window size
- **send & receive buffers**

- **full duplex data:**
  - bi-directional data flow in same connection
  - MSS: maximum segment size
- **connection-oriented:**
  - handshaking (exchange of control msgs) inits sender, receiver state before data exchange
- **flow controlled:**
  - sender will not overwhelm receiver
Socket programming with TCP

Client must contact server
- server process must first be running
- server must have created socket (door) that welcomes client’s contact

Client contacts server by:
- creating client-local TCP socket
- specifying IP address, port number of server process
- when client creates socket: client TCP establishes connection to server TCP

- when contacted by client, server TCP creates new socket for server process to communicate with client
  ➡ allows server to talk with multiple clients
  ➡ source port numbers used to distinguish clients

application viewpoint

TCP provides reliable, in-order transfer of bytes ("pipe") between client and server
TCP Client/server socket interaction

Server (running on **hostid**)

- create socket, port=x, for incoming request:
  ```java
  welcomeSocket = ServerSocket()
  ```
- wait for incoming connection request
  ```java
  connectionSocket = welcomeSocket.accept()
  ```
- read request from connectionSocket
- write reply to connectionSocket
- close connectionSocket

TCP connection setup

Client

- create socket, connect to **hostid**, port=x
  ```java
  clientSocket = Socket()
  ```
- send request using clientSocket
- read reply from clientSocket
- close clientSocket
Stream Jargon

- **stream** is a sequence of characters that flow into or out of a process.
- **input stream** is attached to some input source for the process, e.g., keyboard or socket.
- **output stream** is attached to an output source, e.g., monitor or socket.
Socket programming with TCP

Example client-server app:
1) client reads line from standard input (inFromUser stream), sends to server via socket (outToServer stream)
2) server reads line from socket
3) server converts line to uppercase, sends back to client
4) client reads, prints modified line from socket (inFromServer stream)
Example: TCP Java client

```java
import java.io.*;
import java.net.*;

class TCPClient {

    public static void main(String argv[]) throws Exception {
        String sentence;
        String modifiedSentence;

        BufferedReader inFromUser =
            new BufferedReader(new InputStreamReader(System.in));

        Socket clientSocket = new Socket("hostname", 6789);

        DataOutputStream outToServer =
            new DataOutputStream(clientSocket.getOutputStream());

        BufferedReader inFromUser =
            new BufferedReader(new InputStreamReader(System.in));

        String sentence = inFromUser.readLine();
        String modifiedSentence = sentence.toUpperCase();

        outToServer.writeUTF(modifiedSentence);
    }
}
```
Example: TCP Java client (cont’d)

```java
BufferedReader inFromServer = new BufferedReader(new InputStreamReader(clientSocket.getInputStream()));

sentence = inFromUser.readLine();
outToServer.writeBytes(sentence + '
');
modifiedSentence = inFromServer.readLine();
System.out.println("FROM SERVER: " + modifiedSentence);
clientSocket.close();
```
Example: TCP Java server

```java
import java.io.*;
import java.net.*;

class TCPServer {
    public static void main(String argv[]) throws Exception {
        String clientSentence;
        String capitalizedSentence;
        ServerSocket welcomeSocket = new ServerSocket(6789);

        while(true) {
            Socket connectionSocket = welcomeSocket.accept();
            BufferedReader inFromClient =
            new BufferedReader(new InputStreamReader(connectionSocket.getInputStream()));
            String clientSentence = inFromClient.readLine();
            String capitalizedSentence = clientSentence.toUpperCase();
            System.out.println(capitalizedSentence);
        }
    }
}
```
Example: TCP Java server (cont’d)

- create output stream, attached to socket
  
  ```java
  DataOutputStream outToClient =
  new DataOutputStream(connectionSocket.getOutputStream());
  ```

- read in line from socket
  
  ```java
  clientSentence = inFromClient.readLine();
  ```

- write out line to socket
  
  ```java
  capitalizedSentence = clientSentence.toUpperCase() + '\n';
  outToClient.writeBytes(capitalizedSentence);
  ```

- end of while loop, loop back and wait for another client connection
TCP segment structure

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>source port #</td>
<td>Source port number</td>
</tr>
<tr>
<td>dest port #</td>
<td>Destination port number</td>
</tr>
<tr>
<td>sequence number</td>
<td>Sequence number of bytes transmitted</td>
</tr>
<tr>
<td>acknowledgement number</td>
<td>Acknowledgement number of bytes received</td>
</tr>
<tr>
<td>head len</td>
<td>Length of TCP header</td>
</tr>
<tr>
<td>not used</td>
<td>Bit indicating that the header length is not used</td>
</tr>
</tbody>
</table>
| U, A, P, R, S, F      | Flags indicating:
|                        | URG: urgent data (generally not used)            |
|                        | ACK: ACK # valid (generally not used)            |
|                        | PSH: push data now (generally not used)          |
|                        | RST, SYN, FIN: connection establishment (setup, |
|                        | teardown commands)                               |
| checksum               | Checksum of the TCP payload                     |
| Urg data ptr           | Pointer to urgent data                           |
| Options (variable length) | Options segment (if present)                   |
| application data       | Application data (variable length)               |

Internet checksum (as in UDP)

Counting by bytes of data (not segments!)

# bytes rcvr willing to accept
Multiplexing/demultiplexing

Demultiplexing at rcv host:
delivering received segments to correct socket

Multiplexing at send host:
gathering data from multiple sockets, enveloping data with header (later used for demultiplexing)

= socket = process

<table>
<thead>
<tr>
<th>application</th>
<th>P3</th>
<th>transport</th>
<th>network</th>
<th>link</th>
<th>physical</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>host 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>application</th>
<th>P1</th>
<th>transport</th>
<th>network</th>
<th>link</th>
<th>physical</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>host 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>application</th>
<th>P2</th>
<th>transport</th>
<th>network</th>
<th>link</th>
<th>physical</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>host 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Connection-oriented demux

- TCP socket identified by 4-tuple:
  - source IP address
  - source port number
  - dest IP address
  - dest port number

- recv host uses all four values to direct segment to appropriate socket

- server host may support many simultaneous TCP sockets:
  - each socket identified by its own 4-tuple

- web servers have different sockets for each connecting client
  - non-persistent HTTP will have different socket for each request
Connection-oriented demux (cont)
Principles of Reliable Data Transfer

- important in app., transport, link layers

- characteristics of unreliable channel will determine complexity of reliable data transfer protocol

- Note: slides use the term “packet” but at transport layer, these are segments
Reliable data transfer: getting started

**send side**

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

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

**receive side**

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

- `rdt_rcv()` : called when packet arrives on rcv-side of channel

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

- `rdt_send()` : called from above, (e.g., by app.). Passed data to deliver to receiver upper layer
Reliable transfer over a reliable channel

- underlying channel perfectly reliable
  - no bit errors
  - no loss of packets
- separate FSMs for sender, receiver:
  - sender sends data into underlying channel
  - receiver read data from underlying channel

sender

receiver
What can go wrong (1)?

- Underlying channel may flip bits in segment
- Error detection:
  - Checksum to detect bit errors
- Recovering from errors:
  - *acknowledgements (ACKs)*: receiver explicitly tells sender that segment received OK
  - *negative acknowledgements (NAKs)*: receiver explicitly tells sender that segment had errors
    - sender retransmits segment on receipt of NAK
- *Stop-and-wait*
  - Sender sends one segment, then waits for the receiver to respond
  - We will come back to this later
Handling duplicates

• What happens if ACK/NAK corrupted?
  ➡ sender doesn’t know what happened at receiver!
  ➡ can’t just retransmit: possible duplicate

• Sender retransmits current segment if ACK/NAK garbled

• Sender adds sequence number to each segment
  ➡ For stop-and-go protocol a 1-bit sequence number with modulo-2 arithmetic is sufficient

• Receiver discards (does not deliver up to the application) duplicate segments
NAK-free Protocol

• It is possible to eliminate NAKs as negative acknowledgement
• instead of NAK, receiver sends ACK for last segment received OK
  ➔ receiver must \emph{explicitly} include sequence number of segment being ACKed
• duplicate ACK at sender results in same action as NAK: \textit{retransmit current packet}
What can go wrong (2)?

- Segments (data or ACK) may be lost
- Sender waits a “reasonable” amount of time for ACK
  - Retransmits if no ACK received in this time
  - If segment (data or ACK) simply delayed (not lost)
    - Retransmission will be duplicate, but use of sequence numbers already handles this
    - Receiver must specify the sequence number of segment being ACKed
  - Requires countdown timer
- Sequence numbers
  - For data: byte stream “number” of first byte in segment’s data
  - For ACKs: if pipelined, segment number of next byte expected from other side
    - Cumulative ACK
What can go wrong (3)?

• Data segments may come out of order
• TCP specification does not say what to do
• Two alternatives
  ➡ Receiver discards out of order segments
    ✦ Simplifies receiver design
    ✦ Wastes network bandwidth
  ➡ Receiver keeps out of order segments and fills in the missing ones when they arrive
    ✦ Usually this is what is implemented
TCP sender events:

**data rcvd from app:**
- Create segment with sequence number
- start timer if not already running (think of timer as for oldest unacked segment)
- expiration interval: TimeOutInterval

**timeout:**
- retransmit segment that caused timeout
- restart timer

**Ack rcvd:**
- If acknowledges previously unacked segments
  - update what is known to be acked
  - start timer if there are outstanding segments
TCP sender
(simplified)

NextSeqNum = InitialSeqNum
SendBase = InitialSeqNum

loop (forever) {
    switch(event)

    event: data received from application above
        create TCP segment with sequence number NextSeqNum
        if (timer currently not running)
            start timer
        pass segment to IP
        NextSeqNum = NextSeqNum + length(data)

    event: timer timeout
        retransmit not-yet-acknowledged segment with smallest sequence number
        start timer

    event: ACK received, with ACK field value of y
        if (y > SendBase) {
            SendBase = y
            if (there are currently not-yet-acknowledged segments)
                start timer
        }

} /* end of loop forever */
TCP: retransmission scenarios

Host A
Seq=100, 20 bytes data

Host B
Seq=92, 8 bytes data

loss

X

Host A
Seq=92, 8 bytes data

Host B
ACK=100

lost ACK scenario

Host A
Seq=92 timeout

Host B
ACK=100

premature timeout

Host A
Seq=100, 20 bytes data

Host B
ACK=120

Host A
Seq=92, 8 bytes data

Host B
ACK=120
TCP retransmission scenarios (more)

Cumulative ACK scenario

Host A
Seq=92, 8 bytes data
ACK=100

Host B
X
Seq=100, 20 bytes data
ACK=120

timeout

loss
## TCP ACK generation

<table>
<thead>
<tr>
<th>Event at Receiver</th>
<th>TCP Receiver action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arrival of in-order segment with expected seq #. All data up to expected seq # already ACKed</td>
<td>Delayed ACK. Wait up to 500ms for next segment. If no next segment, send ACK</td>
</tr>
<tr>
<td>Arrival of in-order segment with expected seq #. One other segment has ACK pending</td>
<td>Immediately send single cumulative ACK, ACKing both in-order segments</td>
</tr>
<tr>
<td>Arrival of out-of-order segment higher-than-expect seq. #. Gap detected</td>
<td>Immediately send <em>duplicate ACK</em>, indicating seq. # of next expected byte</td>
</tr>
<tr>
<td>Arrival of segment that partially or completely fills gap</td>
<td>Immediate send ACK, provided that segment starts at lower end of gap</td>
</tr>
</tbody>
</table>
Fast Retransmit

- time-out period often relatively long:
  - long delay before resending lost packet
- detect lost segments via duplicate ACKs.
  - sender often sends many segments back-to-back
  - if segment is lost, there will likely be many duplicate ACKs.
- if sender receives 3 ACKs for the same data, it supposes that segment after ACKed data was lost:
  - fast retransmit: resend segment before timer expires
Resending segment after triple duplicate ACK

Host A

timeout

resend 2nd segment

time

Host B

X
Fast retransmit algorithm:

**event:** ACK received, with ACK field value of y
  
  if (y > SendBase) {
    SendBase = y
    if (there are currently not-yet-acknowledged segments)
      start timer
  }
  else {
    increment count of dup ACKs received for y
    if (count of dup ACKs received for y = 3) {
      resend segment with sequence number y
    }
  }

A duplicate ACK for already ACKed segment  
Fast retransmit
TCP Round Trip Time and Timeout

- 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

- 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 and Timeout

EstimatedRTT = (1 - \(\alpha\))*EstimatedRTT + \(\alpha\)*SampleRTT

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

RTT: gaia.cs.umass.edu to fantasia.eurecom.fr
TCP Round Trip Time and Timeout

Setting the timeout

- **EstimatedRTT** plus “safety margin”
  - large variation in **EstimatedRTT** -> larger safety margin
- first estimate of how much SampleRTT deviates from EstimatedRTT:

  \[
  \text{DevRTT} = (1-\beta) \times \text{DevRTT} + \\
  \beta \times |\text{SampleRTT} - \text{EstimatedRTT}|
  \]

  (typically, \(\beta = 0.25\))

Then set timeout interval:

\[
\text{TimeoutInterval} = \text{EstimatedRTT} + 4 \times \text{DevRTT}
\]
TCP Flow Control

- receive side of TCP connection has a receive buffer:

  - speed-matching service: matching the send rate to the receiving app’s drain rate

  - app process may be slow at reading from buffer

  - flow control: sender won’t overflow receiver’s buffer by transmitting too much, too fast
TCP Flow control: how it works

- spare room in buffer (ignoring out-of-order segments)
  \[ RcvWindow = RcvBuffer - [LastByteRcvd - LastByteRead] \]
- Receiver advertises spare room by including value of \textit{RcvWindow} in segments
- Sender limits unACKed data to \textit{RcvWindow}
  - guarantees receive buffer doesn’t overflow
Recall: TCP is a connection-oriented protocol

- sender, receiver establish “connection” before exchanging data segments
- initialize TCP variables:
  - Sequence numbers
  - buffers, flow control info (e.g. RcvWindow)
- client: connection initiator
  
  ```java
  Socket clientSocket = new Socket("hostname","port number");
  ```
- server: contacted by client
  
  ```java
  Socket connectionSocket = welcomeSocket.accept();
  ```
Three-Way Handshake

**Step 1:** client host sends TCP SYN segment to server
- specifies initial sequence number
- no data

**Step 2:** server host receives SYN, replies with SYNACK segment
- server allocates buffers
- specifies server initial sequence number

**Step 3:** client receives SYNACK, replies with ACK segment, which may contain data
Closing a TCP Connection

client closes socket:
\texttt{clientSocket.close();}

**Step 1:** client end system sends TCP FIN control segment to server

**Step 2:** server receives FIN, replies with ACK. Closes connection, sends FIN.

**Step 3:** client receives FIN, replies with ACK.

- Enters “timed wait” - will respond with ACK to received FINs

**Step 4:** server, receives ACK.
Connection closed.

**Note:** with small modification, can handle simultaneous FINs.
TCP Connection States

TCP client lifecycle

TCP server lifecycle
Principles of Congestion Control

• **Congestion**: informally, “too many sources sending too much data too fast for the network to handle”

• different from flow control!

• manifestations:
  - lost packets (buffer overflow at routers)
  - long delays (queueing in router buffers)
Causes/costs of congestion: Simple Scenario

- two senders, two receivers
- one router, infinite buffers
- no retransmission

- large delays when congested
- maximum achievable throughput
Causes and Effects of Congestion

• With finite buffers at routers, packets may be dropped as $\lambda_{in}$ increases
  ➡ Retransmission needed
  ➡ Offered load $\lambda'_{in} > \lambda_{in}$
    – More work (retransmissions) for given $\lambda_{out}$
    – Unneeded retransmissions: link carries multiple copies of segment

• With multi-hop connections, upstream routers receive two types of traffic:
  ➡ Forwarded traffic from downstream routers
  ➡ Traffic they may receive directly from hosts
  ➡ As $\lambda'_{in}$ increases, more dropped packages and more transmissions
    ✦ $\lambda_{out}$ will approach 0
    – When packet dropped, any upstream transmission capacity used for that packet was wasted!
Approaches to Congestion Control

Two broad approaches towards congestion control:

1. **end-end congestion control:**
   - no explicit feedback from network
   - congestion inferred from end-system observed loss, delay
   - approach taken by TCP

2. **network-assisted congestion control:**
   - routers provide feedback to end systems
     - single bit indicating congestion (SNA, DECbit, TCP/IP ECN, ATM)
     - explicit rate sender should send at
TCP congestion control

- **Approach:** increase transmission rate (window size), probing for usable bandwidth, until loss occurs
  - **Additive increase:** increase \( cwnd \) by 1 MSS (maximum segment size) every RTT until loss detected
  - **Multiplicative decrease:** cut \( cwnd \) in half after loss

- **Algorithm has three components:**
  - **Slow start**
    - Initially set \( cwnd = 1 \) MSS
    - double \( cwnd \) every RTT
    - done by incrementing \( cwnd \) for every ACK received
  - **Congestion avoidance:** After timeout or 2 duplicate ACKS
    - \( cwnd \) is cut in half
    - window then grows linearly
  - **Fast recovery:** After a timeout
    - \( cwnd \) set to 1 MSS;
    - window then grows exponentially
    - to a threshold, then grows linearly